EFFECTS OF APPRAISAL PURPOSE AND RATING FORMAT ON PERFORMANCE APPRAISAL ACCURACY

by

Marta L. Carter-Stuart

Thesis submitted to the Faculty of the

Virginia Polytechnic Institute and State University in

partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

Psychology

APPROVED:

Roseanne J. Foti, Chair

Joseph J. Franchina

Neil M. A. Hauenstein

April, 1989

Blacksburg, Virginia

EFFECTS OF APPRAISAL PURPOSE AND RATING FORMAT ON PERFORMANCE APPRAISAL ACCURACY

by

Marta L. Carter-Stuart

Committee Chair: Roseanne J. Foti

Psychology

(ABSTRACT)

The principle of encoding specificity states that effective information retrieval relies upon consistency of encoding and retrieval cues. The present study generalized this principle to a complex social interaction in order to investigate the relation between certain combinations of pre- and post-observational cues and their effects on information categorization, recognition accuracy, and judgment accuracy. It was hypothesized that two experimental factors, appraisal purpose and rating format categorization, would influence organization, retrieval, and judgment of performance information. Specifically, consistent encoding (purpose) and retrieval (format) cues were expected to result in the most efficient retrieval of information, and consequently in more accurate performance ratings.

Results provided support for information categorization predictions. In summary, subjects asked to form impressions of ratees, organized information by person in free recall

while subjects asked to remember behaviors organized information by behavior. No predicted purpose x format interactions emerged for either recognition or judgment measures.

However, several interesting findings did surface. On the recognition measure, a format effect emerged in which subjects given the behavior-categorized format exhibited greater recognition accuracy than subjects given the person categorized format. Second, a format effect emerged on the judgment measure in which subjects given the person-categorized format exhibited greater judgment accuracy than subjects given the dimension-categorized format. Since the accuracy measures used for the recognition and judgment forms are not statistically comparable, no direct comparison was made to test this evidence of an interaction. However, suggestions are made for a future test of this relationship.

Finally, within subjects analysis revealed a significant ratee by prototype interaction for recognition accuracy. In short, ratings were more accurate on bad behaviors for the two good ratees than for the bad ratee. Likewise, ratings were more accurate on good behaviors for the bad ratee than for the two good ratees. This finding supports an incongruity biased-encoding model (Hastie & Parks, 1986) which states that accuracy is greater for behaviors that are inconsistent (in this case bad behaviors for good ratees and good behaviors for bad ratees) with overall information.

ACKNOWLEDGMENTS

I extend my appreciation to the members of my thesis committee, Dr. Roseanne J. Foti, Dr. Joseph J. Franchina, and Dr. Neil M. A. Hauenstein, for their thoughtful insight and sound advice. In particular I thank Dr. Foti, my advisor, for guidance throughout the past two years. Finally, I am grateful to and to my parents,

, for their unconditional support.

TABLE OF CONTENTS

Abstract ii

Acknowledgments iv

List of Tables vii

List of Figures ix

Introduction 1

Literature Review 3

Overview 20

Hypotheses 22

Method 25

Results 35

Discussion 43

Summary and Conclusions 47

References 49

Tables 55

Figures 68

Appendix A Consent Form 70

Appendix B Script (Proper Bagging Technique) 72

Appendix C Recognition Measure Organized by Person 74

Appendix D Recognition Measure Organized by Behavior 81

Appendix E Judgment Measure Organized by Person . . . 88

Appendix F Judgment Measure Organized by Behavior 93

Appendix G Free Recall Measure 96

Appendix H Manipulation Check 98

Appendix I Signal Detection Equations 100

Appendix J Judgment Accuracy Equations 102

Vita 103

LIST OF TABLES

1. Summary of ANOVA for Impression Purpose Manipulation Check as a Function of Purpose and Format and Cell Descriptive Statistics 55

2. Summary of ANOVA for Behavior Purpose Manipulation Check as a Function of Purpose and Format and Cell Descriptive Statistics 56

3. Summary of ANOVA for Adjusted Ratio of Clustering (ARC) Index for Person Categories as a Function of Purpose and Format and Cell Descriptive Statistics 57

4. Summary of ANOVA for Adjusted Ratio of Clustering (ARC) Index for Behavior Categories as a Function of Purpose and Format and Cell Descriptive Statistics 58

5. Summary of ANOVA for Accuracy (sensitivity) on Recognition Measure as a Function of Purpose, Format, Bagger, and Item Prototypicality 59

6. Cell Descriptive Statistics for Accuracy (sensitivity) on Recognition Measure as a Function of Purpose, Format, Bagger, and Item Prototypicality 60

7. Summary of ANOVA for False Positive Rate on Recognition Measure as a Function of Purpose, Format, Bagger, and Item Prototypicality 61

8. Cell Descriptive Statistics for False Positive Rate on Recognition Measure as a Function of Purpose, Format, Bagger, and Item Prototypicality 62

9. Summary of ANOVA for Elevation of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics 63

10. Summary of ANOVA for Differential Elevation of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics 64

11. Summary of ANOVA for Stereotypical Accuracy of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics 65

12. Summary of ANOVA for Differential Accuracy of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics 66

13. Summary of ANOVA for Overall Accuracy of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics 67

LIST OF FIGURES

1. Bagger x Item Prototypicality Interaction on Recognition Accuracy 68

Effects of Appraisal Purpose and Format on Performance Appraisal Accuracy

Introduction

Performance appraisals have come to serve many functions in organizational settings. Appraisal information is used for deciding promotions, determining pay raises, developing training programs, and providing feedback to employees. In realizing the value of this rating information when it is reliable and valid, researchers have generally attempted to either train raters to reduce psychometric errors or to improve rating formats. Neither strategy alone has proven fruitful (Landy & Farr, 1980; Pulakos, 1984). More recently, performance appraisal purpose has become recognized as a component with power to influence performance appraisal accuracy (Foti & Lord, 1987; Williams, Denisi, Blencoe & Cafferty, 1985). Addressing more complex issues, Pulakos (1986) has begun to investigate potential interactive effects of rating task (purpose) and rater training on performance appraisal accuracy and has suggested that certain combinations of rater training and rating format may be associated with increased rating accuracy.

Pulakos' research may be interpreted as an in-depth approach to performance appraisal, whereby the interactive effects of pre-observational and post-observational cues on performance appraisal accuracy are observed. The present study followed Pulakos' line of reasoning to investigate whether certain combinations of pre- and post-observational cues are linked to performance appraisal accuracy.

Specifically, this study investigated whether type of appraisal purpose results in a differential search for information in the environment and if that search results in improved accuracy of performance evaluations when the rating format used to evaluate ratees is consistent with appraisal purpose. In basic terms, this study applied the principle of encoding specificity to the issue of performance appraisal.

The principle of encoding specificity states that for a retrieval cue to be effective, the information to be retrieved must have been stored in a fashion that matches the retrieval cue itself (Tulving and Osler, 1968; Tulving and Thomson, 1973). In other words, consistency of encoding and retrieval cues enables successful retrieval of information.

In this study, an attempt was made to influence the accuracy of performance appraisal by manipulating the consistency between rating purpose (encoding cue) and rating format (retrieval cue) in order to influence both schema-guided encoding and retrieval of information. Research has

shown that rating purpose can affect how information is schematically encoded (Cohen, 1981) while rating format can affect how information is schematically retrieved for judgment purposes. This study questioned whether greater accuracy in performance ratings occurred when purpose and format were consistent than when they were inconsistent.

Literature Review

Encoding Specificity

It has long been known that how well something is remembered depends not only on *what* it is but also on *how* it is stored in memory. Thus, the presence of many recall strategies in day to day life is not surprising. This idea of how something is stored underlies the principle of encoding specificity which, in its most general form, asserts that something can be retrieved only if it has been stored and that how it was stored dictates how it can be retrieved (Tulving & Thomson, 1973). In more detail, the encoding specificity principle states that the target item of memory must be encoded in reference to a particular cue for that cue to be effective at retrieval (Tulving & Thomson, 1973).

According to Tulving and Thomson (1973), specific encoding operations performed on perceived information determine what is stored, and what is stored determines what type of retrieval cues are effective in providing access to

what is stored. They also state that specific encoding operations performed on an input are not usually directly identifiable but that they can be experimentally manipulated through instructions and other means. This is relevant to the present study in which purpose of performance appraisal was manipulated in order to influence raters' encoding operations.

The literature shows that the effects of encoding operations on the memory trace (what is stored) have been studied in a variety of settings and have used a number of different techniques and paradigms (e.g., Asch, 1969; Tulving & Thomson, 1971; Da Polito, Barker & Wiant, 1972; Tulving and Thomson, 1973; Mandler, 1967; Hyde & Jenkins 1969; Craik, 1973; & Schulman, 1971). For example, Asch (1969) studied the effects of encoding operations on what is stored by changing the context of an item. He found that repetition of a to-be-remembered item may or may not facilitate its recall, depending upon its intralist context. Similarly, recognition of a previously seen list word was shown by Tulving and Thomson (1971) and by DaPolito, Barker, and Wiant (1972) to be influenced by its presentation and test contexts.

In three experiments by Tulving and Thomson (1973), encoding of target words was influenced by the list cues present at input and by the subjects' expectations that they would be tested with those cues. Thus, specific encoding of

target words (input materials) in these studies was responsible for the nature of effective retrieval information.

In another type of experiment, encoding was manipulated by asking the subjects to do different things with the material when it was presented. For instance, Mandler (1967) had one group of subjects sort words into conceptual categories while another group was exposed to the same words under instructions to study and remember them. In a subsequent free-recall test, both groups did equally well.

In a study by Hyde and Jenkins' (1969), subjects in one group made judgments about semantic properties of words, while in another group they studied the same set of words in expectation of a recall test. Both groups recalled the same number of words and did considerably better than a third group that had made judgments about graphemic properties of words prior to the test.

Moreover, Craik (1973) offered an example of an effective manipulation of input in terms of encoding operations performed on presented items. In this study, words were presented tachistoscopically and subjects answered different questions about each word such as, "Is it printed in capital or lower-case letters?" or "Does it belong to the category of fruits?" Depending upon the encoding operation performed at input, large differences in recall of these

words were observed. Finally, Schulman (1971) found that recognition memory for targets from a previously scanned word list depends critically on how those targets were defined. In effect, when a subject searched for semantically defined targets, his later ability to recognize all words scanned was greater than after a search for targets structurally defined. This and the previously discussed studies provide support for the important role that encoding operations play in determining subsequent retrievability of perceived items.

Furthermore, many experiments have demonstrated that recall and recognition of items stored under identical encoding conditions are influenced by the nature of information present in the retrieval environment (Winograd & Conn, 1971; Light, 1972; Tulving & Psotka, 1971). For example, in two experiments performed by Winograd and Conn (1971), homographic nouns (words which are invariant in spelling but which have more than one meaning) were presented to subjects for study without context and then tested for recognition in either a relatively familiar sentence context, a relatively unfamiliar context, or no context. Both experiments showed that supplying a more familiar semantic representation of a homograph originally presented was more effective than an unfamiliar context or than no context at all.

