To the Graduate Council:

I am submitting herewith a dissertation written by Thomas Winslow Mason entitled "Replacing the Employment Interview with Biodata: A Written Structured Interview Scored with Modeled Decisions." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Industrial and Organizational Psychology, and a minor in Statistics.

John M. Larsen Jr., Major Professor

We have read this dissertation
and recommend its acceptance:

Accepted for the Council:

Vice Provost
and Dean of The Graduate School

REPLACING THE EMPLOYMENT INTERVIEW WITH BIODATA: A WRITTEN STRUCTURED INTERVIEW SCORED WITH MODELED DECISIONS

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Thomas Winslow Mason

June 1988

DEDICATION

To Michelle Harold Mason, who had the strength to help me grow, and the courage to let me try; to our beautiful children, Rachael Kathryn, Paige Elizabeth, and Blake Thomas, who were oblivious to graduate work, and made me enjoy life; and to my parents, William Ernest Mason and Evelyn Smith Mason, who taught me faith, love, and perseverance.

ACKNOWLEDGMENTS

Many people generously gave me their assistance in the course of dissertation research. I particularly am indebted to Professor John M. Larsen, who chaired my committee; Professors Robert T. Ladd and Michael C. Rush, who served on my committee; and Professor Ralph G. O'Brien, who served on my committee and directed my work to complete the minor in Statistics. They offered the encouragement and guidance I needed to undertake this most imposing project. I also am indebted to Gary Patrick and Elenore Patrick who provided valuable assistance in obtaining data for this research; and to Lee Ingram of Brewer Ingram Fuller Architects, who supported the printing of this dissertation. I am grateful to everyone at the corporation which provided the data. However, I take full responsibility for any errors or omissions.

ABSTRACT

Three bodies of literature regarding employment interviews, biodata inventories, and decision modeling (or policy capturing) were reviewed, and an integration of major conclusions was presented. It was proposed that a structured interview could provide a good foundation for a biodata inventory. The result was expected to be a biodata inventory developed using theory, that would efficiently gather interview-type information, and could be scored statistically instead of clinically. Also, a scoring method was proposed as an alternative to the often-criticized empirical, rational, and intuitive methods: The decision model of a preferred expert could be used to score the inventory, especially if a well-defined or relevant criterion of performance were unavailable (e.g., for managerial selection).

These proposals were tested by having two experienced interviewers read and rate questionnaires (BSIs) completed by 166 retail store managers as part of a concurrent validation study. The BSI was a biodata inventory developed by one of the raters to simulate the interview he used to assess managerial applicants. The ratings were modeled consistently by small subsets of the 262 variables derived (a priori) from the BSI. The models were discussed in the context of Brunswick's Lens Model paradigm. Differences in prediction among double-crossvalidated and unit-weighted models were assessed using a variation of Efron's (1982) bootstrap method. Validity coefficients were reported for organizational criteria: Raters' global decisions about hiring and promotion were not valid, but statistical combinations of component (or dimension) ratings were moderately valid, and empirically-scored BSIs had strong validity coefficients.

Problems of criterion bias and relevance were discussed, as were suggestions for further research to clarify some of the findings.

TABLE OF CONTENTS

CHAPTER		PAGE
I.	INTRODUCTION	1
II.	A REVIEW OF RELEVANT LITERATURE	10
	The Employment Interview	11
	Validity and Reliability	11
	Dimensions Measured in the Interview	13
	Problems Encountered When the Employment Interview is Used	22
	Suggestions for Improving the Employment Interview	50
	Interview Summary and Conclusions	59
	Biodata Inventories	60
	Biodata Development and Scoring	67
	Modeling Decision Behavior	72
	Clinical Versus Statistical Decisions	72
	Decision Modeling	75
	Bootstrapping	82
	Decision Modeling in Industrial and Organizational Psychology	86
	Summary of Literature Review	93
III.	METHOD	96
	Subjects	96
	Applicants	96
	Raters	96
	Instruments	97
	Biodata Screening Instrument (BSI)	97
	Weighted Application Blank (WAB)	99
	Rating Scales	100
	Criteria	101
	Financial	101
	Ratings	104
	Reliability of Criteria	105
	Composite Criterion	108
	Procedure	111
	Data Collection	111
	Data Analysis	112
	Hypotheses and Research Questions	114
IV.	RESULTS	116
	Coding Accuracy	116
	Demographics	116
	Rating Scale Characteristics	117
	Summary Statistics	117
	Reliability	119
	Criteria	122
	Individual Criteria	122
	Composite Criterion	122

Hypothesis Tests 122
Hypothesis 1 122
Hypothesis 2 125
Hypothesis 3 127
Hypothesis 4 127
Hypothesis 5 129
Hypothesis 6 132
Hypothesis 7 135
Hypothesis 8 136
Hypothesis 9 140
Hypothesis 10 142
Research Questions 144
Question 1 144
Question 2 144
Question 3 148

V. DISCUSSION AND CONCLUSIONS 153
Discussion 153
Qualifications and Limitations 168
Conclusions 170
Suggestions for Future Research 178

BIBLIOGRAPHY 180

APPENDIX 192

VITA 223

LIST OF TABLES

TABLE		PAGE
I-1.	Regression Models Used to Create Scoring Rules	7
II-1.	Validity Coefficients for the Situational Interview as Reported in Latham and Saari (1984) and Latham, Saari, Pursell, and Campion (1980)	56
III-1.	Correlations Between the Same Criteria Reported for Two Departments	103
III-2.	Intercorrelations Among Employee Attitude Scores, Performance Scores, and Store Audit Scores	106
III-3.	Reliability Coefficients for Criteria Computed from First-quarter and Year-end Data: Matched on Managers, Stores, or Managers and Stores	107
III-4.	Crossvalidated Model Statistics: A Corporate Executive's Ratings of Managers Predicted from Seven Performance Cues	110
IV-1.	Rating Scale Characteristics for Global and Dimension Ratings by Two Raters	118
IV-2.	Intraclass Correlation and Pearson Product-moment Correlation Estimates of Interrater Reliability for Two Raters	121
IV-3.	Crossvalidated Model Coefficients for Global Ratings Predicted from Selected Dimension Ratings	124
IV-4.	Comparison of Stepwise Regression Results using Linear and Nonlinear Models to Predict Global Ratings from Dimension Ratings by Two Raters	126
IV-5.	Comparison Between Crossvalidated and Unit-weighted Model Coefficients for Global Ratings Predicted from Selected Dimension Ratings	128
IV-6.	Crossvalidated and Optimal Model Coefficients for Global and Dimension Ratings Predicted from Selected Biodata Screening Instrument Cues	131
IV-7.	Comparisons Among Crossvalidated and Unit-Weighted Model Coefficients for Global and Dimension Ratings Predicted from Selected Biodata Screening Instrument Cues	133
IV-8.	Comparisons Among Unit-Weighted Model Coefficients for Ratings Predicted from Selected Principal Components and from Biodata Screening Instrument Cues	137

IV-9. Intraclass Correlation Estimates of Interrater Reliability for Actual Ratings and Ratings Bootstrapped from Biodata Screening Instrument Cues for Two Raters 139

IV-10. Comparisons Among Prediction Coefficients of Unit-weighted Models of Global Ratings Bootstrapped from Selected Dimension Ratings, Bootstrapped Dimension Ratings, and Biodata Screening Instrument (BSI) Cues 141

IV-11. Comparisons of Use of Biodata Screening Instrument Cues to Model Global and Dimension Ratings for Two Raters 143

IV-12. Systematically Resampled Validity Coefficients for Global Ratings by Two Raters Predicting Organizational Criteria 145

IV-13. Crossvalidated, Unit-weighted, and Optimal Regression Coefficients for Criteria Predicted from Selected Biodata Screening Instrument Cues 146

IV-14. Unit-weighted Coefficients for Criteria Predicted from Selected Actual and Bootstrapped Dimension Ratings 149

CHAPTER I

INTRODUCTION

Interviewing probably is used more than any other personnel selection method, even though researchers have found its mean validity is low (Hunter & Hunter, 1984; Reilly & Chao, 1982). They have offered various reasons for poor validity, including

1. Coefficients often are based on results of studies with small samples (Hunter & Hunter, 1984).

2. Interviews vary in content and structure (even when using only one interviewer; e.g., Latham, Saari, Pursell, & Campion, 1975).

3. Interviewers attend to information which is irrelevant to subsequent job performance (e.g., Mayfield, 1964).

4. Interviewers have inaccurate stereotypes of the ideal applicant (e.g., Hakel, Hollman, & Dunnette, 1970).

5. Interviewing is not the best method by which to assess particular dimensions (Wagner, 1949).

6. Interviewers usually are unreliable (Schmitt, 1976).

7. Interviewers have limited ability to process the information obtained in an interview (Schmitt, 1976).

8. Decisions may be based on perceived attractiveness, gender, race, or comparison with a preceding interviewee (cf. Schmitt, 1976).

The validity of the selection interview has been increased by establishing interviewing guidelines or specific procedures; by restricting interviewing to situations in which job performance reasonably can be inferred from interview performance; and by training interviewers to be more consistent, and to restrict

their judgments to the few dimensions they can measure best (e.g., Arvey, Miller, Gould, & Burch, 1987; Dougherty, Ebert, & Callender, 1986; Latham, Saari, Pursell, & Campion, 1980).

Interviewing job applicants essentially is a way to collect and analyze data (Kahn & Cannell, 1957), so the ultimate goal in this research was to find out whether similar data could be obtained from written responses on a standardized inventory and scored using a regression model derived from the decisions of an interviewer (i.e., a decision model). A written instrument was substituted for the interview for the following reasons:

1. The reliability of inferences (the upper bound on validity) was expected to increase because many interactions between interviewers and applicants were eliminated (Kahn & Cannell, 1957).

2. Information obtained from written instruments was expected to be similar to that collected by interviews, except for applicants' nonverbal cues (e.g., posture, style of speech, eye contact, perfumes, etc.). Nonverbal cues probably provide more irrelevant, biasing information than relevant information.

3. Written instruments, when scored with a single decision model, eliminated effects of the different skills and capabilities of various interviewers.

4. Adapting the structure of an interview to a written instrument standardized the manner in which items were presented to respondents (and judges, for scoring).

5. Written instruments were limited to appropriate dimensions more easily than interviews could have been.

6. Written instruments were administered to a large group by mail.

7. Written responses provided permanent samples of applicants' behavior, not recollections or interpretations. They were referred to repeatedly, and

stored for later organizational research and validation studies by the corporation.

8. Written information was easier to compile statistically than verbal responses would have been.

Furthermore, written information could have been used to prepare interviewers before seeing applicants.[1]

The instrument used in this research was developed as a biographical data (biodata) inventory, except that items allowed open-ended responses instead of multiple choice items. Open-ended items were expected to capture more information than multiple choice items (Kahn & Cannell, 1957). The inventory was created to simulate a selection interview in its form and content, but with the advantages of a standardized instrument.

Decisions about biodata responses usually are nomothetic (i.e., using rules, developed rationally or empirically, to rate or rank respondents), and interview decisions usually are made clinically (i.e., having people, preferably with some expert status, judge responses individually), often during the interview. Both of these methods often have been criticized: Empirical procedures are called atheoretical and serendipitous, and clinical judgments are less valid than empirical decisions.