Light (1972) also investigated the properties of effective retrieval cues and found that certain retrieval cues (homonyms) were more effective than others for words that were encoded under the same conditions. These studies suggest that a stored trace of an item is more accessible through certain cues than others and lend support to to Earhard (1969) who said that memory traces may be said to vary in strength, or quality, or durability, but more importantly they vary in the specificity of code they carry as to the effectiveness of various kinds of retrieval information that govern the recovery of the stored information.

Finally, in the area of performance appraisal, Larson, Lingle, and Scerbo (1984) have found strong preobservation and postobservation cue effects on leader-behavior ratings. In this study, raters watched a videotape of a problem-solving group and later rated the group leader's behavior. Raters' responses on leader-behavior questionnaires were influenced when their perceptions of the group's performance were manipulated before or after watching the tape.

Thus, research has shown it possible to hold constant a to-be-recalled item and observe large differences in its recall and recognition depending upon its encoding conditions. On the other hand, it is possible to hold constant the encoding conditions of the item and observe

large differences in its recall and recognition depending upon retrieval conditions. In summary, these two basic sources of variability of recall, encoding and retrieval conditions, interact to suggest that the effectiveness of a particular cue depends on how the to-be-retrieved item was encoded at input. Data described in the literature (Earhard, 1969; Frost, 1972; Tulving & Thomson, 1971; & Tulving & Thomson, 1973) demonstrate such an interaction.

For example, Earhard (1969) explored the matter of independent versus organized storage of unrelated items by requiring subjects to recall, in alphabetical order, words they had been memorizing for various numbers of trials under free, serial, or cued recall conditions. Results supported the position that appropriate storage facilitates retrieval by providing information as to the retrieval cues and the direction in which to search for the next item during recall.

Furthermore, Frost (1972) examined memory for visually and semantically categorized pictures. In four studies where stimuli were designed so that recall organization could be based on visual, semantic, or both types of categories, free recall and recognition tasks were performed by subjects who expected either recall or recognition memory tests. Subjects who expected recognition recalled by combining visual and semantic categories and could efficiently perform a visual recognition task. However, the subjects who expected recall

clustered by semantic categories only and were efficient only in name recognition. It was concluded that pictures are encoded differently depending on task expectation and that parallel access of visual and semantic memory codes does occur, but when recognition is expected, a visual cue provides faster access, and when expecting recall, verbal access is more efficient.

Moreover, Tulving and Thomson (1971) found that recognition of a single word was impaired when another, associatively related, word accompanied it at the time of the test. Also, they found that recognition of a word presented at input as a member of an associatively related pair was impaired when the other member of the pair was removed or changed at the time of the test. In another study, Tulving and Thomson (1973) found that specific encoding of target materials was responsible for the nature of effective retrieval information. In other words, how information was encoded governed which retrieval cue was most effective.

Finally, in performance appraisal literature, Pulakos (1986) found a significant training by rating format interaction on performance ratings in a study where subjects were randomly assigned to conditions defined by combinations of training and rating task. Mean comparisons revealed that accuracy was generally greater for congruent training and rating task conditions. In the present study, performance

appraisal purpose was manipulated in order to influence raters' encoding operations while rating format acted as a retrieval cue. When appraisal purpose and rating format were consistent, performance appraisal accuracy was hypothesized to be greater than when the two were inconsistent.

Two basic assumptions underlying these predictions are that appraisal purpose affects the schematic encoding and organization of information and that rating format affects the schematic retrieval of information. The following section defines schemata on a general level and discusses the specific types of schemata relevant to this study.

Schemata

A schema may be defined as a cognitive structure which provides observers with an organized knowledge base about a particular concept or type of stimulus and guides the processing of new information and the retrieval of stored information. Furthermore, it contains information about the attributes of a concept and the relationships among these attributes (Fiske & Taylor, 1984; Fiske & Linville, 1980). Perceivers' schemata lend order and coherence to their processing of other people's behavior. In intuitive terms, schemata help to translate the ongoing stream of an actor's behavior into meaningful "chunks" of information, contribute meaning to the extracted information, and facilitate memory for schema-relevant behavior (Cohen, 1981).

Before discussing schemata in more detail, the terms "schema-relevant" and "schema-consistent" should be clarified. People do not necessarily remember information that confirms their schema (schema-consistent) and forget information that disconfirms it (schema-inconsistent). In fact, memory advantages can accrue to both consistent and inconsistent information while non-important (schema-irrelevant) information is easily forgotten. For instance, Hastie (1981) suggests that if people have enough time, they give added attention at input to schema-inconsistent information by elaborating and explaining the inconsistency. This, in turn, strengthens the memory trace.

Schema-consistent information, on the other hand, is favored by normal retrieval routes if the schemata stored in memory contain typical information rather than exceptions (Fiske & Taylor, 1984). Therefore, the impact of inconsistency on memory should depend on whether one concentrates on the inconsistency when it is first encountered and integrates it into memory. In fact, evidence exists which suggests that attentional processes can actually account for which type of information (consistent or inconsistent) will have an advantage at a given time.

For instance, Sentis & Burnstein (1979) found that remembering inconsistent information does seem to depend on whether sufficient study time is allowed. Furthermore,

Brewer, Dull, & Lui (1981) found that inconsistent information required longer encoding time than did consistent or irrelevant information. Consistent information was well recalled (despite its short initial processing time) presumably because it was already stored as part of the knowledge structure, while inconsistent information was well remembered, presumably because of the added attention at input.

In fact, people who encounter an inconsistency normally take time to explain it so that it fits into an existing schema-based impression. This can be accomplished by attributing the inconsistent behavior to temporary situational causes. This makes the behavior irrelevant and it can thus be forgotten. However, if one attributes the inconsistent behavior to dispositional causes, the behavior must then be fit into the existing schema-based impression of the person's personality (Fiske & Taylor, 1984).

In summary, the explanation of inconsistent information takes extra time during information encoding. Furthermore, people presumably remember inconsistent behavior when they attribute it to situational causes. On the other hand, they forget inconsistent behavior and maintain their schema-based impression when they attribute the inconsistent behavior to situational causes. In contrast, if consistent information is stored as part of the social schema, it will be easily

remembered regardless of the encoding processes (Fiske & Taylor, 1984).

A related topic involves misremembrance of non-present, category-consistent attributes. Once a perceiver places a person or event into a particular category, he/she is likely to misremember the presence of category-consistent but never-seen attributes. In other words, people will falsely recognize traits or events that are related to a category but in fact are not contained in a specific instance they encountered. In sum, schemata typically help the perceiver to remember schema-consistent information in more detail than would be possible without the schema. However, the likely interference of nonpresent schema-relevant information is a problem that should not be ignored (Fiske & Taylor, 1984).

The influence of schemata may occur at one or more of the three stages of information processing: encoding, storage, and retrieval (Cohen, 1981). Cohen offers an excellent synopsis of how schemata operate in the course of information processing. Initially, a perceiver's operative schemata guide the encoding of an actor's stream of behavior. Then, "these schemata provide an interpretive framework for determining the 'meaning' of the behavior and will influence what is stored from an actor's behavior. Finally, when a perceiver tries to retrieve information about the actor's behavior from memory, the relevant schemata will

be reactivated and used to fill in forgotten or never seen behavioral features or relationships" (Cohen, 1981, p. 50).

One deduction of the previous discussion of schemata would suggest that the schema used during encoding should be reactivated during retrieval. A proposed area for application of this suggestion is performance appraisal where the type of schema used to direct information encoding is influenced by the purpose of appraisal (Cohen, 1981). However, rating formats used in performance appraisal often do not help to reactivate the type of schema that was used during information encoding. In order to be fully effective retrieval cues, though, rating formats should be consistent with the appraisal purpose. In the present study, encoding cues (appraisal purpose) and retrieval cues (rating format) were manipulated to influence schema-directed encoding and retrieval in order to find combinations of the two that most positively affect performance appraisal accuracy.

All types of schemata can guide perception, memory, and inference of information. Essentially, schemata fall into one of four groups: person schemata, self-schemata, role schemata, and event schemata (scripts). Two types of schemata were relevant to this study, person schemata and script schemata. In the following sections, person and script schemata will be defined, and schemata activation through appraisal purpose and rating format will be discussed.

Person Schemata. Person schemata aid in perception and memory of individuals and schema-relevant information. Placing another person's behavior into the proper trait category, which one learns to do through direct and indirect experience, is similar to categorizing an object (Hastie, 1981; Fiske & Taylor, 1984). In the same manner that we realize that an exercise bicycle is a kind of muscle toning machine and not a vehicle for transportation, our perception of a person who tells us what to do is shaped by our categorizing him by the trait "helpful" as opposed to the trait "insensitive".

According to Foti and Lord (1987), individuals who use a person schema simplify their perception of others by classifying them into preexisting categories based on the similarity of the stimulus to a prototype. A prototype is a collection of the most typical or most highly related features associated with a certain category. Knowledge about any given category is structured around and represented in memory as a prototype. Social prototypes allow people to categorize instances by the family resemblance criterion meaning that while no one insensitive person may possess all the same insensitive traits, having several of them nonetheless identifies the person as an example of an insensitive person.

Scripts or Event Schemata. Scripts provide a guide to behavior that is appropriate to hundreds of stereotypic situations. Examples are riding a bus, visiting a dentist, placing an operator-assisted telephone call, asking for directions, and so on. Through direct or vicarious experiences, each person acquires hundreds of such cultural stereotypes along with his idiosyncratic variations (Bower, Black, & Turner, 1979). Schank and Abelson (1977) use the term "script" to refer to a person's memory structure which encodes his/her general knowledge of a certain situation-action routine.

A script consists of several components. these components are (1) roles to be played, (2) standard props or objects, (3) ordinary conditions for entering upon the activity, (4) a standard sequence of scenes or actions wherein one action enables the next, and (5) some normal results from performing the activity successfully. Scripts aid in the planning and execution of conventional activities as well as enable understanding when a person observes or reads about someone performing another instance of a conventional activity (Bower, Black, & Turner, 1979).

A weak script, defined by Giola and Poole (1984) serves to organize expectations about the behavior of people in certain situations without specifying the exact sequence of these behaviors. For example, performance appraisals

(Feldman, 1981) appear to involve weak script-based understanding and behavior. Both managers and subordinates are likely to have cognitively structured expectations about what is supposed to happen in general during performance appraisal, but they find it difficult to predict a specific order for these events (Giola & Poole, 1984).

The two following sections will discuss the feasibility of manipulating purpose of appraisal and rating format to influence either script-guided or person-guided encoding and retrieval of information.