Researchers studying clinical judgment generally found judges were able to collect relevant information in a clinical setting (e.g., the interview), but they were unable to combine that information to make decisions as reliable and valid as decisions from statistical models (see Slovic & Lichtenstein, 1971 and Hogarth, 1980 for overviews). Schmitt (1976) believed interviewers' capabilities

---

[1]Some form of interview probably would be used in most hiring situations; therefore, this type of instrument would be used for screening applicants prior to interviewing.

to gather and process information were the critical variables in the usefulness of interviews. He reviewed studies in which most individuals could not recall accurately information collected during interviews, could not infer proper relationships among cues and future job performance, and could not combine the cues for accurate prediction. However, purely empirical methods lack face validity, and scoring keys must be updated periodically.

An alternative to clinical judgment and empirical scoring was the subject of the present research. The purpose was to investigate whether it would be feasible to predict the success of job applicants with models of the clinical decisions of two individuals. The major emphasis was on modeling decisions and applying the models to make employment decisions more efficient and reliable. Although the judges used were experienced in interviewing, it was not presumed that either of these particular judges would yield the consummate model. But if models of these judges proved successful, models of proven judges could be applied to decisions, perhaps in areas other than interviewing.

This research was undertaken because a large corporation needed a way to screen prospective managers to increase the probability of success of those interviewed, so recruiters could interview fewer applicants from an applicant pool with a higher base rate of success. Two concurrent studies were undertaken for the corporation, both of them by a consultant in managerial selection and the author. Managers of the corporation also were planning two predictive validations.

The same population of managers and the same performance criteria were used in each of the concurrent studies. The goal of the first study was to assess the predictive validity of a weighted application blank developed from applications filled out by store managers when they first applied for their

positions. The second study was focused on evaluating the predictive validity of a written instrument (the Biodata Screening Instrument or BSI) that simulated the employment interview the consultant used to assess applicants for managerial positions. It is the latter of these two studies with which the research reported here is concerned.

The BSI originally was scored by an empirical procedure (cf. Mitchell & Klimoski, 1982; England, 1971), but criticisms of empirical and rational methods suggested the search for an alternative method. So in this research, the BSI was scored using regression models derived from the decisions of two judges who read applicants' responses. In theory, models should be better than the judges themselves in making consistent, valid decisions. They also should capture some of the unique ways judges used applicants' responses: Each judge's decisions implied an intuitive theoretical model of the correlations among the items on the BSI and the judges' perception of an ultimate criterion. As noted in Chapter II, a judge's model might be used in lieu of an empirical validation, when the criterion is inadequate or nonexistent.

The judges were (a) the consultant who developed the first draft of the instrument, and (b) a company recruiter. For the purposes of this research, each judge's decisions about applicants[2] were regressed on potential predictors. Instead of evaluating the accuracy of each judge, applicants' ratings predicted from each judge's decision model were validated (Slovic & Lichtenstein, 1971; Dudycha & Naylor, 1966). The goal was to see if nomothetic decisions could be made by replacing judges with their own decision models. Models were expected to be more consistent and valid than judges themselves.

---

[2]Data used in this study were from a concurrent validation study, so applicants were current managers.

The judges' primary decisions were (a) a rating of whether each applicant should be hired, and (b) a rating of how much potential each applicant had for advancement in the company. Other variables were the judges' ratings of each applicant on eight dimensions (e.g., intellectual effectiveness, emotional characteristics, etc.). In addition, each applicant's responses were coded or transcribed from the BSI to a computer data set, and comprised a third set of independent variables. These three sets of variables were used alternately as dependent and independent variables to arrive at models of judges' decisions.

For example, each judge's ratings of applicants' potential were regressed on the judge's eight dimension ratings for the same applicants. Predicted ratings of potential from the resulting regression model were obtained, and their validity for predicting various criteria was assessed. Table I-1 shows the models investigated in this research.

Researchers studying decision modeling refer to the process of substituting predicted decisions for actual decisions as *bootstrapping*.[3] In the research reported here, a bootstrapped rating (or decision) was a rating (decision) predicted from a judge's model used in place of the judge.

Two levels of ratings were gathered from judges, and decisions were modeled at two levels: For each judge, the eight dimension ratings were used to model the two global ratings (i.e., the employment decision and the promotion decision), and the coded BSI responses were used to model the eight dimension ratings. Next, the eight dimension ratings were bootstrapped from the BSI items, and the global ratings were bootstrapped from the bootstrapped dimension

---

[3]Bootstrapping also is used in statistical literature to refer to a procedure in which random samples are repeatedly drawn from a group of observations to estimate parameters. In this research, the statistical procedure is referred to as *systematic resampling* to avoid confusion with decision bootstrapping.

Table I-1. Regression Models Used to Create Scoring Rules

Model	Dependent Variable	Independent Variables
1	Employment Decision[1]	Eight Dimension Ratings[1]
2	Employment Decision	Coded BSI Responses
3	Promotion Decision[1]	Eight Dimension Ratings
4	Promotion Decision	Coded BSI Responses
5	Each Dimension Rating	Coded BSI Responses

[1]Applicants' ratings from each judge.

ratings. The result was a complete model built from the cues on the BSI (see Figure I-1). This yielded a model for synthesizing global employment decisions, and also allowed future investigations of the utility of the eight second-level ratings for various purposes, such as employment, placement or classification.

In summary, a biodata instrument was developed for collecting data similar to that collected in selection interviews, but with fewer limitations than interviews. Next, judges with interviewing experience inferred and rated various dimensions from applicants' responses on the BSI. Models were derived by regressing the judges' ratings on coded information taken from the BSIs. The judges' models were used to predict a rating for each applicant on the various rating scales. These bootstrapped ratings were compared to (a) the judges' actual ratings, (b) scores from the BSI scored empirically, and (c) scores from a weighted application blank (WAB).

These results contributed to research on biographical inventories, interviewing, and decision making in personnel selection. The next chapter presents a review of relevant literature in these areas.

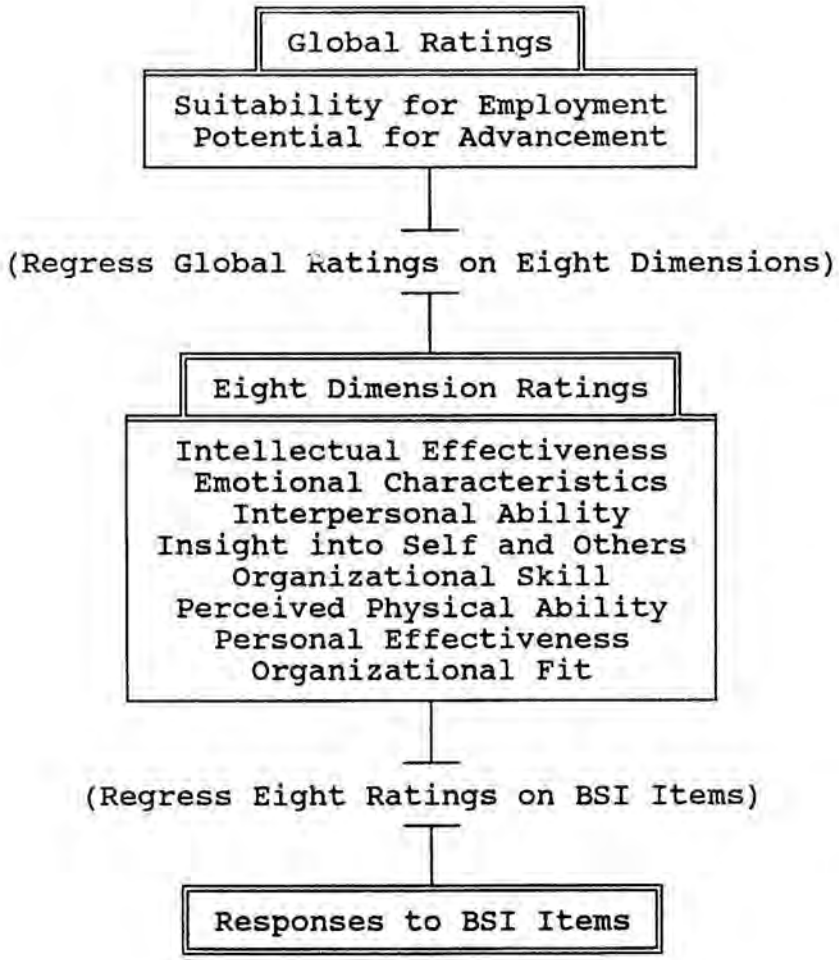

Figure I-1. Flowchart of Biodata Screening Instrument (BSI), Ratings, and Decision Models. Ratings were actual or bootstrapped from a judge's model. A direct (though imperfect) relationship between the characteristics of applicants and their responses on the BSI was presumed but not specified.

CHAPTER II

A REVIEW OF RELEVANT LITERATURE

An overview of employment interviewing, biodata inventories, and decision modeling is presented in this chapter. Many studies have been published about all three topics, so this overview is presented in three sections, described next. But most of the chapter concerns studies of employment interviewing, because it was the basis for developing the BSI.

In the first section, four reviews of different aspects of employment interviewing are presented:

1. A brief summary is presented of research concerning the reliability and validity of employment interviewing.

2. A review of research that addressed what employment interviews measure is presented. The interview should be adapted to a written instrument only if it offers some unique contribution as a tool for measurement or analysis.

3. Studies of problems with using employment interviews to collect data and make decisions were reviewed. The goal was to suggest what benefits could be gained from adapting the interview to a written format and scoring it using decision models.

4. Researchers' suggestions for improving employment interviews were reviewed because they provided the initial ideas used in this research.

Next, a review of studies of biodata inventories is presented. In this section, studies of the reliability and validity of biodata inventories were reviewed. Also, some ways that biodata instruments have been developed and scored are outlined.

Finally, a brief history of decision modeling is presented to support the scoring method used in this research. Some approaches to modeling decisions, the benefits of replacing judges with their models, and ways models have been applied to organizational decision making were reviewed.

The Employment Interview

Validity and Reliability

Even though interviewing is perceived as important for making employment decisions, studies have yielded validity coefficients on the borderline of practical significance. For example, Hunter and Hunter (1984) estimated the mean validity of the employment interview for predicting supervisors' performance ratings at about .14. This means that, on the average, a company selecting employees solely on the basis of interviews would predict only about 2% of the variance in supervisors' subsequent ratings of applicants' job performance (based on 10 studies that totaled 2694 interviews). Mean validity coefficients for predicting promotion, training success, and tenure using interviewers were .08, .10, and .03, respectively (based on 5, 9, and 3 studies that included totals of 1744, 3544, and 1925 interviews, respectively). So less than 1% of the variation in promotion, training success, or tenure was predicted by interviews. Reilly and Chao (1982) found 12 previously unreviewed studies that reported predictive validity coefficients for the interview. The mean validity coefficient was .19, based on 987 total interviews predicting supervisory ratings. The highest validity coefficients Reilly and Chao (1982) found were associated with structured interviews (cf. Latham et al., 1980). Arvey et al. (1987) also reported impressive validity coefficients for selecting sales clerks using a structured interview (corrected coefficients of .42 and .61).

The reliability of the interview rarely has been assessed. Arvey and Campion (1982) reviewed a study in which the reliability of three interviewers varied from .54 to .66 across seven rating dimensions. Latham et al. (1980) found reliability coefficients consistently above .70 among interviewers; however, their study involved a situational interview (discussed later) instead of the traditional employment interview. While the reliability coefficients reviewed by Wright (1969) were also above .70, all of the reported coefficients were for patterned or structured interviews (also discussed later). Ulrich and Trumbo (1965) reported the few reliability coefficients they found ranged from 0 to .90, and were "lower than usually accepted for devices used for individual prediction" (p. 108). In the three studies that Mayfield (1964) found which addressed reliability, *intra*interviewer reliability was satisfactory for two interviews separated by a short time period. But he concluded that *inter*interviewer reliability was poor unless a structured interview was used, because different interviewers interpreted and used the same information differently. Finally, when Wagner (1949) reviewed early employment interview studies, only two individual trait ratings had reliability coefficients consistently above .70. They were intelligence (.96, .90, .87, .77, and .62) and sociability (.87 and .72). The reliability coefficients for ratings of overall ability were considerably lower (.85, .71, .68, .61, .55, .48, .43, .26, .24, and -.20).

In summary, employment interviews were most reliable when highly structured. But even then reliability coefficients were just adequate (if .70 is the cut off point for good reliability). The aggregate validity of the employment interview is very poor: Unless interviews are highly structured and relevant to the job in question, they account for one-tenth to one-twentieth of

the variance in job performance of many other predictors (Hunter & Hunter, 1984).

Dimensions[4] *Measured in the Interview*

Out of nearly 100 traits Wagner (1949) considered, only four were measured well in the interview: academic grades, intelligence or mental ability, sociability, and overall ability. Quantitative results for motivation to work or career motivation were not reported in the literature he reviewed. He concluded

> . . . when the interview was stripped of those functions which could better be performed by other methods, all that remained was "social interaction," i.e., ability to deal with people, as revealed in an interview situation. It was felt that assigning a specific, unique purpose to the interview would result in better measurement of "social interaction" since no time would be lost in irrelevant questioning or evaluation. (p. 38)

Arvey and Campion (1982) explained that the employment interview persists as a selection tool because some (unspecified) individuals believe it is a valid way to measure interpersonal skills (presumably sociability, verbal fluency, ability to communicate, insight, etc.). Arvey and Campion (1982) proposed that the interview is more appropriate for measuring dimensions that are easily observed than for measuring dimensions that require a greater leap of inference. That is, interviewers may assess validly some dimensions that influence performance in the interview, but those same dimensions may be unrelated to performance on the job. They included motivation to work as a dimension appropriately measured in the employment interview. They did not find the interview measured this dimension well in absolute terms, but it is probably the only context in which motivation to work can be predicted with any accuracy. Certainly, motivation to work is important for selection: Measuring it, alone,

---

[4]The term "dimension" refers to applicants' traits, knowledge, skills, abilities, or other characteristics of interest to the interviewer.

could justify the use of employment interviews. Arvey and Campion (1982) supported this conclusion with research that found potential applicants and professional interviewers were more willing to discuss motivator than hygiene factors in the employment interview (see Landy & Trumbo, 1980, for a discussion of motivator and hygiene factors). If motivators exist and if they are related to motivation to work (both are unresolved issues), then applicants might disclose cues relevant to work motivation. But even then, interviewers must be able to perceive and interpret such disclosures.

Denton (1964) presented evidence that a systematic interview and a questionnaire patterned after that interview measured leadership ability and job attitudes (similar to interpersonal skills and work motivation). These dimensions were valid for predicting sales success. His study showed that the same data collected in the employment interview could be collected with a written instrument. The interview and the questionnaire also appeared to measure intelligence.

Palacios, Newberry, and Bootzin (1966) reported that *inter*interviewer reliability coefficients of ratings of employment potential and interpersonal interaction, among other dimensions, were greater than .75. The applicants were blind adults categorized according to what their jobs were (e.g., whether they worked independently) and where the jobs were located. Employment potential and interpersonal interaction (again, similar to work motivation and sociability) differed significantly among the groups of applicants. Tape recorded data and scoring protocols (similar to using written interviews and scoring models) probably were responsible for the high reliability of the ratings. The study also illustrated the benefits of somehow recording responses for later use.

Ulrich and Trumbo (1965) reviewed the literature and concluded interviewing was an appropriate way to assess personal relations and career motivation:

> In other words, perhaps the interviewer should seek information on two questions: "What is the applicant's motivation to work?" and "Will he adjust to the social context of his job?" Such an approach would leave the assessment of abilities, aptitudes, experience, and biographical data to other, and, in all likelihood, more reliable and valid sources. Since the two traits in question probably represent areas in which some of the least successful assessment attempts have been made, the interview may have as much utility as any alternative device currently available. (Ulrich & Trumbo, 1965, p. 113)

Mayfield (1964), however, concluded after his review that only intelligence was validly measured in the interview.

Gifford, Ng, and Wilkinson (1985) selected motivation to work and social skills as the two dimensions their subjects (who watched videotapes of actual interviews) were to rate. They chose these dimensions because they felt that (a) they were supported in the literature, and (b) they were qualities required for the specific position they studied (although they did not report that a job analysis had been conducted). Their subjects' ratings were valid for assessing social skills, but not for assessing motivation to work. However, no verbal or aural cues were presented because this was a study of nonverbal cues. Ratings of social skills predicted applicants' reports of social behavior (the operational criterion), but little was revealed about inferring work motivation from nonverbal cues. The criteria of motivation to work and social skills were questionnaire items similar to items on the BSI intended to assess similar dimensions. Gifford et al. (1985) presumed that the dimensions of interest could be assessed better by a questionnaire than by asking applicants directly, a major premise of the BSI research.

James, Campbell, and Lovegrove (1984) administered the California Personality Inventory (CPI) to applicants for police work, and tested differences

on specific scales of the CPI between those who were accepted and those who were rejected by the interview process. They interpreted these differences as how well the interview board differentiated personality dimensions in the interview. They acknowledged the circularity in using a single personality instrument, but felt the CPI was an adequate referent. The scales with significant differences between groups were represented well by two principal components, even though the interview boards had been instructed to assess eight dimensions. The principal components were correlated with orthogonal CPI scales: The first component seemed to be an overall character dimension, and the second measured educational achievement. The first dimension was related to social conformity (an interpersonal characteristic) for men, and was uninterpretable for women. "[A]s the major device for enacting the qualitative aspects of selection policy, the interview appeared unable to encompass multifaceted features of job suitability" (James et al., 1984; p. 132). Instead, it seemed useful for one or two dimensions.

Arvey and Campion (1982) discussed the validity of interviews for predicting personality traits. One presumes that they were actually referring to the validity of interviewers' judgments for predicting applicants' performance on personality instruments. However, one must first determine how well the criterion is related to the dimension of interest (cf. Ulrich & Trumbo, 1965). With few exceptions (e.g., Jackson, Peacock, & Holden, 1982; Jackson, Peacock, & Smith, 1980), researchers have not attempted to validate the personality constructs they purported to study. Even when researchers attempted to validate the personality constructs their subjects measured, they generally relied on occupational stereotypes or a single standardized measure, and they studied personality traits only (e.g., Jackson et al., 1980; Jackson et al., 1982).

Although the issues of measuring the ultimate criterion and construct validity are not elaborated here, they are critical to interview research.

Jackson et al. (1980) investigated whether interviewers could infer and use personality traits. In their studies, subjects successfully used implied personality information presented on paper (in interviewing vignettes) to select (or place) applicants for various occupations. The vignettes were comprised of items from the Personality Research Form (PRF) that were hypothesized to measure specific traits. The traits chosen were those that had high common loadings with items taken from the Strong Vocational Interest Blank (SVIB). The items selected from the SVIB were those that discriminated among responses from subjects in different occupational groups. This study followed others in which stable and replicable associations among such hypothetical candidates (i.e., contrived personality profiles) and occupations were perceived by subjects (Jackson et al., 1982; Jackson & Siess, 1970; Rothstein & Jackson, 1980). However, the premises of these studies depended on the unproven generalizability and validity of two instruments: the PRF and the SVIB. This program of research also depended on at least two unsupported hypotheses: Implicit personality theories (IPTs) of subjects are related to job performance, and the interview is the best way to measure personality traits. The researchers assumed that the dimensions measured best in the interview were also measured by the instruments they used. They first should have studied the PRF and the SVIB directly. These studies illustrated that interviewers (and students) applied their IPTs when they were presented with obvious personality cues. Unfortunately, they addressed only the question of whether IPTs were used by interviewers, instead of whether specific dimensions were appropriately

measured in the employment interview. Also, there was no evidence that the same interviewers would perceive these traits in an ecologically valid setting.

Schmitt (1976) noted that Grant and Bray (1969) found assessment center ratings of managers' personal characteristics, career motivation, work motivation, and control of feelings were based in part on information gathered in interviews. The ratings may have measured predictors of either job performance or organizational fit (i.e., attitudes or characteristics other than knowledge, skills, or abilities that render one applicant more suitable than another in a specific organization), because about half of the personal characteristics were significantly related to salary progress. Organizational fit is consistent with the findings of Klimoski and Strickland (1977), that assessment centers measure future progress within the organization (i.e., organizational fit), instead of effectiveness or job performance. Personal characteristics perceived by interviewers may influence ratings because they intuit that the applicant will "fit in" the organization. Wagner (1949) quoted Bingham regarding the predictive validity of rating halo:

> "Finally, it is not the rater alone whose reactions to the candidate are in question. He is but typical of others -- clients, subordinates, fellow employees -- who will react to the prospect, not as a bundle of isolated traits, but as a person with certain duties. The judgments and responses of all these people will unconsciously and inevitably manifest a halo effect which is in part at least, valid" (Wagner, 1949, p. 40).

This conception of halo supports the idea that employment interviewers can measure organizational fit. Further evidence was found in studies of similarity effects (discussed more fully later). Researchers inferred similarity effects when interviewers made different decisions depending on whether they perceived applicants as having attitudes similar to their own. Schmitt (1976) suggested that interviewers can use written biographical data to infer attitudes similar to their own. While generally presented as bias, similarity effects may aid

interviewers in recognizing applicants that will remain with the company, be more satisfied, or adapt to prevailing norms. In other words, if an applicant is similar to the interviewer, and the interviewer fits the organization, then the applicant should also fit the organization. But organizational fit may be unrelated to an applicant's abilities, skills, or motivation to work.

Wright (1969) reviewed work by Maier and his associates investigating whether untrained raters could detect which interviewees were deceitful. These studies reported relatively good accuracy, but the manipulation was a specific incident (i.e., a student requesting a grading correction on a returned test) that afforded the subjects a single judgment of whether the student was lying. The question of whether the employment interview is appropriate for judging personal integrity remains untested, because interviews are not discussions of an isolated incident.

Sterrett (1978) considered in his study eight traits he believed were commonly considered in the hiring process: ambition, motivational drive, self confidence, self organization, responsibility, verbal ability, intelligence, and sincerity. However, he offered no evidence that these dimensions were validly inferred in the interview. Although the primary focus of the study was on whether applicants' body language affected how interviewers perceived them, he approached the interview as most interview researchers have, presuming that certain traits are measured in the interview, even though few studies have tried to define them.

In the how-to-do-it category, Half (1985) cautioned potential interviewers about the difficulty of assessing personality in the interview and suggested that they concentrate on two to three specific dimensions related to the job. Of the impressions gained in the interview, he emphasized the importance of

experience, education, intelligence, appearance and personality, innate ability, and motivation. He gave specific guidance on assessing industriousness, intelligence, temperament, creativity and resourcefulness, confidence, and motivation and drive, and encouraged interviewers to attempt to assess how well applicants will fit the organization. He offered no empirical support for his recommendations and conclusions.