Schematic-guided Encoding Through Appraisal Purpose

Landy and Farr (1980) posit in their model of performance appraisal that appraisal purpose has a substantial effect on cognitive processes of the rater, and, thus, on the rating outcome. This idea has been expanded by Denisi and his colleagues (Williams, Denisi, Blencoe, & Cafferty, 1985; Williams, Denisi, Meglino, & Cafferty, 1986) who have looked at the impact of appraisal purpose on how raters search for and retrieve information. Their general consensus was that the purpose for an appraisal may act to cue raters to search for, weigh, and encode information in different ways.

The observational goal (Cohen, 1981) serves the function of focusing the perceiver on those categories or features of behavior that are relevant to the purpose at hand. In

effect, it serves the important function of schema selection resulting in a schema choice which will help the perceiver achieve his goal. According to Cohen (1981), perceivers with different observational goals will, at times, chunk up the stream of behavior differently, interpret and store the information differently, and retrieve from memory different information about the actor.

Foti & Lord (1987) illustrated the impact of observational purpose in a study where they influenced schematic processing by manipulating observational purpose and knowledge of task goals to assess the impact of prototype (person) and script (event) processing on recall and recognition accuracy of leader behavior. Subjects watched a videotaped recording of a school board meeting. Findings showed that subjects who received a "memory" observational purpose or knowledge about group goals processed information using a script schema, while subjects who received an "impression-formation" observational purpose or no knowledge of the group goals used a leader prototype schema. The results of the study lent support to the contention that memory for complex information is highly dependent on the cognitive schema into which new information has been integrated.

Schematic-guided Judgment and Recognition Through Rating Format

Inference is the process of collecting and combining often diverse and complex information into a judgment (Fiske & Taylor, 1984). Hamilton (1981) pointed out that the general process of making judgments includes encoding and organization of stimulus information in terms of a schematic structure resulting in a cognitive representation of the target person. Thus, subsequent judgments made about that person will be based on the perceiver's cognitive representation and retrieval of that representation, not on the actual stimulus.

In addressing the potentially important implications of the rating instrument itself for training, Denisi, Cafferty, and Meglino (1984) and Feldman (1986) suggest that the nature of the scales dictates what types of performance information raters must acquire. Pulakos (1986) suggests that efforts to increase accuracy should not proceed independent of these rating task demands. Along this line, studies by McIntyre, Smith, and Hassett (1984) and by Pulakos (1984) imply that training raters to search for and collect relevant information in a manner that is consistent with the rating instrument does seem to facilitate accuracy.

According to Pulakos (1986), although no one scale type stands alone as "best", certain combinations of rating format and training can be associated with higher quality performance evaluations than other such combinations. Pulakos (1986) found a significant format x training interaction in a study where training was developed that oriented raters to collect appropriate performance information and use it in a manner that would facilitate accurate ratings on a rating format designed to match the rating task. These results provide indirect support for the principle of encoding specificity with rater training and rating format taking on the roles of encoding and retrieval cues respectively.

In sum, while different appraisal purposes may result in raters focusing on and encoding different types of information, the manner in which raters are required to evaluate ratees may act as a type of retrieval cue. Thus, a judgment scale organized by person should reactivate a person schema while a scale organized by performance dimension should reactivate a script schema.

Overview

This study was an attempt to show the importance of schemata when generalizing the principle of encoding specificity to a complex social interaction, namely performance appraisal. In effect, schematic processing was

influenced in order to create conditions of consistency or inconsistency between encoding and retrieval cues. The prediction of major importance was that when cues are successfully manipulated to be consistent, performance appraisal accuracy will be significantly greater than when cues are inconsistent.

Specifically, the encoding conditions in this study were developed so as to instill in raters an appropriate or inappropriate (as the case may be) cognitive orientation for accurately performing either a behavior oriented or a person oriented rating task. For example, presenting raters with the purpose of paying attention to each person in a group of ratees (encoding cue), would require collecting ratings with a person-categorized rating format (retrieval cue) in order to make comparisons concerning differences between ratees. Likewise, inducing a script that includes a set of expectations for appropriate sequence of ratee behaviors upon raters would require that ratings be collected using a behavior-categorized format (retrieval cue) in order to make the necessary evaluation of the ratees' strengths and weaknesses.

Presenting raters with the purpose of paying attention to each person in the group and using a behavior format would constitute an inconsistent encoding-retrieval match. Likewise, presenting raters with a behavior script and using a person rating format would constitute an inconsistent

encoding-retrieval match. In summary, consistent encoding (purpose of ratings) and retrieval cues (type of rating format) were expected to result in the most efficient retrieval of information, and consequently more accurate performance ratings. Dependent measures used in this study included: free recall (measure of memory organization), behavior recognition (measure of accuracy and false positive rate), judgment (measure of elevation, differential elevation, stereotypical accuracy, differential accuracy, and overall accuracy), and a manipulation check. These measures will be defined in more detail in the methods section.

Hypotheses

The present study used three types of performance appraisal purpose (encoding cues). They included an impression formation condition, a behavior memory condition, and a control condition which was a combination of impression formation and behavior memory. Also, two rating formats (retrieval cues) were used, person categorization and behavior categorization. The following hypotheses were generated:

Hypothesis 1: On the free recall measure, individuals in the impression formation condition will organize information by person to a greater degree than individuals in both the behavior memory and control conditions. Moreover,

individuals in the behavior memory condition will organize information by behavior to a greater degree than individuals in both the impression and control conditions. Finally, individuals in the control condition should organize information by person to a smaller degree than subjects in the impression formation condition but to a greater degree than individuals in the behavior memory condition. Likewise, they should organize information by behavior to a smaller degree than individuals in the behavior memory condition but to a greater degree than individuals in the impression formation condition.

Hypothesis 2: On the recognition measure, individuals in the impression formation condition using the person-categorized format will exhibit greater recognition accuracy and fewer false positives than individuals in both the behavior memory and the control conditions. Moreover, individuals in the behavior memory condition using the behavior-categorized format should exhibit greater recognition accuracy and fewer false positives than individuals in both the impression formation and the control conditions. Finally, individuals in the control condition using the person-categorized format should exhibit less recognition accuracy and more false positives than individuals in the impression formation condition but more recognition accuracy and fewer false positives than

individuals in the behavior memory condition. Likewise, individuals in the control condition using the behavior-categorized format should exhibit less recognition accuracy and more false positives than individuals in the behavior memory condition but more recognition accuracy and fewer false positives than individuals in the impression condition.

Hypothesis 3: On the judgment measure, individuals in the impression formation condition using the person-categorized format measure will exhibit greater judgment accuracy than individuals in both the behavior memory and control conditions. Moreover, individuals in the behavior condition using the dimension-categorized format should exhibit greater judgment accuracy than individuals in both the impression formation and control conditions. Finally, individuals in the control condition using the person-categorized format should exhibit less judgment accuracy than individuals in the impression formation condition but greater judgment accuracy than individuals in the behavior memory condition. Likewise, individuals in the control condition using the dimension-categorized format will exhibit less judgment accuracy than individuals in the behavior memory condition, but greater judgment accuracy than individuals in the impression formation condition.

Method

Subjects

Subjects were 108 undergraduate students from a large, southeastern university. They were randomly assigned to one of six experimental conditions with the stipulation that both males and females were evenly distributed across conditions.

Design

The study consisted of a 3 (rating purpose) x 2 (rating format) between groups factorial design.

Stimulus Materials

A 20-minute videotape served as the stimulus material. The videotape contained actual footage of 3 grocery baggers at work in a local supermarket. The videotape was edited to achieve two objectives. First, since the baggers were not always in the checkout area due to other duties, the tape was edited so that each bagger was on the screen for approximately the same amount of time (about 12 minutes). Second, the tape contained approximately the same number of prototypically good and prototypically poor behaviors for each bagger. The tape was approximately 20 minutes in length to represent the full range of possible behaviors on each performance dimension that defines a grocery bagger. A training manual issued by the store at which the videotape was filmed was consulted to determine the prototypically good and poor bagger behaviors.

Obtaining True Scores

Performance Evaluation. Borman (1977) noted that the accuracy of ratings can be determined by comparing each subject's ratings to "true scores," which represent the expected value of the rating obtained from an expert who is evaluating ratee behavior under optimal rating conditions. True scores were developed for each ratee and performance dimension by the author and a research assistant both of whom were very familiar with the performance demands of the job of grocery bagger. Furthermore, both raters became thoroughly familiar with each ratee's performance by watching the videotape several times and taking notes before the actual rating sessions. The rating sessions themselves, which involved watching the entire videotape and focusing on one performance dimension per bagger at a time (for a total of 15 actual rating sessions), provided ample opportunity to stop the videotape and review notes when necessary. It was hoped that maximum opportunity to review relevant performance-related behavior would lead to highly informed true score ratings. Agreement was virtually unanimous and rating never differed by more than 1 point in which case the mean of the two scores was adopted as the "true score". These scores were used to subsequently assess raters' elevation, differential elevation, stereotypical accuracy, and differential accuracy scores.

Recognition. True scores for behavior frequency were determined in a manner similar to that of the evaluation true scores. Again, the raters watched the entire videotape focusing on each present behavior per bagger at a time. These viewings plus one viewing to assure that all nonpresent events were in fact not present added up to 31 rating sessions for recognition true scores. Again agreement was virtually unanimous. These scores were used to later determine recognition accuracy and false positive rates.

Procedure

Subjects reported in groups of 3-5 for testing. First, they were asked to read and sign a consent form (see Appendix A). Next, all subjects were given verbal instructions containing a short script that described the proper technique for grocery bagging (see Appendix B), a list of five performance dimensions relevant to the job of grocery bagger (see Appendix B), and one of the three purposes for viewing the videotape. After viewing the videotape, subjects completed a short filler task, the Picture Number Test MA-1 (Ekstrom, French, Harmon, & Derman, 1976), to eliminate the effects of short-term memory. They then completed measures of free recall, recognition for bagger behaviors, judgment of bagger performance, and a manipulation check. Subjects were then be debriefed concerning the true purpose of the study.

Manipulations

Rating purpose and rating format were manipulated in a 3 x 2 factorial design.

Purpose of ratings. This factor was manipulated by verbal instructions given to the subjects prior to viewing the videotape. In the person condition, subjects were told "to try and form an impression of each of the grocery baggers." In the behavior condition, subjects were told to " try and remember as many specific behaviors as possible, concentrating on the baggers, not the customers or cashiers." In the control condition, subjects were told to "try and remember as many specific behaviors as possible in order to form an impression of each bagger."

Rating format. In both rating conditions, subjects were given two-part forms. The first part was a recognition measure (see Appendices C & D) while the second part was a judgment measure (see Appendices E & F). In the person-organized condition, both measures were organized by bagger. For example, subjects were required to complete behavior recognition questions (see Appendix C) and performance dimension ratings (see Appendix E) about one bagger on all behaviors and dimensions before rating the next bagger, and so on. These recognition and rating formats were expected to cue the retrieval of information by person.

In the behavior-organized condition, the measures were organized by behavior or by dimension. For instance, subjects were required to complete behavior recognition questions (see Appendix D) and performance dimension ratings (see Appendix F) for all 3 baggers before rating them on the next behavior or dimension, and so on. These recognition and rating formats were expected to cue the retrieval of information by behavior.

Summary of Encoding-Retrieval Combinations

The following combinations of encoding and retrieval cues were considered to be consistent and inconsistent patterns for encoding specificity.

Purpose	Format	Match Type
Impression Formation	Person	Consistent
Behavior Memory	Behavior	Consistent
Impression Formation	Behavior	Inconsistent
Behavior Memory	Person	Inconsistent

Dependent Measures

Free recall. The effects of encoding cues (appraisal purpose) on memory organization were inferred from responses on the free recall measure. Free recall was assessed by asking subjects to write everything they could remember from the videotape on a blank sheet of paper (see Appendix G). The measure of subjective organization used to score the free recall measure was the Adjusted Ratio of Clustering (ARC)

index (Roenker, Thompson, and Brown, 1971; Murphy and Puff, 1982) which indicates the degree of clustering in the recalled items by person and by behavior. The ARC score represents the proportion of actual category repetitions above chance to the total possible category repetitions above chance for any given recall protocol. Chance clustering is set at zero and perfect clustering at 1.00. The computational formula for the ARC score is as follows (Murphy & Puff, 1982, p. 120):

Terms

N = number of items recalled

r = number of category repetitions

c = number of different categories represented in recall

n_i = number of items recalled in the ith category

Max = maximum value of r for an output

E(r) = expected value of r for an output

Definitions

$$Max = N-c$$

$$E(r) = \frac{\sum n_i^2}{N} - 1$$

Measure

$$ARC = [r - E(r)]/[Max - E(r)]$$

Separate calculations were made using persons and behaviors as the categories of anticipated clustering.

Recognition Memory Test. Lord (1985) supported the use of recognition memory tasks in studying cognitive processes in performance appraisal. This approach makes it possible to employ the methods of Signal Detection theory in measuring accuracy (Swets & Pickett, 1982). Signal detection theory procedures were originally developed in psychophysics as a means for discriminating between (a) changes in subjects' sensitivity to a signal presented in a field of background noise, and (b) changes in their response criterion for reporting the signal's presence.

In applying signal detection theory to the present problem, it is assumed that a subject's report about whether a bagger engaged in a particular behavior involves two processes: (a) an attempt to retrieve stored evidence from memory concerning whether or not the bagger performed the behavior, and (b) the establishment of decision criteria to be used in weighing the retrieved evidence and responding to the questionnaire item. Signal detection theory views these two processes as separate and independent. In short, encoding cues may increase a subject's tendency to report that particular types of behavior were enacted by the bagger either because (a) the cues selectively increase the availability in the subject's memory of evidence concerning the occurrence of the behaviors (via either selective encoding or selective retrieval), or (b) the cues cause the

subject to lower his/her evidence criterion for concluding that such behaviors were enacted by the bagger (i.e., a general response bias process) (Larson, Lingle, and Scerbo, 1984).

The signal detection theory procedures used were based on an analysis of each subject's Memory Operating Characteristic (MOC) curve (Swets, 1973). An MOC curve is a theoretical curve derived from a plot of a subject's hit rate (i.e., the rate at which behaviors actually enacted by the bagger are in fact reported as having been enacted by him) against his/her false alarm rate (i.e., the rate at which behaviors not enacted by the bagger are reported as having been enacted by him).

MOC curves can be estimated from single administrations of recognition memory tests that are in the form of rating scales. Both parametric and nonparametric indices of memory sensitivity and response bias have been developed for use with such questionnaires. Of these, the parametric indices are potentially more sensitive, however, they require strong assumptions of normality and homogeneity of variance (Pastore & Scheirer, 1974).

In order to develop the recognition memory test for this study, the author and a research assistant recorded critical incidents from the videotape. Ten incidents that had occurred on the tape [5 prototypical (good) and 5 antiprototypical

(bad)], and an additional ten behavioral incidents (5 prototypical and 5 antiprototypical) that were judged to be plausible, but that had not occurred on the tapes were generated.

This set of 20 items was randomly intermixed and used to form the Recognition Memory test. Subjects' task was to indicate how frequently (if at all) each incident had occurred (on a 5-point Likert scale) on the tape they had seen. Since the true status of each item was known, it was possible to compute measures of both response bias and sensitivity based on Signal Detection Theory (see Appendix I for computations from Grier, 1971).

A measure of sensitivity is especially useful since it is independent of any response bias that might be present, whereas a simple raw count of true positives (hits), and false positives (false alarms) could be strongly affected by general response tendencies. Both parametric and nonparametric memory sensitivity measures are independent of response bias. However, the parametric measures are much more stringent in their assumptions. Therefore, a commonly used nonparametric index of memory sensitivity was employed in the present research. It is the area under the MOC curve, known as A'. A' can range from .5 to 1.0 with the two extremes reflecting chance and perfect recognition accuracy, respectively.

The parametric measure of response bias known as false positive rate was employed in this study. It was used instead of its nonparametric equivalent, B''. B'' tends to be undefined or unstable in situations with perfect hit rates. This data did contain perfect hits, and in order to avoid instability of the response bias measure, the parametric version was chosen for use. Furthermore, only parametric measures of response bias, guarantee a response bias index that is completely independent of memory sensitivity levels (Larson, Lingle, and Scerbo, 1984).

Performance Evaluation Test. The second rating scale asked participants to evaluate each lecturer's performance on 5 different performance dimensions selected from existing teacher evaluation forms. Participants used a Likert-type scale ranging from very poor (1) to very good (5) to rate each performance dimension.

When raters evaluate a number of ratees on multiple performance dimensions, each rater's Overall Accuracy can be rewritten as the sum of four components: (a) accuracy in the overall level of rating (Elevation), (b) accuracy in discriminating among ratees (Differential Elevation), (c) accuracy in discriminating among dimensions of performance (Stereotype Accuracy), and (d) accuracy in discriminating among ratees within each performance dimension (Differential Accuracy) Borman, 1977; Cronbach, 1955). Formulas for these accuracy measures are listed in Appendix J from Cronbach, 1955).

Differential Elevation, Stereotype Accuracy, and Differential Accuracy all reflect accuracy in making discriminations among ratees, performance dimensions, or both, and therefore may all be affected by bias in the observation, encoding and retrieval of information about ratee behavior. For example, if one ratee is better than another in performing most but not all aspects of the job, memory distortion may lead raters to give that ratee higher ratings on all dimensions (differential elevation). Elevation, on the other hand, is not likely to be systematically affected by the purpose of observation or by memory biases (Murphy & Balzer, 1986) and is conceptually similar to leniency/severity bias.

Experimental manipulation checks. A manipulation check for each purpose factor was included on a final questionnaire (see Appendix H). Subjects' perceptions of the extent to which they attempted to form an impression of the ratees and the extent to which they attempted to remember specific behaviors were each be assessed by a 5-point Likert scale.

Results

Manipulation Effectiveness

The two items designed to assess predicted experimental manipulation effects were presented to each subject in one questionnaire. Each item was subjected to a one-way Analysis of Variance (ANOVA) with viewing purpose designated as the

independent variable. Summary tables for these analyses are presented in Tables 1 & 2.

Impression Purpose. Table 1 summarizes the ANOVA and also indicates the mean response on the impression formation manipulation check for the three purpose conditions.

The expected main effect of viewing purpose emerged for this manipulation. Individuals in the impression purpose condition reported trying to form impressions to a greater degree than did those assigned to the behavior purpose condition and to those assigned to the control condition (see Table 1).

Insert Table 1 about here

One-tailed t-tests revealed that group means for all purpose groups were significantly different each other (see Table 1). In general, it appears that the impression purpose had the desired effect on subordinates.

Behavior Purpose. Table 2 summarizes the ANOVA and also indicates the mean response on the behavior memory manipulation check for the three purpose conditions.

The expected main effect of viewing purpose emerged for this manipulation. Individuals in the behavior purpose condition reported trying to remember specific behaviors to a greater degree than did those assigned to the impression

purpose condition and to those assigned to the control condition (see Table 2).

Insert Table 2 about here

One-tailed t-tests revealed that the group mean for the behavior group was significantly different from that of the impression group (see Table 2). The group mean for the control group was significantly different from the behavior memory group mean but not significantly different from the impression group mean. In general, it appears that the behavior purpose manipulation had the desired effect on subordinates.

Free Recall Measure

Both ARC indices (one for clustering by person and one for clustering by behavior) were subjected to 3 (purpose) x 2 (format) ANOVAs to examine the differences between groups on recall clustering scores. Viewing purpose was the independent variable of interest. Hypothesis 1 predicted that individuals in the impression purpose condition would organize information more by person on the free recall measure, while individuals in the behavior purpose condition would organize information more by behavior and individuals in the control condition would fall somewhere in between the other two conditions. These hypotheses were supported for

both person (see Table 3) and behavior (see Table 4) clustering.

Insert Tables 3 and 4 about here

One-tailed t-tests indicated that individuals' mean person clustering scores in the impression condition were significantly greater than person clustering scores in the behavior condition. Individuals' person clustering scores in the control condition fell in between and were significantly different than both of the other two groups (see Table 3).

One-tailed t-tests also revealed that individuals' mean behavior clustering scores in the behavior condition were greater than behavior clustering scores in the person condition. Moreover, individuals' clustering scores in the control condition were significantly different from those in the impression condition but were not significantly different than those in the behavior memory group (see Table 4).

Finally, two-tailed t-tests revealed that for individuals in the control group mean behavior clustering scores and mean person clustering scores were significantly different, $\underline{t}(35)=.95$, $\underline{p}<.001$. So, the control subjects clustered information more by person. This is supported by the schema literature which says that people naturally cluster people-related information by person (Fiske & Taylor, 1984).

Recognition Memory

Hypothesis 2 predicted that individuals in the impression purpose would exhibit greater recognition accuracy and fewer false positives on the person-categorized recognition measure than individuals in both the behavior and the control conditions with individuals in the control condition falling somewhere between the two. Furthermore, individuals in the behavior purpose condition were predicted to exhibit greater recognition accuracy and fewer false positives on the behavior-categorized recognition measure than individuals in both the impression and the control conditions, with individuals in the control condition falling somewhere between the two.

Recognition scores were transformed into sensitivity and false positive rate values (see Appendix I). Each of these scores was subjected to a 3 (purpose) x 2 (format) x 3 (ratee) x 2 (item prototypicality) ANOVA. As with previous analyses, the between subjects factors were purpose and format. Within subjects factors were ratee and item prototypicality.