Fear (1984) and Fear and Ross (1983) proposed that interviews are appropriate for assessing the "3 M's": motivation, maturity, and mental ability. The characteristics of motivation are conditioning to work, high energy level, initiative, and conscientiousness. The characteristics of maturity are judgment, willingness to "stay put", self-insight, and relationships with people. Grades in relation to effort, academic likes and dislikes, reasons for leaving high school, and quality of response to in-depth questions all are supposed to measure mental ability. The authors quite strongly implied that these characteristics (and the traits they represent) can be assessed virtually without error in the interview (p. 206). Two traits that stood alone for these authors were tough-mindedness (a component of leadership) and adaptability (*"because it is so important in hiring people without previous experience*"; Fear & Ross, 1983; p. 209). Interestingly, Fear (1984) reported studies of the effectiveness of interviews in which about two-thirds of the inferences were based on verifiable or biographical information, and only about one-third were based on intuition or impressions. Yet he did not suggest that verifiable or biographical information could be gathered more efficiently by biodata inventories or weighted application blanks.

In a book based on considerable experience, Rohrer, Hibler, and Replogle (1965), a firm specializing in executive assessment, listed five aspects of

behavior predicted or identified by clinical methods. These broadly defined dimensions are those that "foreshadow managerial ability" (p. 20). The authors believed that they must be assessed by clinical methods because they deal with "intangibles and high-level abstractions" (p. 5) best measured by psychologists' evaluation interviews. These five aspects are "[a] intellectual competence, [b] emotional stability, [c] skill with people and the ability to interact effectively with others, [d] insight into and understanding of human behavior, and [e] the ability to plan and organize effectively" (p. 20).

Researchers (except Mayfield, 1964) seemed to agree that the interview should be useful in assessing interpersonal and motivational dimensions of applicants. However, there are obstacles, including (a) inadequate criteria to validate inferences, (b) interviewers' inability to gather evidence of specific dimensions and reliably interpret it, and (c) unknown relationships among dimensions, inferences, and job performance (Ulrich & Trumbo, 1965). This last obstacle is most serious when the criterion used for job performance is irrelevant to ultimate job performance (i.e., some inferences might be valid for predicting an ultimate criterion, but an appropriate operational measure must be used to estimate that relationship). In summary, there is a consensus that the employment interview can uniquely contribute to employment decisions. The dimensions that most researchers concluded were measured best with the interview are social skills and motivation to work. Authors of self-help literature, however, condoned using the interview to assess almost any dimension of interest.

Even if dimensions are assessed accurately in the employment interview, interviewers must be able to collect and use that information correctly. Some problems associated with employment interviewing are presented next.

Problems Encountered When the Employment Interview is Used

According to Kahn and Cannell (1957), employment interviewing is a way to quickly gather information, and it requires a process to discard irrelevant data and retain relevant "bits". But final decisions depend on the interaction of two roles, interviewer and applicant, each of which influences the other. This picture of the employment interview provided the structure for the following review of research on the inadequacies of the interview. First, studies of the interview as a way to collect data were reviewed. Researchers investigated how effective the employment interview was for gathering information used to make employment decisions. Next, studies of problems inherent in making decisions in the interview were reviewed. Such problems involve the interaction between participants, as well as interviewers' prediction errors.

Problems with interviewing to collect data. Wagner (1949) did not specifically address employment interviewing as a way to collect data, but he included the following quotation he attributed to Wonderlic (no reference given):

> "As a procedure, interviewing is not useful for obtaining facts; all important information obtained in an interview should be verified by one means or another -- by checking references, by credit investigation, by inspection, etc. Interviewing is not a good device for measuring mental ability or job skills, nor is it particularly useful in discovering hidden aptitudes of possible candidates." (Wagner, 1949; p. 34)

The question of how accurately data are collected is different from the question of how reliable or valid interviews are; the latter question is concerned with how well interviewers infer dimensions, but the former addresses how available the raw facts are, and how well interviewers collect them. Ulrich and Trumbo discussed several studies that explicitly looked at the accuracy of data collected during employment interviews. One of the studies, by Keating, Paterson, and Stone (1950), verified stated experience information

obtained during interviews. Better than 90% of the data were verified by prior employers for wages, tenure, and duties. Schmitt (1976) cited research by Carlson (1967) that found interviewers, on average, were able to answer only about 50% of the questions asked immediately after an interview, even though the questions were factual. Schmitt (1976) suggested interviewers' talk inhibited data collection, because it blocked information that might have invalidated or expanded their first impressions or presumptions. Reilly and Chao (1982) summarized the validity and reliability of employment interviewing and concluded that as a device for collecting data, making inferences, and selecting employees, it falls far short of standardized instruments. Arvey and Campion (1982) reviewed only one study that investigated how well interviewers collected data: The highest accuracy rating was about 79%.

In studies by Weiss, Dawis, and their colleagues (e.g., Weiss & Dawis, 1960; cited in Ulrich & Trumbo, 1965), the accuracy of information suffered when applicants perceived some responses as less socially desirable than others. Ulrich and Trumbo (1965) also reviewed studies that found bias associated with the social desirability of responses. While this kind of bias was not specific to employment interviews, employment interviews and a questionnaire collected different information in other studies (particularly that of Metzner & Mann, 1952, and Bennett, Allport, & Goldstein, 1954). These studies supported Kahn and Cannell's (1957) postulation that the information obtained in the interview will differ as a function of the psychological dynamics of a particular interview.

Of Mayfield's (1964) 15 conclusions about employment interviews, five concerned data collection:

1. In unstructured interviews, not all topics will be covered similarly. In the studies he reviewed, biographical items or factual items were covered most

consistently. The kind of material covered least consistently involved applicants' attitudes.

2. Applicants' responses depended on how questions were asked. Without exactly the same questions (and presumably the same presentation style), different data will be collected.

3. Interviewers will perceive applicants' responses differently depending on their own attitudes and opinions. This conclusion was extrapolated from the opinion research literature.

4. Interviewers talked more than applicants in unstructured interviews, especially after they decided to hire the applicant.

5. Interviewers tended to make their employment decisions early, in unstructured interviews.

Perhaps when interviewers found evidence confirming preexisting positive judgments, data collection stopped either because attempts to influence the applicant began (Anderson, 1960), or because irrelevant topics were discussed (Ulrich & Trumbo, 1965). Anderson (1960) first investigated the relationships among the amount of time the interviewer and applicant each talked, the amount of time neither talked, and the decision to accept or reject the applicant. There was less time free of talk and more time taken by the interviewer, when the interviewer accepted the applicant. The interviewer determined the amount of time the applicant talked and the amount of time neither talked.

Schmitt (1976) proposed several ways data collection is affected in the employment interview. For example, interviewers lose interest and stop collecting data after they make up their minds, usually early in the interview. Forcing the interviewer to use a data form may increase attentiveness, and help

the interviewer gather more accurate information. Also, racial differences between interviewers and applicants can affect the quantity and quality of the information available to the interviewer: The amount of information disclosed by the applicants depended on the race of interviewers and applicants. He also cited evidence of gender differences, where in spite of equal qualifications, interviewers placed male and female applicants in different jobs (e.g., Dipboye, Fromkin, & Wiback, 1975). But it was unclear whether interviewers' decision processes were faulty, or they were unable to collect equivalent data on different sexes. For example, an interviewer might expect males and females to have different qualifications.

Snyder and Swann (1978) investigated the strategies interviewers used to question applicants, and found interviewers appeared to seek information to confirm their hypotheses about particular applicants. More recently, Sackett's (1982) three studies extending Snyder and Swann's (1978) research to other populations and dimensions weakly affirmed their findings. But he concluded that the kinds of hypotheses held by interviewers affected questioning strategies. His conclusion differed from Snyder and Swann's (1978) in that interviewers continued to use the strategies initially chosen, instead of adjusting them to search for confirming information, as they had hypothesized. A later study by McDonald & Hakel (1985) supported Sackett's (1982) conclusions: Interviewers in a maximally controlled experiment did not engage in a search for confirming evidence, nor did they look specifically for negative information. Perhaps interviewers essentially duplicated a weighted application blank or a biodata inventory by finding and following one strategy. This means that an interviewer's strategy might be simulated with a written format, improving consistency and data collection.

Palacios, Newberry, and Bootzin (1966) recorded, transcribed, and scored applicants' responses to interviewers' questions. Interrater reliability was good, and they concluded that their scoring method was successful. They attributed high reliability to recording and transcribing applicants' responses. In other words, they relieved the interviewer of most of the responsibility for collecting data. These results were similar to those obtained by Denton (1964): When questions asked by interviewers were adapted to a questionnaire, responses were collected, scored, and used better. In the most specific test of employment interviews to that time, Walsh (1967) found no evidence that the interview was any better or worse than a questionnaire or a personal history form (i.e., a biodata inventory) for obtaining self-report information. There was no difference even when a financial incentive ($15.00 in 1965) was paid to subjects to provide distorted information. However, all the information requested was verifiable, unlike some that might be requested in an employment interview. These three studies support questionnaires as alternatives to employment interviews for collecting the same data as interviews, except permanently and consistently.

Reilly and Chao (1982) recognized that the interview is a multipurpose tool, because it also can be used to disseminate information (probably better than it collects it). However, they particularly noted its low utility. It must be extremely valid before it can be useful, because it is so expensive: At least one administrator (or even a panel of administrators) must be present for each applicant (Schmitt, 1976). It will be cost effective only if it collects unique data (e.g., motivation, social skills), and the data predict well enough to overcome the cost advantage of other methods, such as biodata inventories. In the BSI research, the goal was to measure inexpensively the same dimensions as

the interview, while improving the collection of other information. This was expected to lead to an increase in utility over that of the interview.

In criticizing the study of a highly structured interview format by Bonneau (1957), Ulrich and Trumbo (1965) questioned the utility of the interview for collecting data. They wondered why an interviewer would be used in such a "perfunctory role . . . one is hard pressed to say why the interview rather than a questionnaire was used" (p.105), and concluded that

> even when it appeared that the interview was the sole basis for the decision, it was not clear that the same predictive efficiency could not have been achieved from other, probably less expensive, sources. . . .
> . . . . . .
> the interview is obviously a costly procedure for gathering data, with cost increasing as some function of length. (pp. 113-114)

In summary, there are not many studies on which to base a conclusion about the employment interview as a means of collecting information. However, in the few relevant studies conducted, interviews were no better for gathering information about applicants than were paper-and-pencil methods. If one were to perform a utility study comparing the interview and written methods, there is no evidence that the interview would be as cost-effective as other methods. Therefore, adapting the interview to a written format should yield less costly information with which to make employment decisions. Problems that occur when interviewers attempt to make these decisions have been widely studied, and a review is presented next.

Problems of decision-making in the employment interview. Wagner (1949) reviewed research "designed to study the value of the interview not only to discover information but as a means of carefully synthesizing all the data into an evaluation or prediction of over-all ability, proficiency, or potential job success" (p. 23). He concluded that interviewers are too unreliable to be left with the task of predicting applicants' success, except when (a) rough screening

is adequate, (b) there are too few applicants for an empirical study, or (c) the dimensions measured best by the interview are relevant to the job. Since Wagner (1949), research on the employment interview has proceeded from macroanalytic studies of predictive validity to microanalytic investigations of the content of the interview and the process by which interviewers make decisions about applicants (Arvey & Campion, 1982).

The most influential studies of decision-making in the employment interview were conducted at McGill University, and reported by Webster (1964). This program of microanalytic research covered 10 years and resulted in several separate publications about how interviewers make employment decisions. The hallmark of these studies was their focus on investigating how interviewers made decisions and how they were affected by applicants' cues, instead of computing the validity of the interviewer's predictions about applicants' future tenure or work behavior.

In the first of these studies, Springbett (1958) studied whether the order in which applicants and their applications were presented made a difference in interviewers' decisions. Interviewers either saw the applicant before they read the application, or they read the application first. Several findings of interest emerged:

1. Interviewers made decisions early in the interview. The applications and applicants' physical appearance "provide[d] information in the first two or three minutes of the interview" (p. 22) that correctly predicted 85% of the time whether applicants would be accepted or rejected.

2. Interviewers gave more weight to applications than to applicants' physical appearance.

3. Interviewers primarily sought to reject applicants, and searched for negative information.

4. Ninety per cent of the time, interviewers rejected applicants when either appearance or applications were not rated favorably.

5. There was a primacy effect for the application: Applicants more often were accepted when applications were read first, when the applicant's appearance and application form were both rated favorably.

Springbett (1958) rationalized the results for each of the findings in cognitive terms. He proposed that interviewers made decisions early in each interview because the application and the applicant's appearance introduced a perceptual set (Gibson, 1941), and interviewers tended to perceive information that confirmed their initial hypotheses. He suggested that applications imposed structure on the way interviewers gathered information, and this accounted for the primacy effect for applications. He postulated that negative reinforcement caused the search for negative information: Interviewers rarely received feedback about successful applicants they selected, or the potentially successful applicants they rejected, only about the failures they selected. So interviewers employ the search for negative information to avoid the negative reinforcement of replacing failed employees. Interviewers had less confidence in applicants' appearance than in the applications, since the application and appearance both had to be favorable for a positive rating, when the application was read first. He proposed that the concrete nature of the information on the applications influenced the interviewer more than the tentative, subjective information contained in appearance.

In another of the McGill studies, Sydiaha (1959) tested Meehl's (1954) hypothesis that the clinicians (in this case, interviewers) use both systematic

(i.e., test, biographical, mechanically obtained) and unsystematic (i.e., qualitative, intuitive, idiosyncratic) information, while statisticians (when using empirical models) have only systematic information available. He questioned whether clinical and statistical decisions relied on the same systematic information. He also posited that clinicians were under the illusion that they used unsystematic data, and they actually made decisions with systematic data only. He tested these hypotheses by correlating the actual decisions of the interviewers with (a) models of their decisions, and (b) an empirical model. The results were significant and supported his hypotheses:

1. Interviewers' actual decisions correlated better with ratings produced by interviewers' models than with those produced by the empirical model.

2. Predictions produced by interviewers' models and by the empirical model were not colinear.

3. A single model adequately described the decisions of the various interviewers used in the study.

4. Interinterviewer reliability was high, and there was no evidence of rating halo (i.e., colinearity among the dimensions on which interviewers rated applicants).

He concluded that (a) interviewers made different decisions than linear models, and (b) all interviewers (at least in his study) used substantially similar decision methods. This last conclusion is not consistent with other studies of employment interviewing or decision modeling, that have suggested that different interviewers use different processes or criteria to judge applicants (cf. Arvey & Campion, 1982; Mayfield, 1964; Schmitt, 1976). For example, in an early decision modeling study, Hakel, Dobmeyer, and Dunnette (1970) compared the importance 22 recruiters (who were also CPAs) and 20 male psychology

students placed on scholastic standing, business experience, and interests. The interviewers all used the information differently, and the impact of unfavorable information depended on the content area (good scholastic standing compensated for poor work experience and inappropriate interests). The importance interviewers placed on the various content areas mediated their evaluation of the applicants.

Sydiaha (1961) conducted two studies addressing the interaction between interviewers and applicants. He used Bales' interaction analysis coding scheme to categorize various events in the interview. Decisions to accept the applicant were associated with less questioning, more problem solving, and positive social interaction (for both interviewer and applicant). Unlike earlier results (Sydiaha, 1959), different interviewers made different decisions, and they used different information. The second study discriminated between interviewers' and applicants' events, and produced results consistent with the first study. Both interviewers' and applicants' behaviors accounted for significant amounts of variance in whether the applicant was accepted or rejected by the interviewer (about 38% combined). As in the first study, different interviewers made different decisions. Sydiaha (1961) drew two conclusions from these studies that continue to influence studies of how decisions are made in the employment interview:

1. When presented with (possibly relevant) information about an applicant's interpersonal skills, interviewers differ in their ability to gather and use that information.

2. The "social atmosphere" created by the interaction between the interviewer and the applicant may be a source of irrelevant information. To the extent that such irrelevant data were used, the resulting decision would be

erroneous. Sydiaha (1961) raised the related question of whether an applicant would be able to create a favorable social atmosphere by faking, and thus lead the interviewer to an unwarranted favorable decision. Such attempts might be avoided by using written instruments when feasible.

Sydiaha (1962) next investigated whether interviewers made empathic decisions based on applicants' inferred attitudes, intentions, or tendencies. This investigation was the basis for subsequent studies of attitude similarity. These were concerned with the effects of interviewing applicants with attitudes similar to those of the interviewer. Sydiaha's (1962) findings suggested that only specific interviewers made empathic decisions, because correlations between empathy (operationalized as accuracy of predicting applicants' attitudes, assumed similarity of attitudes, and actual similarity of attitudes) and decisions to accept the applicant varied from -.45 to .84. Interviewers sometimes attributed characteristics to applicants that the applicants did not attribute to themselves, and often perceived greater similarity between themselves and applicants than appeared justified. Sydiaha (1962) concluded that interviews would be effective if a consistent interviewing method were used, and if the information collected is combined statistically, instead of by "apparent relevance" (p. 349). The alternative proposed later is to use the model of a good interviewer.

Finally, Rowe (1963) tested her hypothesis that interviewers vary on the trait of "category width." She proposed that interviewers will accept or reject significantly different proportions of applicants depending on how broad the interviewers' conceptions of "successful applicant" are. Although her hypothesis was supported, the interviewers all tended to select the same (paper) applicants. She concluded that (a) applicants can be scaled on how acceptable they are to a group of interviewers, and (b) interviewers will select different numbers of

acceptable applicants depending on the interviewers' category width. Her results support the feasibility of a written substitute for the employment interview, and using decision modeling to score it. This would allow companies to rank order applicants on their substitute interview performance (or predicted interview performance, if it were used to screen applicants).

These five publications (i.e., Springbett, 1958; Sydiaha, 1959, 1961, 1962; Rowe, 1963), among others, laid the foundations for later studies of the difficulty of making valid and reliable decisions in the employment interview. This research covered the following major topics: (a) primacy effects; (b) the effects of application forms, tests, or other factual data on interviewers' decisions; (c) the interview as a search for negative information; (d) using questioning strategies that only confirm the interviewer's first impressions; (e) differences in interviewers' decision behavior (and ways to deal with those differences); and (f) social, interpersonal, similarity, racial, gender, and age effects on interviewers' decisions. Unfortunately, the microanalytic nature of the McGill studies and those that followed them did not help integrate results into a unified model of how interviewers make decisions (Schmitt, 1976; Wright, 1969). However, this research suggested ways the employment interview could be improved (discussed later). Some of the results could not be generalized, and others appeared specific to the situation, or spurious. But many have helped improve the effectiveness and efficiency of employment interviews (cf. Latham et al., 1980).

Before reviewing research following the McGill studies, a note about methods is justified. When they attempted to exert experimental control, many researchers studying the components of interviewing used contrived applicants that existed only on paper. Gorman, Clover, and Doherty (1978) concluded that

there were enough differences in the cues available to interviewers between the "paper-people analog" (p. 165) and the real person, that studies using paper applicants may be "nothing more than interesting demonstrations, from a practical viewpoint" (p. 168). It is with the caveat -- anything short of social interaction may not generalize to real employment interviews -- that the following review is presented. On the other hand, the research reported in this paper proposes an alternative selection procedure, and is not just the simulation of an employment interview. It uses written interviews completed by real managers in a concurrent study of validity.

Schmitt (1976) found consistent support in the literature for the hypothesis that positive and negative information have different effects on the decisions of interviewers. However, it appeared that there were also interactions among positive and negative information, order of presentation, and whether information was impressionistic or factual. Arvey and Campion (1982) also found research generally consistent with earlier studies, and concluded that "interviewers tend[ed] to produce ratings or evaluations which [were] influenced by contrast, primacy-recency, first impressions, personal feelings, and other factors" (p. 297). But they also reviewed studies where interviewers took longer to make decisions than previously thought (cf. Springbett, 1958), and length of time depended on whether the applicant was perceived to be of high quality. Unfortunately, these results were based primarily on interviewer introspection, a process that may not represent the decision accurately.

More recently, Belec & Rowe (1983) investigated whether order of presentation of positive and negative information affected how often the interviewers attributed applicants' prior successes and failures to internal causes (e.g., ability and effort). When positive information followed negative

information, successful outcomes significantly more often were attributed to internal characteristics. Under those same conditions, failures were attributed to external factors significantly more often than in the other conditions (when positive information preceded negative information, or when negative information was both preceded and followed by positive information). When positive information followed negative information, more decisions to hire were made, and applicants were rated as more suitable for the job in question. The researchers concluded that there is a recency effect for causal attributions on consequent ratings. They also concluded that their findings supported previous hypotheses that contrasting information will result in a recency effect (cf. Arvey & Campion, 1982; Schmitt, 1976).

The critical manipulation in these studies was to have subjects make judgments throughout the (simulated) interview, increasing the contrast between, and salience of, positive and negative information. The results of this work were consistent with research by Tucker and Rowe (1979). They suggested that what interviewers expected of a given applicant affected the number of internal attributions for the applicant's past successes and failures. Rowe (1984) commented that "sex, race, attractiveness, and the like may create expectancies and effects . . . such that women, minority group members, or individuals similarly disadvantaged will be more likely to be blamed for their failures and less likely to be given credit for their successes" (p. 331). She also concluded that (a) negative information received more weight because of the same reinforcement hypothesis Springbett (1958) proposed, and (b) interviewers used prototypes to compare and contrast applicants. She posited that it is easier to find that an applicant does not match the prototype (by using negative information) than it is to confirm all of the requisites and conclude that the

applicant does match. (An appropriate analogy is that it is easier to reject a false null hypothesis than to confirm a true one.) She arrived at a two-stage model wherein the interviewer first ascribes characteristics to the applicant, and then determines how close the applicant is to the prototype or ideal applicant: The closer the applicant, the better the suitability rating. Her model implied that the more job-related the interviewer's prototype, the better the interviewer's prediction will be (assuming that the interviewer can accurately perceive the correspondence between an applicant and the prototype). In fact, some researchers concluded that when interviewers had job analysis information, they were better able to select applicants (e.g., Langdale & Weitz, 1973; Schmitt, 1976; Wiener & Schneiderman, 1974).

A study by Dipboye, Stramler, and Fontenelle (1984) supported Rowe's (1984) model. Poor credentials led interviewers to attribute failure to internal causes, and to categorize the applicant as less of a match to interviewers' ideal prototypes. The interviewers perceived the applicants in accordance with a prototype that was not ideal. In other words, the interviewers matched poor applicants with a poor applicant prototype. Powell (1986) also found evidence that expectations created prior to the interview resulted in differences in ratings. However, of all the variables studied, the actual qualifications of the applicant had the strongest relationship with interviewers' rated likelihood of offering the applicant a job.