Accuracy/sensitivity rate. The relevant ANOVA summary and descriptive statistics found in Tables 5 and 6 indicate that no significant purpose x format interaction emerged for rater sensitivity.

Insert Tables 5 and 6 about here

However, a significant bagger by prototype interaction emerged for recognition accuracy (see Figure 1). Ratings were more accurate on antiprototypical items for Baggers A and B (the better baggers) than for Bagger C (the worst bagger). Likewise, ratings were more accurate on prototypical items for Bagger C than for Baggers A and B.

Finally, a significant format x prototype interaction emerged for recognition accuracy. Basically, for those subjects given person-categorized forms, recognition accuracy was significantly better for antiprototypical items than for prototypical items. However, recognition accuracy was not significantly different for subjects given behavior-categorized rating forms.

Insert Figure 1 about here

False Positive Rate. The ANOVA summary and descriptive statistics found in Tables 7 and 8 indicate that no significant purpose x format interaction emerged for false positive rate. However, an unpredicted format effect did emerge for false positive rate. The false positive rate for person format was greater than the false positive rate for

behavior format. This indicates that raters were more likely to say that a nonpresent event occurred when they were given a rating format organized by person than when they were given a rating format organized by behavior. Interestingly, the difference between false positive rate as a function of rating format was somewhat greater for Bagger A than for Baggers B and C. This resulted in a format x bagger interaction but is more parsimoniously interpreted as a format effect.

Insert Tables 7 and 8 about here

Finally, a bagger by prototypicality interaction emerged. For Baggers A and B, the false positive rate was much lower for antiprototypical items than for prototypical items. However, for Bagger C (the worst bagger) there was little difference between the false positive rates for antiprototypical and prototypical items.

Judgment Accuracy

Hypothesis 3 predicted that individuals in the impression formation group using the person-categorized judgment measure would exhibit more judgment accuracy than individuals in the behavior condition with individuals in the control condition falling somewhere between the two.

Likewise, it was predicted that individuals in the behavior purpose condition using the dimension-categorized judgment measure would exhibit more judgment accuracy than individuals in the impression condition, with individuals in the control condition falling somewhere between the two.

Judgment scores were transformed into elevation, differential elevation, stereotypical accuracy, differential accuracy, and overall accuracy values (see Appendix J). Each of these five judgment scores was subjected to a 3 (purpose) x 2 (format) ANOVA. No purpose x format interaction emerged for any of the judgment measures (see Tables 9-13).

However, an unpredicted format effect did emerge for elevation. As can be seen in Table 9, individuals who received rating formats organized by person were less susceptible to elevation than individuals who received rating formats organized by behavior. In other words, raters who used person-categorized rating forms had mean ratings across all ratees and dimensions that were closer to the true mean rating across all ratees and dimensions than raters who used dimension-organized rating forms.

Insert Tables 9-13 about here

To summarize the results of this study, subjects asked to form impressions of ratees organized information by person

in free recall while subjects asked to remember behaviors organized information by behavior. No predicted purpose x format interactions emerged for either recognition or judgment accuracy.

However, several interesting findings did emerge. Two format effects, one for recognition accuracy and one for judgment accuracy, were found. In essence, for the judgment measure, people who received person categorized rating forms exhibited more judgment accuracy than those who received dimension-categorized rating forms. On the other hand, for the recognition measure, people who received behavior-categorized rating forms exhibited more recognition accuracy than those who received person-categorized rating forms. Finally, within subjects analysis revealed a significant bagger by prototype interaction for recognition accuracy. Specifically, for Baggers A and B (the good baggers) recognition accuracy was greater for antiprototypical items than for Bagger C (the bad bagger). Likewise, for Bagger C, recognition accuracy was greater for prototypical items than for Baggers A and B.

Discussion

Through the principle of encoding specificity, the present study examined the relation between certain combinations of pre- and post-observational cues and their

effects on information categorization, recognition, and performance appraisal accuracy. Results provided clear support for hypothesis 1. However, no support was found for hypotheses 2 and 3. Each of these hypotheses will be discussed below.

Hypothesis 1 predicted that individuals in the impression formation condition would organize information more by person on the free recall measure, while individuals in the behavior memory condition would organize information more by behavior, with individuals in the control condition falling in between the two. Complete support was found for this hypothesis suggesting that performance appraisal purpose can significantly impact on the way raters organize information. This finding was no surprise and supports previous research with similar findings (e.g., Foti & Lord, 1987).

Hypothesis 2 predicted that consistent encoding (purpose) and retrieval (rating format) cues would result in greater recognition accuracy. This hypothesis was not supported. These findings are inconsistent with the findings of prior encoding specificity investigations (e.g., Schulman, 1971) indicating that in this study the hypothesized definition of consistent cues and their effects might be inaccurate.

Data analysis did reveal a format effect for recognition accuracy. Specifically, raters who used recognition measures organized by behavior produced fewer false positives than raters whose forms were organized by person. This supports previous investigations of retrieval effects (e.g., Winograd & Conn, 1971; Light, 1972; Tulving & Psotka, 1971) and suggests the need for future research looking at category effects of behavioral recognition measures used in performance appraisal situations.

Finally, a significant within subjects bagger by item prototypicality interaction emerged for recognition accuracy. Recognition accuracy was lower on the two good baggers for prototypical items. Likewise, accuracy was lower on the worst bagger for antiprototypical items. This supports the incongruity biased-encoding model (Hastie & Park, 1986) which says that accuracy is greater for items that are inconsistent because information that is incongruent (contradictory) is given special processing elaborated through attributions that enhances its memorability. This special processing is most likely due to the subject's effort to explain why the surprising act was performed by the ratee. In other words the subject tries to make an attribution as to the cause of this behavior. Moreover, when the subject's memory is tested, it is especially likely that he/she will find incongruent information when searching long-

term memory. This seems especially likely since the raters had ample time to process the incongruities they were viewing (Sentis & Burnstein, 1979). In this case, prototypical behaviors are inconsistent for the poor bagger while antiprototypical behaviors are inconsistent for the good baggers.

Hypothesis 3 predicted that consistent encoding (purpose) and retrieval (rating format) cues would result in greater judgment accuracy. This hypothesis was not supported, again indicating that the definition of consistency and the actual measures used in this study to test encoding-specificity hypotheses should be thoroughly investigated for flaws as opposed to deserting the idea of applying this theory to performance appraisal situations.

Data analysis did reveal a rating format effect for judgment accuracy. In effect, individuals who received rating forms organized by person were less susceptible to elevation. In other words, they were less likely to exhibit leniency/severity error than individuals who received rating forms organized by dimension. This finding, like the format effect for recognition accuracy, supports previous investigations of retrieval effects (e.g., Winograd & Conn, 1971; Light, 1972; Tulving & Psotka, 1971).

The recognition format effect and judgment format effect suggest a 2-way interaction that should be investigated in

future research endeavors. In essence, studying the interactive effects of type of rating (recognition vs. judgment) and format organization (behavior vs. person) might prove to be a more appropriate way to apply the idea of encoding specificity to the problem of performance appraisal accuracy. For example, people asked to complete a judgment measure such as a rating scale might be more accurate using a person-categorized rating form. This information would only serve to reinforce the categorization of currently used judgment measures. However, people who are asked to complete a behavior recognition measure such as a Behavioral Observation Scale (BOS), which asks raters to consider how frequently the ratee has been observed behaving in a particular manner, might be more accurate using a behaviorally-organized format. Presently, raters using the BOS consider each ratee separately on all behavioral dimensions before moving to the next ratee. Evidence from this study suggests that this procedure be changed.

Summary and Conclusions

To summarize, the present study aimed to shed more light on the information processes involved in performance appraisal situations. The principle of encoding specificity (Tulving & Thomson, 1973) was applied to an actual performance appraisal situation.

Three important findings emerged. First, subjects clearly organized performance information differently as a result of appraisal purpose.

Second, and particularly relevant to the performance appraisal literature, categorization of rating format was found to affect both recognition and judgment accuracy. A behaviorally-organized format resulted in greater recognition accuracy whereas a person-categorized format resulted in greater judgment accuracy. Since the accuracy measures used for the recognition and judgment forms are not statistically comparable, no direct comparison was made to test the evidence of an interaction. Computation of Cronbach's (1955) accuracy scores for the recognition measure is advised in order that the appropriate comparisons might be made.

Third, an item prototypicality x ratee interaction emerged for recognition accuracy, supporting the incongruity biased-encoding model (Hastie & Park, 1986) which states that accuracy is greater for items that are incongruent with the overall image or schema of an individual.

Finally, the lack of evidence for the two proposed encoding specificity hypotheses (Hypotheses 2 and 3) suggests the need for further investigation of encoding-specificity theory as it applies to performance appraisal accuracy.

References

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. Journal of Verbal Learning and Verbal Behavior, 17, 1-12.

Asch, S. E. (1969). Reformulation of the problem of association. American Psychologist, 24, 92-102.

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.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. Cognitive Psychology, 11, 177-220.

Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. Journal of Personality and Social Psychology, 41, 656-670.

Cohen, C. E. (1981). Goals and schemata in person perception. In N. Cantor & J. F. Kihlstrom (Eds.), Personality, Cognition, & Social Interaction. Hillsdale, N. J.: Lawrence Erlbaum Associates.

Craik, F. I. M. (1973). A "levels of analysis" view of memory. In P. Pliner, L. Krames, & T. M. Alloway (Eds.), Communication and Affect: Language and Thought. New York: Academic Press.

Craik, F. I. M. & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 1972, 11, 671-684.

Crocker, J., Binns, D., & Weber, R. (1983). Person memory and causal attributions. Journal of Personality and Social Psychology, 44, 55-66.

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

DaPolito, F., Barker, D., & Wiant, J. (1972). The effects of contextual changes on the component recognition. American Journal of Psychology, 85, 431-440.

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

Earhard, M. (1969). Storage and retrieval of words encoded in memory. Journal of Experimental Psychology, 80, 412-418.

Ekstrom, R. B., French, J. W., Harmon, H. H., & Derman, D. Manual for kit of factor-referenced cognitive tests. Princeton, New Jersey: Educational Testing Service, 1976.

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

Feldman, J. M. (1986). Instrumentation and training for performance appraisal: A perceptual-cognitive viewpoint. In K. M Rowland and J. R. Ferris (Eds.), Research in Personnel and Human Resource Management (Vol. 4). Greenwich, CT: JAI.

Fiske, S.T. & Linville, P.W. (1980). What does the schema concept buy us? Personality and Social Psychology Bulletin, 6(4), 543-557.

Fiske, S. T. & Taylor, S. E. (1984). Social Cognition. Reading, Massachusetts: Addison-Wesley Publishing Co.