Intuitively, it would seem that application forms could be used as interview guides (or at least as previews) to help interviewers perceive relevant qualifications. Several recent studies addressed that topic. In a study by Dipboye, Fontenelle, and Garner (1984), previewing application forms had no significant effect on the style or strategy of interviewing, or on interviewers'

accuracy in describing the applicant. However, reliability decreased in some of the interviewers' ratings of applicants. Interviewers who previewed applications gathered more information, but they did not recall more of that information. Those who did not preview applications had better recall of information on the application than those who previewed them. Perhaps those who previewed applications tended not to discuss that kind of information during interviews, so it was not as salient as it was for interviewers who did. Unfortunately, we do not know what information was salient at the time the interviewers made their decisions (cf. Tucker & Rowe, 1977).

Their results were consistent in part with the results of an earlier study by Tucker and Rowe (1977), in which length of time to decide, information obtained by the time they made their decisions, and confidence in their decisions were compared for interviewers who had or had not previewed applications. Unlike Dipboye, Fontenelle, and Garner (1984), Tucker and Rowe (1977) measured information recalled when the interviewer made the final decision. Both groups of interviewers took that same amount of time to decide to accept or reject applicants. Interviewers who did preview the application were able to recall more relevant application and nonapplication information than were those who did not preview the application. Perhaps those interviewers who previewed the application form did not spend time collecting redundant data. In spite of recalling less relevant information, those who did not preview the application expressed the same confidence in their ratings as those who did. If applications created expectations, then at least they were based on hard (although not necessarily veridical) data.

All the preceding studies that addressed attributions depended on models that involved some sort of prototype. Other recent studies addressed the use

of stereotypes by interviewers. Schmitt (1976) reviewed a number of studies that supported the position held by Rowe (1984; i.e., that interviewers compare each applicant to their own prototype of an ideal worker or applicant), finding that interviewers may actually judge applicants in comparison to some prototype (that may or may not be consistent throughout the interview). However, using prototypes may lead to different evaluations for groups of applicants if the prototype includes aspects of race, gender, age, or nonverbal behavior. The same would occur if interviewers have stereotypes about groups of applicants, based on sex, race, national origin, handicaps, or behavior.

Recent interviewing research addressed the effects of race, gender, age, and nonverbal behavior, more than other topics. Although not explicitly presented as studies of categorization, studies of bias against subgroups looked at whether interviewers made different judgments for members of certain stereotyped categories. Arvey (1979) described three ways in which stereotypes may operate in the employment interview. The first involves general negative abstractions that the interviewer may ascribe to a particular group. The second is the preconception that certain groups inherently may be more suited for some position than others. The third is the belief that certain groups are assessed better on certain dimensions than others. An example of the third instance is a tendency to evaluate female applicants on typing skills or attractiveness while evaluating male applicants on supervisory ability. Arvey (1979) reviewed research supportive of all three of these mechanisms at work in employment interviews. He also presented a plausible alternative to group stereotyping, in which nonverbal behaviors that are interpreted by, or confusing to, interviewers influence their decisions. The effect would be greater when interviewers and

applicants have fewer common social experiences (e.g., different races, national origins, sexes, or socioeconomic levels).

Arvey (1979) found consistent evidence that the employment interview is biased against females. For example, females consistently received lower ratings (or lower salaries, or fewer job offers) than males. Also, much evidence supports the hypothesis that females receive lower suitability ratings for masculine oriented jobs. In some cases, females received higher ratings than males for traditionally feminine jobs. However, the qualifications of the individual accounted for much more variance in suitability ratings than did gender. Finally, Arvey (1979) found studies in which sex accounted for less than one percent of the variance in ratings of competence. However, males were chosen much more often than females when subjects were instructed to select just one applicant to hire.

Arvey and Campion (1982) reviewed more studies of sex bias. In a study by Heilman and Saruwatari (1979), more attractive female candidates were rated higher for clerical jobs and lower for management jobs than were unattractive females. Attractiveness may have increased interviewers' perception of femininity, and increased their tendency to place females in traditionally feminine jobs. They encouraged more studies dealing with the context in which decisions are made. In this regard, they described a study by Heilman (1980) in which the percentage of female incumbents in a given job class affected the evaluations received by female applicants for that position.

Arvey (1979) found only three studies dealing with the race of the applicant (black or white). In none of these studies did the race of the applicant influence practically the interviewers' ratings of the applicants. Arvey (1979) noted that in two of those studies, the statistical power may have been too low

to detect racial effects. Interestingly, in a study that classified subjects according to racial prejudice, those in the racially prejudiced group gave lower ratings to both blacks and whites.

The studies of race bias reviewed by Arvey (1979) were supported by later studies investigating the combined effects of race, applicants' quality, and interviewers' prejudice. Mullins (1982) hypothesized that blacks would be rated lower than whites, particularly by racially prejudiced interviewers. Once again, applicants' quality accounted for most of the variance in ratings, although race exerted a significant effect in the direction opposite that hypothesized. Black applicants were rated higher than white applicants, even by the highly prejudiced interviewers. Again, these results are consistent with those summarized by Arvey (1979).

Arvey and Campion (1982) also found three additional studies of race or national origin. In one of the two studies in which black candidates were favored over white candidates (Mullins, 1978), applicants' qualifications were manipulated, and they were the most important variables in explaining interviewers' decisions. In the third study (Kalin & Rayko, 1978), "applicants with foreign accents were given lower ratings for the higher status jobs, but higher ratings for the lower status jobs" (Arvey & Campion, 1982; p. 304).

Arvey (1979) found only two studies of age effects. Age was a strong influence in both. Rosen and Jerdee (1976) discussed the effects of age stereotyping in managerial decisions. Subjects evaluated older employees as "deficient in on-the-job performance, potential for development, certain interpersonal skills, vitality, and propensity for risk taking. . . . [although they] were rated higher than younger persons on integrity" (p. 428).

Handicapped applicants were perceived as having greater motivation or courage than non-handicapped applicants (Arvey, 1979). But there was no evidence that handicapped applicants would actually be hired because of those perceptions.

Many of the studies reviewed by Arvey and Campion (1982) dealt with the influence of applicants' nonverbal cues. The cues involved eye contact, body language, voice characteristics, facial expressions, and so forth. The experimental manipulations varied from the position of applicants' eyes in photographs to entire repertoires of behaviors (e.g., eye contact, voice level, friendliness, distance from interviewer, etc.). There were consistent significant effects for nonverbal behavior, so Arvey and Campion (1982) raised the question of how the cues were used by interviewers. Interviewers may use them in conjunction with verbal content, or separately. They might measure them as skills relevant to the job, use them to categorize applicants, or they might influence how interviewers perceive applicants' personal attractiveness. Nonverbal cues are probably used to categorize applicants in both relevant and irrelevant ways. Arvey and Campion (1982) concluded that nonverbal cues explained less variance in interview outcomes than did verbal content, but were important variables, nevertheless.

More recent studies have addressed these related issues of stereotyping in the employment interview (i.e., based on race, gender, age, and nonverbal cues). For example, in research reviewed earlier, interviewers held relatively stable, consistent stereotypes of the traits and interests of successful workers in specific occupations (cf. Jackson et al., 1982; Jackson et al., 1980; Peacock, 1982; Rothstein, 1984). These studies used ratings of occupation-specific personality prototypes for various purposes. They provided clear evidence that

interviewers created stereotypes of applicants and compared them to occupational prototypes. That might be a valid way to collect and evaluate information about occupational or organizational fit. But if those stereotypes or prototypes involve race, sex, age, handicaps, or certain nonverbal behaviors, interviewers will base their decisions on irrelevant, and perhaps harmful, information.

Giles and Feild (1982) measured how well male interviewers could describe the job characteristics desired by white male and white female applicants. The study addressed sex stereotyping directly instead of inferring it from ratings or employment decisions. Job characteristics were not rated differently by males and females (except that males reported more of a preference to supervise others), but male recruiters made significantly more errors of preference for females than males (9 out of 13 for females vs. 5 out of 13 for males, $p < .05$). The authors concluded the results were not practically significant. But they were consistent with the findings of other studies on sex bias in the employment interview.

Reid, Kleiman, and Travis (1986) studied sex bias in the employment interview as a process of attribution. They hypothesized that interviewers' and applicants' sex would influence favorable attributions of past success (i.e., attributions to ability and effort), and male interviewers would attribute past success most favorably for male applicants. They expected a larger effect for jobs categorized as typically masculine. The results were consistent with others reported in this review, but their hypotheses were only partially supported: (a) male and female interviewers made different decisions, and (b) interviewers based their decisions on cues (e.g., sex, gender congruence with the job, etc.) other than applicants' qualifications. There also was evidence that interviewers'

attributions, particularly ability, were related to applicants' suitability ratings. Interestingly, there were different attributions for male and female applicants (particularly by male interviewers) on effort, luck, and task ease. Most notable was that effort attributions by male interviewers were higher for females in nontraditional jobs (perhaps due to a perception that any female who can survive in a masculine job must be working very hard at it). However, perhaps due to a poor manipulation for ability, there were no sex-based differences on ability -- the attribution most highly related to interviewers' decisions.

In a study specifically investigating the effect of applicants' attractiveness on interviewers' decisions, Gilmore, Beehr, and Love (1986) found recruiters would have hired females significantly more often than males, and students would have hired males and females with equal frequency. Although this effect is in the opposite direction of other studies on sex bias, the important findings were that (a) there was a sex effect for professional recruiters, and (b) recruiters and students did not yield the same research results. The latter of these was inconsistent with previous results (where students were more lenient) when professional recruiters and students were compared on psychometric characteristics of their ratings (cf. Arvey & Campion, 1982; Dipboye, Fromkin, & Wiback, 1975). One plausible explanation of why professional recruiters favored female applicants is that the recruiters may have yielded to some (intrinsic or extrinsic) goal of affirmative action. It would have been an interesting hypothesis to test with these interviewers and a sample of minority and majority applicants. But it would not have been relevant for selecting effective employees.

In a study of questioning strategy, McDonald and Hakel (1985) found applicants' suitability accounted for most of the variance in interviewers'

ratings, but race and sex effects both were significant. These results contradicted those of previous studies of race and sex effects (cf. Arvey, 1979), but both were negligible.

Among the results of an investigation of nonverbal cues, Parsons and Liden (1984) reported nonverbal cues of black and white applicants were perceived differently as were those of male and female applicants. Also, female interviewers gave higher overall ratings than male interviewers, but only eight interviewers were studied. Contrary to earlier studies, white applicants were given better ratings than black applicants, and female applicants received higher ratings than male applicants. The results depended on whether effects of nonverbal cues were removed from overall ratings. But cue ratings were obtained from interviewers, and probably shared bias with overall ratings. Their findings were interesting in light of studies where applicants' qualifications more important than sex, interviewer or applicant race, or the expectations of interviewers in explaining interviewers' decisions (cf. Arvey, 1979; Mullins, 1982; Powell, 1986). Parsons and Liden (1984) found nonverbal cues were strongly related to how interviewers rated applicants' qualifications, even after statistically removing the effects of objective biographical information obtained with a written instrument. They also concluded that nonverbal cues contributed to (or perhaps arose from) halo. However, the most influential nonverbal cues were associated with stable dimensions (e.g., speech patterns or verbal skills) instead of changeable characteristics (e.g., clothing, cleanliness). There were differences in how male and female interviewers perceived male, female, black, and white applicants. Since interviewers' and applicants' attributes affected the outcomes of interviews, nonverbal cues should be controlled, at least initially, when applicants' qualifications are most important (i.e., the social atmosphere is

not relevant to performance on the job). Obviously, this could be accomplished with a written instrument, such as the BSI.

Arvey and Campion (1982) reviewed nine studies of nonverbal cues involving eye contact, smiling, body posture or orientation, distance from interviewer, head nodding, or speech and vocal characteristics. They concluded that nonverbal cues influenced interviewers' decisions, and posited that nonverbal cues had less influence on final decisions than did verbal content. They believed nonverbal cues and verbal skills confounded the results because interviewers probably considered them redundant.

Rasmussen (1984) directly addressed the question of whether nonverbal cues, verbal content, and resume credentials effected interviewers' decisions. Resume credentials had the greatest impact on student interviewers. However, he did not consider that resumes may have been particularly salient for students, a plausible explanation for his findings. How applicants' nonverbal behaviors influenced interviewers depended on verbal content: With positive verbal content (i.e., the script used by the videotaped applicant/actor presented a nearly ideal candidate), more nonverbal behaviors (e.g., eye contact, smiling, gesturing with hands, and nodding the head) were associated with higher ratings of applicants' suitability. With poor verbal content, nonverbal behaviors were negatively related to ratings. But resume and verbal content always influenced interviewers much more than did nonverbal cues. Rasmussen (1984) concluded that the effect of nonverbal cues was secondary to that of cues with more face validity, but they were still important.

Gifford et al. (1985) used Brunswick's lens model to study relationships among recorded nonverbal cues, applicants' self-reported social skills, applicants' self-reported career motivation, and interviewers' ratings of social skills,

motivation, and hireability. Interviewers used the same nonverbal cues to infer social skills, that also differentiated applicants' self-reported social skills. But the nonverbal cues they used to evaluate motivation were not the same ones that explained applicants' self-reported motivation. Nonverbal cues accounted for about 50% of the variance in self-reported applicant skills and motivation, so they were useful for inferring applicants' opinions, if not their actual traits. But the correlations between interviewers' ratings of social skills and ratings of motivation (.85) and self-reported social skills and self-reported motivation (.28) suggested that interviewers intuited a greater connection between the two than might exist.

In a similar study, Beehr and Gilmore (1982) looked at the effects of applicants' attractiveness on interview outcomes. Specifically, they followed a study by Cash, Gillen, and Burns (1977), in which ratings of employability and suitability were related to attractiveness and the sex-congruence of jobs, but not the decision to hire. Beehr and Gilmore (1982) controlled gender effects, and varied how attractive applicants were and whether attractiveness was relevant to interviewers' ratings of whether to hire applicants and predicted job performance. Interviewers rated attractive candidates significantly higher when the job required attractiveness than when it did not, and they rated attractive candidates higher than unattractive applicants when attractiveness was required. So when there were attractive and unattractive candidates, and when the job required attractive workers, interviewers correctly used that information.

Gilmore et al. (1986) studied the effects of attractiveness on students and professional recruiters who interviewed male and female applicants. As noted above, there was an interaction involving kind of interviewer and applicant's gender. Unlike the results of Beehr and Gilmore (1982), attractive applicants

consistently were rated higher on whether they would be hired, predicted job performance, and perceived personality dimensions relevant to the job (but not some general personality rating). There were no effects for kind of job or the interaction between attractiveness and kind of job, so attractiveness unduly influenced interviewers' decisions when attractiveness was irrelevant.

Forsythe, Drake, and Cox (1985) studied sex stereotypes, job prototypes, and nonverbal cues. They investigated whether interviewers' decisions depended on female applicants' style of dress. This study was limited in that it used videotaped presentations (16 presentations of four applicants in four costumes) and had no verbal content (i.e., the audio track was not presented). However, up to a point, higher ratings were associated with more masculine clothing. But there was a main effect for person modeling the clothes ($p < .01$). So, regardless of the style of dress, without any verbal content or sound, with all models performing similarly, the person significantly affected the hiring decision. Unfortunately, the authors did not control that effect or the possible interaction between person and clothing.

In similar studies, Baron (1983, 1986) investigated the effects of self-presentation through nonverbal cues. In one (Baron, 1983), male interviewers rated applicants lower who wore scents (females wore perfume and males wore cologne) than those who did not. They also gave lower ratings of intelligence and friendliness (two dimensions consistently regarded as measured well in the employment interview) to applicants who wore scents. Female interviewers, however, rated differently. Both sets of interviewers rated applicants who wore scents as better dressed, but male interviewers liked them less, and female interviewers liked them more. Male interviewers reported they were better interviewers in the absence of scents and were affected more by the grooming

and personal appearance of applicants, so he proposed male interviewers were less able to ignore such irrelevant nonverbal cues and might have been irritated by the their use, because it distracted them. To control other nonverbal cues, Baron (1986) manipulated both the use of scents and other nonverbal cues (i.e., smiling, eye contact, body posture), but used only female applicants. When used alone, either scents or nonverbal cues increased the assigned ratings. However, when used together, male interviewers rated applicants lower than in any other condition (although significantly different only from certain conditions). Baron (1986) concluded male interviewers were less adept than female interviewers at ignoring certain nonverbal cues, males were aware of this and realized the consequences, so they reacted more negatively when they perceived them.

Research on similarity effects, like that on the effects of sex, race, and nonverbal cues, may be described as a form of stereotyping. For example, in the absence of adequate job information, interviewers may compare their stereotype of an applicant to the most available prototype -- themselves. Schmitt (1976) reviewed five studies extending Sydiaha's (1962) study of empathy (similarity of attitudes). Interviewers consistently rated applicants with attitudes similar to those of the interviewer higher than applicants with dissimilar attitudes. He suggested these effects were resistant to additional information. However, similarity effects appeared to be specific to individual interviewers.

Orpen (1984) studied the relationships among interviewers' and applicants' attitudes (as measured by the Survey of Attitudes Scale), and interviewers' predictions of applicants' attitudes). A total of 614 real applicants were interviewed by one of 24 interviewers. The interviewers selected candidates for real jobs. The results supported his hypotheses: Interviewers hired those

applicants they liked, those the interviewers assumed had attitudes similar to their own, and those who did have similar attitudes. However, his study did not investigate causality; his hypotheses were not the only ones supported by the evidence.

A smaller study (five interviewers and 37 applicants) by Dalessio and Imada (1984) used a structured board interview (i.e., a structured interview where more than one interviewer is present). They asked how appealing were (on a scale of one to nine) each of seven college majors, ten personality traits, eleven interests, and six preferences. They measured each interviewer's ideal prototype, each interviewers' perceptions of each applicant's attitudes, and each interviewers' self-reported attitudes. There were no similarity effects for interviewers' and applicants' attitudes. They concluded interviewers compared applicants' perceived attitudes to the inferred attitudes of their ideal worker prototypes, instead of their own attitudes. However, there were two questionable methodological issues. First, with only 37 subjects, the large number of correlations may well have yielded spurious results. Second, the researchers never measured the attitudes of the applicant. There may be a large difference between assumed and actual relationships (cf. Gifford et al., 1985). Interviewers might have rated those applicants higher who were similar, regardless of whether they accurately identified specific similarities, as they were required to do in that study.

In summary, the decisions of interviewers were affected by the verbal content of applicants' responses, applicants' resumes or applications, and a plethora of nonverbal cues. The first two (verbal content and applications) are relevant primarily for determining an applicant's qualifications. Social skills, work motivation, and intelligence (the three dimensions that appear to be

measured in employment interviews) may be inferred from applicants' experience or their responses to behavioral questions, and social skills also may be inferred from nonverbal cues. However, many nonverbal cues (e.g., perfume or cologne, posturing, inflection, etc.) might be applicants' attempts to create an advantageous social atmosphere.

The BSI studied later in this research took advantage of the constraints inherent in a written instrument. It was expected to reduce the salience of information about gender (to some extent), race, and national origin; thus, should have reduced subgroup bias. It also appeared useful for inferring social skills, because studies have shown that they are cued by information other than nonverbal behavior. It standardized (as much as possible) the order in which applicants read and responded to questions, and the order in which their responses were presented to judges for rating.

It is not surprising that the advantages of written instruments parallel researchers' suggestions for improving the employment interview. Their suggestions generally are related to standardizing interviewers' questions and perceptions, and applicants' responses. They are summarized in the next section.

Suggestions for Improving the Employment Interview

As presented above, employment interviews have poor validity and reliability because (a) they are appropriate only for certain dimensions; (b) information gathered depends on the dynamics of the interaction between interviewers and applicants; and (c) decisions of the interviewers depend on the relevance of information collected and recalled, and on interviewers' beliefs about applicants' future behavior. So advice offered by previous researchers has been directed to answering the following three questions:

1. What dimensions should be assessed?
2. How can the needed information be collected?
3. How can decisions based on employment interviews be improved?

On the first of these questions there has been a virtual consensus. Researchers have concluded the dimensions appropriately evaluated in the interview are intelligence, social or interpersonal skills, and work motivation (cf. Arvey & Campion, 1982; Wagner, 1949). Although standardized tests probably assess intelligence better than any other method, interviews may be adequate for screening prior to testing (if the testing program is expensive or limited), or for differentiating applicants after screening. Wagner (1949) championed the idea of limiting the employment interview to certain dimensions: "[M]ore and more of the factors that were formerly measured in the interview can be handled better by other means. The interview should be confined to evaluating factors that cannot be measured better by other means" (p. 43). Except for some of the self-help authors (e.g., Half, 1985; Fear & Ross, 1983), most researchers have echoed Wagner's (1949) advice (e.g., Arvey, 1979; Arvey & Campion, 1982; Schmitt, 1976), and have studied only those dimensions. However, some have assumed general personality factors are measured in the interview, and have used those dimensions that interviewers implicitly ascribe to certain occupations (cf. Jackson et al., 1982; Jackson et al., 1980). In general, research dating back to the early 1900s has not supported such general personality assessments in the employment interview. The early research established that even general suitability ratings were unreliable and of little value (Mayfield, 1964).

In summary, improving the employment interview should start with determining which dimensions are relevant, and limiting the dimensions assessed

to some specific subset for which the interview is an adequate measure. Only then can predictive validity be evaluated.

A second way to improve the employment interview -- collecting consistent, relevant information -- has been addressed in specific studies, which focused on improving interrater reliability for two reasons. First, the reliability of any procedure for gathering data is directly related to the upper limit of the validity of that information for predicting a criterion (e.g., future behavior, job performance, or tenure). If more reliable information predicted future behavior, then its observed validity should be relatively closer to its true (population) validity. Second, high interrater reliability suggests that similar information was gathered by different raters, and lends credence to their conclusions. Wagner (1949) proposed how to increase reliability and validity:

> [The] interview, regardless of its length or purpose, should be conducted according to a standardized form. . . . [to prevent] aimless rambling, lengthy digressions, and the possibility of omitting important areas. . . .
>
> . . . . . .
>
> The interviewer is no longer expected to possess a special ability which enables him to make a valid estimate simply by looking at an applicant or by talking with him a few minutes. . . . If the interviewer is to be capable of making any valid predictions or evaluations he must have at his disposal every possible bit of information. (pp. 42-43)

Mayfield (1964) also concluded material is not sufficiently covered in unstructured interviews. Although *intra*interviewer reliability was adequate, and interviewers used consistent techniques across applicants, *inter*interviewer reliability was too low. This, along with the finding that interviewers inadequately covered topics, led him to conclude that standardized interviews would increase interinterviewer reliability. The kind of material covered best was factual, biographical information -- that which is probably collected best by a biodata inventory. He also concluded that the form of the question affected

the response, and if interviews are used at all, more reliable information will be collected if they are structured.