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.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. Psychological Bulletin, 75(6), 424-429.

Hamilton, D. L. (1981). Cognitive Representations of Persons. In E. T. Higgins, C. P. Herman, & M. P. Zanna (Eds.), Social Cognition: The Ontario Symposium (Vol. 1). Hillsdale, N.J.: Erlbaum.

Hastie, R. (1984). Causes and effects of causal attribution. Journal of Personality and Social Psychology, 46(1), 44-56.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. Psychological Review, 93(3), 258-268.

Landy, F. & Farr, J. (1980). Performance rating. Psychological Bulletin, 87(1), 72-107.

Light, L. L. (1972). Homonyms and synonyms as retrieval cues. Journal of Experimental Psychology, 96, 255-262.

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(1), 147-156.

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

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

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

Roenker, D. L., Thompson, C. P. & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. Psychological Bulletin, 76(1), 45-48.

Schank, R. C. & Abelson, R. P. (1977). Scripts, Plans, Goals and Understanding. Hillsdale, New Jersey: Lawrence Erlbaum.

Schulman, A. J. (1971). Recognition memory for targets from a scanned word list. British Journal of Psychology, 62, 335-346.

Sentis, K. P., & Burnstein, E. (1979). Remembering schema consistent information: Effects of a balance schema on recognition memory. Journal of Personality and Social Psychology, 37, 2200-2211.

Swets, J. A. (1986). Indices of discrimination and diagnostic accuracy: Their ROCs and implied models. Psychological Bulletin, 99, 100-117.

Tulving, E. & Osler, S. (1968). Effectiveness of retrieval cues in memory for words. Journal of Experimental Psychology, 77, 593-601.

Tulving, E. & Psotka, J. (1971). Retroactive inhibition in free recall: Inaccessibility of information available in the memory store. Journal of Experimental Psychology, 87, 1-8.

Tulving, E. & Thomson, D. M. (1971). Retrieval processes in recognition memory: Effects of associative context. Journal of Experimental Psychology, 87, 116-124.

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80(5), 352-373.

Williams, K. J., Denisi, A. S., Blencoe, A. G. and Cafferty, T. P.. (1985). The role of appraisal purpose: Effects of purpose on information acquisition and utilization. Organizational Behavior and Human Decision Processes, 35, 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(2) 189-195.

Winograd, E. & Conn, C. P. (1971). Evidence from recognition memory for specific encoding of unmodified homographs. Journal of Verbal Learning and Verbal Behavior, 10, 702-706.

Table 1

Summary of ANOVA for Impression Purpose Manipulation Check as a Function of Purpose and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	26.76	2	26.39	.001
Format (F)	.05	1	.10	.758
P x F	1.60	2	1.58	.212
Error Between	51.71	102		

Impression Purpose	4.19 (.71)	a
Behavior Purpose	3.03 (.81)	b
Control Purpose	3.92 (.60)	c

Note: N = 108. Standard deviations are in parentheses after the means. Cell sizes are equal (36). High numbers indicate a greater attempt to have formed an impression. Means with common letters do not differ at the $p = .05$ level of significance.

Table 2

Summary of ANOVA for Behavior Purpose Manipulation Check as a Function of Purpose and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	8.02	2	6.61	.01
Format (F)	.14	1	.23	.64
P x F	.12	2	.10	.90
Error Between	62.82	102		

Impression Purpose	3.39 (.84)	a
Behavior Purpose	4.06 (.67)	b
Control Purpose	3.69 (.79)	a

Note: N = 108. Standard deviations are in parentheses after the means. Cell sizes are equal (36). High numbers indicate a greater attempt to remember behaviors. Means with common letters do not differ at the p=.05 level of significance.

Table 3

Summary of ANOVA for Adjusted Ratio of Clustering (ARC) Index for Person Categories as a Function of Purpose and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	3.10	2	6.64	.01
Format (F)	.41	1	1.75	.19
P x F	.35	2	.76	.47
Error Between	23.82	102		

Impression Purpose	.78 (.47)	a
Behavior Purpose	.37 (.49)	b
Control Purpose	.59 (.49)	c

Note: N = 108. Standard deviations are in parentheses after the means. Cell sizes are equal (36). Values can range from 0 to 1 with high numbers indicating a greater degree of clustering by person. Means with common letters do not differ at the p = .05 level of significance.

Table 4

Summary of ANOVA for Adjusted Ration of Clustering (ARC) Index for Behavior Categories as a Function of Purpose and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	3.14	2	7.28	.001
Format (F)	.22	1	1.00	.320
P x F	.85	2	1.98	.143
Error Between	21.99	102		

Impression Purpose	.20 (.40)	a
Behavior Purpose	.61 (.49)	b
Control Purpose	.44 (.50)	b

Note: N = 108. Standard deviations are in parentheses after the means. Cell sizes are equal (36). Values can range from 0 to 1 with high numbers indicating a greater degree of clustering by behavior. Means with common letters do not differ at the p = .05 level of significance.

Table 5

Summary of ANOVA for Accuracy (Sensitivity) on Recognition Measure as a Function of Purpose, Format, Bagger, and Item Prototypicality.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	.01	2	.16	.854
Format (F)	.08	1	3.16	.079
P x F	.03	2	.52	.595
Error Between	2.45	102		
Bagger (B)	.34	2	17.14	.001
P x B	.01	4	.29	.885
F x B	.00	2	.09	.915
P x F x B	.02	4	.38	.819
Error Within	2.00	204		
Prototypicality (Pr)	.09	1	6.59	.01
P x Pr	.00	2	.11	.893
F x Pr	.06	1	3.99	.05
P x F x Pr	.03	2	.99	.376
Error Within	1.45	102		
B x Pr	.17	2	11.62	.001
P x B x Pr	.01	4	.32	.862
F x B x Pr	.02	2	1.54	.217
P x F x B x Pr	.01	4	.41	.803
Error Within	1.45	204		

Table 6

Cell Descriptive Statistics for Accuracy (Sensitivity) on Recognition Measure as a Function of Purpose, Format, Bagger, and Item Prototypicality.

	Impression Formation Condition		Behavior Memory Condition		Control Condition		
	Person Format (n=16)	Behavior Format (n=20)	Person Format (n=16)	Behavior Format (n=20)	Person Format (n=17)	Behavior Format (n=19)	$\bar{M}$
Bagger A							
Proto	.518	.609	.542	.583	.528	.572	.559
Aproto	.631	.618	.624	.621	.621	.612	.621
							.590
Bagger B							
Proto	.500	.588	.544	.535	.513	.558	.540
Aproto	.576	.562	.573	.561	.540	.580	.565
							.553
Bagger C							
Proto	.528	.567	.517	.533	.552	.562	.543
Aproto	.525	.545	.526	.524	.511	.531	.527
							.536
Means	.546	.582	.554	.560	.544	.569	

Note: N = 108. Higher numbers indicate a greater degree of recognition accuracy.

Table 7

Summary of ANOVA for False Positive Rate on Recognition Measure as a Function of Purpose, Format, Bagger, and Item Prototypicality.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	.06	2	.13	.874
Format (F)	2.06	1	8.56	.01
P x F	.94	2	1.95	.147
Error Between	24.57	102		
Bagger (B)	2.08	2	62.54	.001
P x B	.11	4	1.58	.180
F x B	.15	2	4.50	.01
P x F x B	.13	4	1.96	.102
Error Within	3.39	204		
Prototypicality (Pr)	4.66	1	40.54	.001
P x Pr	.01	2	.05	.952
F x Pr	.21	1	1.80	.183
P x F x Pr	.50	2	2.19	.118
Error Within	11.73	102		
B x Pr	1.57	2	33.05	.001
P x B x Pr	.22	4	2.32	.058
F x B x Pr	.11	2	2.32	.101
P x F x B x Pr	.21	4	2.20	.070
Error Within	4.83	204		

Table 8

Cell Descriptive Statistics for False Positive Rate on Recognition Measure as a Function of Purpose, Format, Bagger, and Item Prototypicality.

	Impression Formation Condition		Behavior Memory Condition		Control Condition		
	Person Format (n=16)	Behavior Format (n=20)	Person Format (n=16)	Behavior Format (n=20)	Person Format (n=17)	Behavior Format (n=19)	$\bar{M}$
Bagger A							
Proto	.725	.500	.537	.520	.682	.505	.578
Aproto	.348	.343	.321	.343	.336	.316	.335
							.454
Bagger B							
Proto	.760	.450	.615	.617	.755	.544	.624
Aproto	.473	.307	.411	.414	.437	.278	.387
							.506
Bagger C							
Proto	.771	.467	.573	.642	.686	.518	.610
Aproto	.583	.583	.583	.475	.755	.491	.578
							.594
Means	.610	.442	.507	.502	.609	.442	

Note: N = 108. Higher numbers indicate a greater false positive rate (less accuracy).

Table 9

Summary of ANOVA for Elevation of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	.18	2	.08	.92
Format (F)	4.53	1	4.04	.05
P x F	.59	2	.27	.77
Error Betwn	114.29	102		

Impression Condition		
Person Format	.54 (.73)	n=16
Behavior Format	1.16(1.33)	n=20
Behavior Memory Condition		
Person Format	.69 (.79)	n=16
Behavior Format	.99(1.30)	n=20
Control Condition		
Person Format	.62 (.80)	n=17
Behavior Format	.92(1.08)	n=19

Note: N = 108. Standard deviations are in parentheses after the means. Also, high numbers indicate lower judgment accuracy.

Table 10

Summary of ANOVA for Differential Elevation of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	1.05	2	1.55	.22
Format (F)	.00	1	.00	.97
P x F	.07	2	.10	.91
Error Betwn	34.63	102		

Impression Condition		
Person Format	.55 (.58)	n=16
Behavior Format	.55(.56)	n=20
Behavior Memory Condition		
Person Format	.49 (.50)	n=16
Behavior Format	.44(.45)	n=20
Control Condition		
Person Format	.67 (.48)	n=17
Behavior Format	.73(.83)	n=19

Note: N = 108. Standard deviations are in parentheses after the means. Also, high numbers indicate lower judgment accuracy.

Table 11

Summary of ANOVA for Stereotypical Accuracy of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	.06	2	.32	.70
Format (F)	.00	1	.00	.91
P x F	.04	2	.21	.80
Error Betwn	9.66	102		

Impression Condition		
Person Format	.52 (.27)	n=16
Behavior Format	.52(.22)	n=20
Behavior Memory Condition		
Person Format	.58 (.27)	n=16
Behavior Format	.53(.27)	n=20
Control Condition		
Person Format	.47 (.31)	n=17
Behavior Format	.52 (.44)	n=19

Note: N = 108. Standard deviations are in parentheses after the means. Also, high numbers indicate lower judgment accuracy.

Table 12

Summary of ANOVA for Differential Accuracy of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	.60	2	2.67	.07
Format (F)	.15	1	1.31	.26
P x F	.49	2	2.19	.12
Error Between	11.47	102		

Impression Condition		
Person Format	.38 (.30)	n=16
Behavior Format	.39 (.31)	n=20
Behavior Memory Condition		
Person Format	.52 (.28)	n=16
Behavior Format	.56 (.37)	n=20
Control Condition		
Person Format	.69 (.47)	n=17
Behavior Format	.43 (.25)	n=19

Note: N = 108. Standard deviations are in parentheses after the means. Also, high numbers indicate greater judgment accuracy.

Table 13

Summary of ANOVA for Overall Accuracy of Judgment Scores as a Function of Purpose and Format and Cell Descriptive Statistics.

Source of Variation	Sum of Squares	DF	F	Significance Of F
Purpose (P)	.66	2	.21	.81
Format (F)	3.20	1	2.05	.16
P x F	1.19	2	.38	.68
Error Betwn	158.80	102		

Impression Condition		
Person Format	1.99(1.15)	n=16
Behavior Format	2.63(1.29)	n=20
Behavior Memory Condition		
Person Format	2.28 (.91)	n=16
Behavior Format	2.52(1.33)	n=20
Control Condition		
Person Format	2.44(1.14)	n=17
Behavior Format	2.60(1.49)	n=19

Note: N = 108. Standard deviations are in parentheses after the means. Also, high numbers indicate lower judgment accuracy.

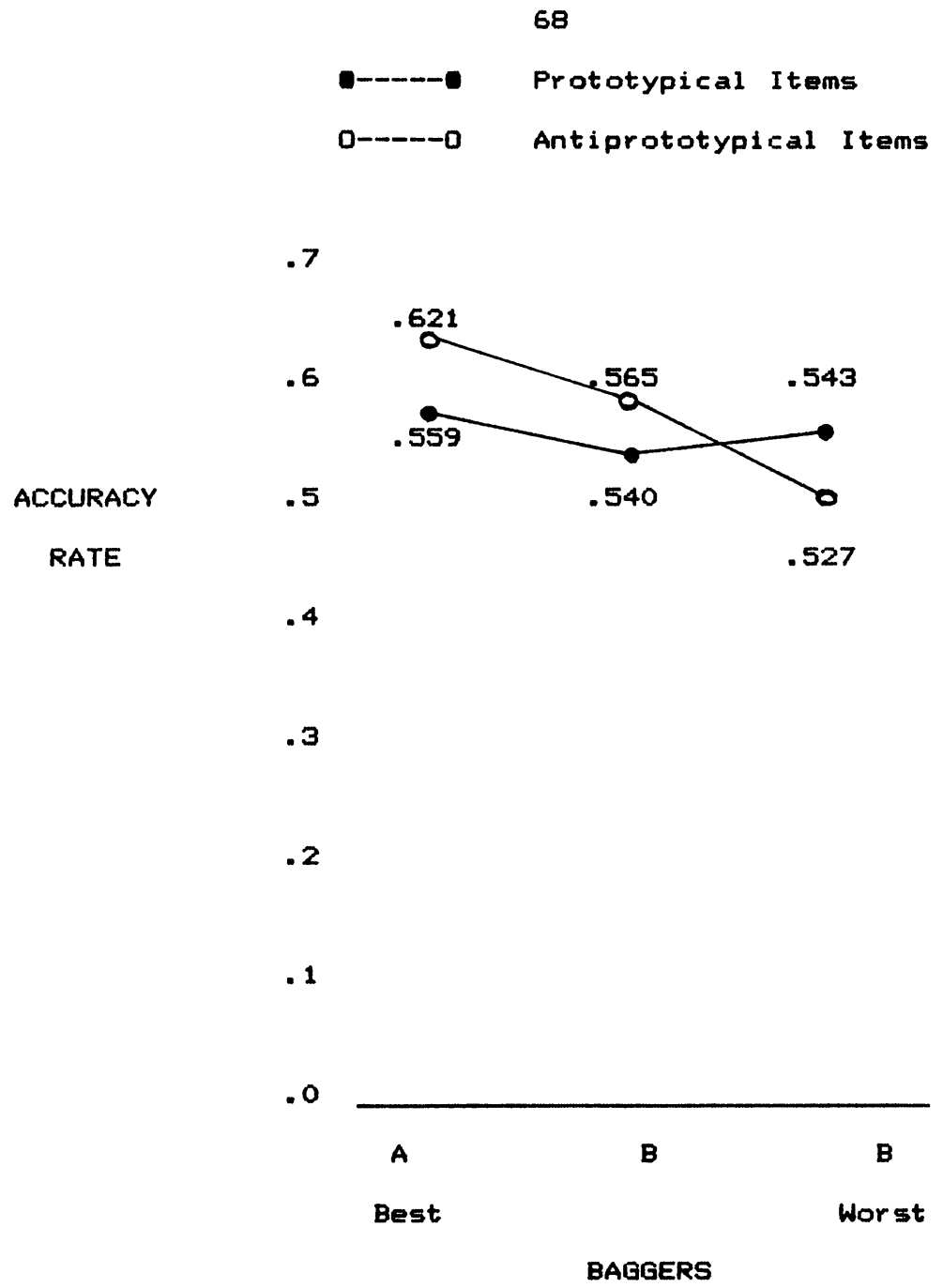

Figure 1.

Bagger x Item Prototypicality Interaction on Recognition Accuracy.

APPENDIX A

CONSENT FORM

INFORMED CONSENT

Project Title: Performance Appraisal Study

Description and Explanation of Procedure: You will be asked to watch a videotape and to complete questionnaires.

Your participation in this activity is STRICTLY VOLUNTARY. If you do not wish to participate, you will not be required to take part in the experiment. You may also discontinue your participation AT ANY TIME without suffering a penalty. The experiment will last approximately 1 hour, and you will receive one credit for your participation.

Except for this consent form, neither your name nor your student identification number will be associated with the survey materials. If you have any questions regarding any part of the evaluation, please feel free to contact Dr. Helen Crawford. She may be reached by stopping by room 5070C in the psychology department in Derring Hall, x6520. If you would like a copy of this consent form, simply ask the administrator and one will be provided.

This study has been approved by the Human Subjects Committee and the Institutional Review Board. Please direct any questions about the study to:

Marta Carter-Stuart
Experimenter

Helen Crawford
Chair, HSC

Chuck Waring
Chair, IRB

I have read and understand the requirements of the study and my rights as a study participant. I agree to participate on a voluntary basis.

_____ _____

NAME (PLEASE PRINT) SIGNATURE

_____ _____

STUDENT ID NUMBER DATE

APPENDIX B

SCRIPT AND PERFORMANCE DIMENSIONS

The Proper Technique For Grocery Bagging

First, the bagger should open a bag by sliding his hand inside and moving his hand back and forth until the bag is open. He should never just snap the bag open without putting his hand inside first. The bagger should use small plastic bags for frozen and cold foods. Also, the bagger should never pick up a full bag only by the top. Furthermore, the bagger should never wait for the customer to pick up his or her own bag.

The bagger should help out at other stations whenever possible. He should also clear the front area of small shopping baskets, and he should bring in shopping carts from outside. The bagger should never throw any food or objects in the air. Finally, the bagger should never engage in horseplay with another person.

Performance Dimensions Associated With Grocery Bagging And Behaviors That Underlie These Dimensions

1. BAGGING PROCEDURE
   opens bag correctly/incorrectly
   uses plastic bags for frozen food
   picks up full bag correctly/incorrectly

2. CUSTOMER RELATIONS
   hands bags directly to customer

3. PROFESSIONAL DEMEANOR
   throws objects
   engages in horseplay

4. PERFORMANCE OF "OTHER" DUTIES
   helps out at other stations
   clears front of small baskets
   brings in carts from outside

5. OVERALL PERFORMANCE

APPENDIX C

RECOGNITION MEASURE ORGANIZED BY PERSON

The following is a list of 20 behaviors which may or may not have occurred in the film you just observed. For each behavior, please circle a number on the corresponding line to indicate how frequently each bagger performed that behavior.

BAGGER #1 (long-sleeved white shirt)

1. Picked up full bags by the top.

1 2 3 4 5
NEVER ALWAYS

2. Combed hair.

1 2 3 4 5
NEVER ALWAYS

3. Opened bag by sliding hand inside to open it.

1 2 3 4 5
NEVER ALWAYS

4. Picked up full bags by side and bottom.

1 2 3 4 5
NEVER ALWAYS

5. Bagged poisonous items such as ammonia separately from other groceries.

1 2 3 4 5
NEVER ALWAYS

6. Helped out at other stations.

1 2 3 4 5
NEVER ALWAYS

7. Ran inside the store.

1 2 3 4 5
NEVER ALWAYS

8. Waited for customer to pick up his/her own bag.

1 2 3 4 5
NEVER ALWAYS

9. Opened bag by snapping it open.

1 2 3 4 5
NEVER ALWAYS

10. Threw objects in the air.

1 2 3 4 5
NEVER ALWAYS

11. Engaged in horseplay (with another person).

1 2 3 4 5
NEVER ALWAYS

12. Threw garbage on the floor.

1 2 3 4 5
NEVER ALWAYS

13. Used plastic bag for frozen/cold food.

1 2 3 4 5
NEVER ALWAYS

14. Tightened caps on poisonous items.

1 2 3 4 5
NEVER ALWAYS

15. Brought in shopping carts from outside.

1 2 3 4 5
NEVER ALWAYS

16. Cleared front area of small shopping baskets.

1 2 3 4 5
NEVER ALWAYS

17. Threw away bags that tore while being opened.

1 2 3 4 5
NEVER ALWAYS

18. Threw groceries on floor.

1 2 3 4 5
NEVER ALWAYS

19. Stepped into cashier's space.

1 2 3 4 5
NEVER ALWAYS

20. Bagged boxed cakes in a separate bag with the bottom down.

1 2 3 4 5
NEVER ALWAYS

BAGGER #2 (short-sleeved checked shirt)

1. Picked up full bags by the top.

1 2 3 4 5
NEVER ALWAYS

2. Combed hair.

1 2 3 4 5
NEVER ALWAYS

3. Opened bag by sliding hand inside to open it.

1 2 3 4 5
NEVER ALWAYS

4. Picked up full bags by side and bottom.

1 2 3 4 5
NEVER ALWAYS

5. Bagged poisonous items such as ammonia separately from other groceries.

1 2 3 4 5
NEVER ALWAYS

6. Helped out at other stations.

1 2 3 4 5
NEVER ALWAYS

7. Ran inside the store.

1 2 3 4 5
NEVER ALWAYS

8. Waited for customer to pick up his/her own bag.

1 2 3 4 5
NEVER ALWAYS

9. Opened bag by snapping it open.

1 2 3 4 5
NEVER ALWAYS

10. Threw objects in the air.

1 2 3 4 5
NEVER ALWAYS

11. Engaged in horseplay (with another person).

1 2 3 4 5
NEVER ALWAYS

12. Threw garbage on the floor.

1 2 3 4 5
NEVER ALWAYS

13. Used plastic bag for frozen/cold food.

1 2 3 4 5
NEVER ALWAYS

14. Tightened caps on poisonous items.

1 2 3 4 5
NEVER ALWAYS

15. Brought in shopping carts from outside.

1 2 3 4 5
NEVER ALWAYS

16. Cleared front area of small shopping baskets.

1 2 3 4 5
NEVER ALWAYS

17. Threw away bags that tore while being opened.

1 2 3 4 5
NEVER ALWAYS

18. Threw groceries on floor.

1 2 3 4 5
NEVER ALWAYS

19. Stepped into cashier's space.

1 2 3 4 5
NEVER ALWAYS

20. Bagged boxed cakes in a separate bag with the bottom down.

1 2 3 4 5
NEVER ALWAYS

BAGGER #3 (short-sleeved white shirt)

1. Picked up full bags by the top.

1 2 3 4 5
NEVER ALWAYS

2. Combed hair.

1 2 3 4 5
NEVER ALWAYS

3. Opened bag by sliding hand inside to open it.

1 2 3 4 5
NEVER ALWAYS

4. Picked up full bags by side and bottom.

1 2 3 4 5
NEVER ALWAYS

5. Bagged poisonous items such as ammonia separately from other groceries.

1 2 3 4 5
NEVER ALWAYS

6. Helped out at other stations.

1 2 3 4 5
NEVER ALWAYS

7. Ran inside the store.

1 2 3 4 5
NEVER ALWAYS

8. Waited for customer to pick up his/her own bag.

1 2 3 4 5
NEVER ALWAYS

9. Opened bag by snapping it open.

1 2 3 4 5
NEVER ALWAYS

10. Threw objects in the air.

1 2 3 4 5
NEVER ALWAYS

11. Engaged in horseplay (with another person).

1 2 3 4 5
NEVER ALWAYS

12. Threw garbage on the floor.

1 2 3 4 5
NEVER ALWAYS

13. Used plastic bag for frozen/cold food.

1 2 3 4 5
NEVER ALWAYS

14. Tightened caps on poisonous items.

1 2 3 4 5
NEVER ALWAYS

15. Brought in shopping carts from outside.

1 2 3 4 5
NEVER ALWAYS

16. Cleared front area of small shopping baskets.

1 2 3 4 5
NEVER ALWAYS

17. Threw away bags that tore while being opened.

1 2 3 4 5
NEVER ALWAYS

18. Threw groceries on floor.

1 2 3 4 5
NEVER ALWAYS

19. Stepped into cashier's space.

1 2 3 4 5
NEVER ALWAYS

20. Bagged boxed cakes in a separate bag with the bottom down.

1 2 3 4 5
NEVER ALWAYS

APPENDIX D

RECOGNITION MEASURE ORGANIZED BY BEHAVIOR

The following is a list of 20 behaviors which may or may not have occurred in the film you just observed. For each behavior, please circle a number on the corresponding line to indicate how frequently each bagger performed that behavior.

1. Picked up full bags by the top.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5

NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5

NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5

NEVER ALWAYS

2. Combed hair.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5

NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5

NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5

NEVER ALWAYS

3. Opened bag by sliding hand inside to open it.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5

NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5

NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5

NEVER ALWAYS

4. Picked up full bags by side and bottom.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

5. Bagged poisonous items such as ammonia separately from other groceries.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

6. Helped out at other stations.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

7. **Ran inside the store.**

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

8. **Waited for customer to pick up his/her own bag.**

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

9. **Opened bag by snapping it open.**

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

10. Threw objects in the air.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

11. Engaged in horseplay (with another person).

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

12. Threw garbage on the floor.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

13. Used plastic bag for frozen/cold food.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

14. Tightened caps on poisonous items.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

15. Brought in shopping carts from outside.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

16. Cleared front area of small shopping baskets.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

17. Threw away bags that tore while being opened.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

18. Threw groceries on floor.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

19. Stepped into cashier's space.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

20. Bagged boxed cakes in a separate bag with the bottom down.

1) Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

2) Bagger 2 (checked shirt)

1 2 3 4 5
NEVER ALWAYS

3) Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
NEVER ALWAYS

APPENDIX E

JUDGMENT MEASURE ORGANIZED BY PERSON

The following is a list of 5 performance dimensions. Please circle a number on the corresponding line to indicate your judgment of each bagger on that dimension.

BAGGER 1 (long-sleeved white shirt)

1. BAGGING PROCEDURE
(e.g., opened bag by sliding hand into it, used plastic bags for frozen foods, etc.)

1 2 3 4 5
VERY POOR VERY GOOD

2. CUSTOMER RELATIONS
(e.g., handed groceries directly to customer)

1 2 3 4 5
VERY POOR VERY GOOD

3. PROFESSIONAL DEMEANOR
(e.g., threw objects in air, engaged in horseplay, etc.)

1 2 3 4 5
VERY POOR VERY GOOD

4. PERFORMANCE OF "OTHER" DUTIES
(e.g., helped out at other stations, cleared front of small baskets, etc.)

1 2 3 4 5
VERY POOR VERY GOOD

5. OVERALL PERFORMANCE

1 2 3 4 5
VERY POOR VERY GOOD

BAGGER 2 (short-sleeved checked shirt)

1. BAGGING PROCEDURE
(e.g., opened bag by sliding hand into it, used plastic bags for frozen foods, etc.)

1	2	3	4	5
VERY POOR				VERY GOOD

2. CUSTOMER RELATIONS
(e.g., handed groceries directly to customer)

1	2	3	4	5
VERY POOR				VERY GOOD

3. PROFESSIONAL DEMEANOR
(e.g., threw objects in air, engaged in horseplay, etc.)

1	2	3	4	5
VERY POOR				VERY GOOD

4. PERFORMANCE OF "OTHER" DUTIES
(e.g., helped out at other stations, cleared front of small baskets, etc.)

1	2	3	4	5
VERY POOR				VERY GOOD

5. OVERALL PERFORMANCE

1	2	3	4	5
VERY POOR				VERY GOOD

BAGGER 3 (short-sleeved white shirt)

1. BAGGING PROCEDURE
(e.g., opened bag by sliding hand into it, used plastic bags for frozen foods, etc.)

1 2 3 4 5
VERY POOR VERY GOOD

2. CUSTOMER RELATIONS
(e.g., handed groceries directly to customer)

1 2 3 4 5
VERY POOR VERY GOOD

3. PROFESSIONAL DEMEANOR
(e.g., threw objects in air, engaged in horseplay, etc.)

1 2 3 4 5
VERY POOR VERY GOOD

4. PERFORMANCE OF "OTHER" DUTIES
(e.g., helped out at other stations, cleared front of small baskets, etc.)

1 2 3 4 5
VERY POOR VERY GOOD

5. OVERALL PERFORMANCE

1 2 3 4 5
VERY POOR VERY GOOD

APPENDIX F

JUDGMENT MEASURE ORGANIZED BY DIMENSION

The following is a list of 5 performance dimensions. Please circle a number on the corresponding line to indicate your judgment of each bagger on that dimension.

1. BAGGING PROCEDURE
(e.g., opened bag by sliding hand into it, used plastic bags for frozen foods, etc.)

Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 2 (short-sleeved checked shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

2. CUSTOMER RELATIONS
(e.g., handed groceries directly to customer)

Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 2 (short-sleeved checked shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

3. PROFESSIONAL DEMEANOR
(e.g., threw objects in the air, engaged in horseplay, etc.)
Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 2 (short-sleeved checked shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

4. PERFORMANCE OF "OTHER" DUTIES
(e.g., helped out at other stations, cleared front of small baskets, etc.)
Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 2 (short-sleeved checked shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

5. OVERALL PERFORMANCE
Bagger 1 (long-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 2 (short-sleeved checked shirt)

1 2 3 4 5
VERY POOR VERY GOOD

Bagger 3 (short-sleeved white shirt)

1 2 3 4 5
VERY POOR VERY GOOD

APPENDIX G

FREE RECALL MEASURE

Please write down everything you can remember about the baggers from the videotape you have just seen.

APPENDIX H

MANIPULATION CHECK

Circle the most appropriate number for each of the following:

1. To what extent did you attempt to form impressions of the baggers while watching the videotape?

1	2	3	4	5
NOT AT ALL				TO AN EXTREME DEGREE

2. To what extent did you attempt to remember specific behaviors while watching the videotape?

1	2	3	4	5
NOT AT ALL				TO AN EXTREME DEGREE

APPENDIX I

SIGNAL DETECTION EQUATIONS

$$\text{False Positive Rate} = x = 1 - \bar{m}_{np}$$

To calculate the false positive rate (x), incorrect item responses to nonpresent behavior recognition items (saying that they did occur) are set equal to zero. Next, correct item responses are set equal to 1. Finally, the mean of these converted item responses ($\bar{m}_{np}$) is computed and subtracted from 1.

$$\text{Sensitivity} = A' = 1/2 + \frac{(y-x)(1+y-x)}{4y(1-x)}$$

The value of x is computed as above. To calculate y (the hit rate), incorrect responses to present behavior recognition items (responding incorrectly to the number of times that they occurred) are set equal to zero. Next, correct item responses are set equal to the number of times that particular behavior occurred. Finally, the mean of these converted item responses ($\bar{m}_{p}$) is computed. This mean is equal to y.

APPENDIX J

JUDGMENT ACCURACY EQUATIONS

For a rater who evaluates n ratees on k items or dimensions, scores on Elevation (EL), Differential Elevation (DEL), Stereotype Accuracy (SA), and Differential Accuracy (DA) are given by the square roots of the following terms (values are normally reported in their squared state):

$$EL^2 = (x_{..} - t_{..})^2$$

$$DEL^2 = 1/n \sum_i [(x_{i.} - x_{..}) - (t_{i.} - t_{..})]^2$$

$$SA^2 = 1/k \sum_j [(x_{.j} - x_{..}) - (t_{.j} - t_{..})]^2$$

$$DA^2 = 1/kn \sum_i \sum_j [(x_{ij} - x_{i.} - x_{.j} + x_{..}) - (t_{ij} - t_{i.} - t_{.j} + t_{..})]^2$$

where x_{ij} and t_{ij} = rating and true score for ratee i on item j; $x_{i.}$ and $t_{i.}$ = 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 (Cronbach, 1955). The sum of these four components yields Overall Accuracy. Lower values of these 5 measures reflect greater accuracy.

The 6 page vita has been removed from the scanned document

The vita has been removed from the scanned document

The vita has been removed from the scanned document

The vita has been removed from the scanned document

The vita has been removed from the scanned document

The vita has been removed from the scanned document