Ulrich and Trumbo (1965) claimed that Bonneau (1957) used interviewers as "systematic data-collectors" (Ulrich & Trumbo, 1965; p. 105) and yielded predictive validity of .65. Without that structure, other groups of interviewers attained validity coefficients of .33 and .42 for the same applicants and criterion, assessing the same dimension. Ulrich and Trumbo (1965) concluded their review by affirming Wagner's (1949) position: "[T]he highest validities and the greatest gains in validity over other predictors involved interviews described as systematic, designed, structured, or guided" (p. 112). But they questioned whether the same predictive validity might not be attained with less costly measures. Denton (1964) concluded that the interview's predictive validity could be maintained or improved, and its cost could be reduced, by adapting it to a written format (essentially giving it maximum structure). He believed that the utility of his instrument would exceed that of the interview, because they measured the same dimensions (he used two raters to judge applicants' responses on four dimensions).

Suggestions in the third area, improving interviewers' decisions, have centered on structuring and standardizing interviews. Wright (1969) supported the use of structured interviews: He called for more research to make structured interviews less expensive and more useful. He concluded that structuring interviews reduced interviewers' bias. Schmitt (1976) suggested that structured interviews were more reliable because they standardized the sequence in which information was presented, also standardizing primacy and recency effects. They probably caused the interviewer to make numerous decisions, and use more of the available information (cf. Belec & Rowe, 1983). Schmitt (1976)

concluded that structured interviews increased reliability by forcing interviewers to record information, making it available later. This might have reduced or eliminated order effects, contrast effects, or excessive weighting of certain information. He suggested that allowing applicants to talk more provided a larger sample of behavior to judge. Perhaps structured interviews are more reliable and valid because they constrained interviewers' influence and allowed applicants to offer more information.

Schuh (1973) found the content of the rating form used in structured interviews influenced interviewers' decisions. Specifically, experienced interviewers made essentially the same decisions whether the form emphasized important or unimportant information. Novice interviewers made the same decisions as experienced interviewers when they used the form that emphasized important items, but different decisions when they used the form that emphasized unimportant items. Arvey and Campion (1982) cited other (inaccurately referenced) work by Schuh in which interviewers who were not interrupted and took notes had higher listening accuracy (79%) than those who were interrupted and/or did not take notes.

Latham et al. (1980) studied the situational interview -- a variation of the standardized interview -- in which interviewers read examples of situations relevant to the job (i.e., critical incidents), and rated applicants' responses on behaviorally anchored scales. They developed the situational interview to force interviewers to rely on the stated intentions of applicants, instead of past behavior (cf. Ghiselli, 1966), to predict future performance. Critical incidents and behavioral anchors were developed from examples of behavior that

interviewers believed discriminated performance levels among workers.[5] The results were as expected: interrater reliability was greater than .70, and all validity coefficients exceeded .30. Decisions made from information gathered in the situational interview explained two to three times more variance in job performance than did information from standard interviews (e.g., 11% vs. 4%). Latham and Saari (1984) investigated whether the intentions stated by applicants were related to subsequent behavior on the job. Responses to questions in the situational interview were significantly correlated with behavior observed by peers and supervisors. The responses accounted for the same variance in job behavior as did questions regarding experience and formal training, in addition to a significant amount of unique variance. A predictive study of the situational interview, reported in the same article, produced a validity coefficient of only .14 ($p < .05$). When the researchers interviewed the interviewers, they found the interviewers had interviewed incorrectly. Instead of making a series of specific, behavior-related decisions, they used the interview structure as a guide to form an overall impression of applicants. When they reinterviewed a random sample of those hired correctly using the situational interview, the validity coefficient was .40. So in spite of the conceptual closeness of the predictors and criteria, there was a difference between the situational interview and the inadvertent guided interview. In all four of these studies (i.e., Latham & Saari, 1984; Latham et al., 1980), significant validity coefficients were obtained with small sample sizes, suggesting that effect sizes were large (see Table II-1). Standardizing the

---

[5]The success of these items was important when deciding how to create items for the BSI: This study illustrated that interviewers chose valid behaviors to assess, instead of having to use the shotgun approach to develop items. It was also important that interviewers were familiar with the content of the items.

Table II-1. Validity Coefficients for the Situational Interview as Reported in Latham and Saari (1984) and Latham, Saari, Pursell, and Campion (1980)

Study	df	Validity Coefficient	Significance Level	Kind of Study
Latham et al. (1980)	47	.50	.00	Concurrent
Latham et al. (1980)	61	.30	.00	Concurrent
Latham et al. (1980)[1]	28	.39	.00	Predictive
Latham et al. (1980)[2]	54	.33	.00	Predictive
Latham and Saari (1984)	27	.40	.00	Concurrent

[1]Female applicants.

[2]Black applicants.

employment interview and constraining interviewers to specific questions resulted in prediction comparable to that of other, more highly regarded, predictors.

After finding that exposure to applicants' application forms prior to interviewing reduced interinterviewer reliability, Dipboye, Fontenelle, and Stramler (1984) recommended that interviews be standardized. They contended that previewing the application form caused the interviewer to gather more nonapplication information. However, they believed that the additional information decreased interinterviewer reliability, because different interviewers made different decisions with the same information. So standardizing the interview would limit nonapplication information to that which is valid or manageable, maintain adequate reliability, and increase validity.

Janz (1982) compared structured interviews that were similar to situational interviews, to unstructured interviews, and he compared the behavior of interviewers and applicants. Validity coefficients were significantly higher for structured interviews than for unstructured interviews (.54 vs. .07). However, the statistical method used to assess validity was questionable. Also, interinterviewer reliability was lower for structured than unstructured interviews. There were more questions and answers about experience and intended behavior, and fewer about credentials and self-perceptions, in the structured interview. The results favored the hypothesis that structured interviews were more valid, but the interviewers using the structured interview received more training than did interviewers in the unstructured condition. The differences could have been caused by the different training received.

Orpen (1985) used the same interview format as Janz (1982) but administered equal amounts of training to interviewers in the structured and

unstructured conditions. Orpen (1985) used global supervisory ratings and sales (in dollars) as criteria, unlike Janz (1982), who used criteria that were similar to items on the interview. Orpen (1985) found the validity coefficient of the structured interviews exceeded that of the unstructured interviews, despite the differences between predictors and criteria. These studies suggested two further ways to improve the interview: Train the interviewers and increase interviewers' knowledge about the job for which applicants are being interviewed. Researchers who successfully structured interviews explicitly tied behavioral aspects of the jobs to interview questions, and administered extensive training to interviewers. However, these studies also showed that the most important part of the training was having a specific instrument or structure to teach. Closely tying the structure of the interview to job information and criteria is critical. This was supported by Arvey and Campion (1982): Training focused on increasing correspondence between interviews and job gave better results than training to reduce psychometric errors (e.g., halo, leniency, etc.). Schmitt (1976) reviewed studies in which interinterviewer reliability increased and the use of irrelevant dimensions decreased as interviewers received more job information. Neither Arvey and Campion (1982) nor Schmitt (1976) found consistent evidence that professional or experienced interviewers were more reliable or more valid than novices.

Dougherty, Ebert, and Callender (1986) found training effects when they modeled the decisions of three interviewers. After training, interviewers' employment decisions correlated less with test scores or application blank ratings, and more with ratings related to interpersonal skills and motivation to work (the two dimensions that the researchers believed were appropriately measured in the interview).

In summary, increasing the structure of employment interviews, and increasing interviewers' job knowledge appear to be the most useful ways of improving the validity of employment interviews. These methods probably worked because they increased the relevance of information extracted by interviewers, and insured that all interviewers extracted similar information in every employment interview.

Interview Summary and Conclusions

This review has shown employment interviews were rife with problems, whether conducted by experts or novices. They were valid for poorly defined dimensions -- interpersonal skills and work motivation -- and are often misapplied. Their costs were high, and their utility low. Reviews have been disparaging, at the least. Interviewers could not collect data as well as applications could. Employment interviews were shown to magnify the idiosyncracies of human judgment: Biases, stereotypes, and errors of judgment abounded. Even appearance, gestures, and perfumes influenced employment decisions.

But employment interviews can contribute unique information to employment decisions. Assessing motivation and social skills probably can not be accomplished better by other means, so interviewing has a place in the employment process. But interviews can be improved: by constraining interviewers to ask certain questions relevant to the job; by training interviewers to ignore irrelevant stereotypes and nonverbal cues; by forcing interviewers to make decisions throughout the interview; by fixing the order of questions and topics; by collecting biographical and application information with written instruments; by allowing applicants to express themselves, all of which can be accomplished by using structured or situational interviews. However, it

was shown earlier in this review that little may be gained from face-to-face interviewing: Even social skills were inferred from the verbal content of applicants' responses during interviews. The suggestions of researchers, that interviews can be replaced by written instruments, has been supported by the evidence reviewed here (e.g., Denton, 1964; Dunnette, 1961; Wagner, 1949; Ulrich & Trumbo, 1965; Mayfield, 1964).

From the results of studies he reviewed, Wagner (1949) suggested that interviewers might be capable of discovering more information than can be coded for an empirical validation. But he questioned whether subjective integration of more information would be as valid as an empirical prediction based on a limited amount of information: "The question which arises is whether the interviewer can do as well as or better than statistical procedures" (Wagner, 1949; p. 24).

The BSI is an instrument with items written by an experienced interviewer (cf. Latham et al. 1980). The items request behavioral and attitudinal information that is relevant to the job, and familiar to interviewers. The BSI capitalizes on the findings and conclusions of previous researchers, and it is scored by a combination of statistical and clinical methods: It is essentially a biodata inventory developed to simulate a structured interview.

Biodata Inventories

Some evidence of the reliability and validity of biodata inventories is presented in this section, followed by an overview of how they are developed and scored. Three aspects of reliability have been investigated, including (a) how accurate and truthful are responses about factual or verifiable information, (b) the test-retest reliability of items, and (c) interobserver reliability (i.e., how

often respondents and other individuals agree regarding respondents' biographical history).

Owens (1984) noted that items usually included in biodata inventories were not subject to contamination due to response sets. For example, typical options for responses do not appear to be more socially desirable than others (Nunnally, 1970). Factual or historical information is required by most items, and there is little opportunity for the respondent to adopt an acquiescent response set. Responses involve the recollection of events and should be stable,[6] unless the respondent's memory fails (Nunnally, 1970) or an item is threatening.

Mosel and Cozan (1952) were the first researchers to report a systematic check of how accurate responses were on application forms. They questioned 61 men and 65 women about how much they made per week, how long they had been employed, and what their job duties were. They requested the same information from former employers. They did not inform applicants of their verification plans, nor did they tell employers what any applicant had reported. Employers' and applicants' information were compared for the previous one or two years. Correlations between employers' and applicants' responses for weekly salary and duration of employment, and the fraction of employers and applicants agreeing on reported job duties were as follows:

	0-12 months		13-24 months	
Item	Males	Females	Males	Females
Weekly salary (*r*)	.93	.94	.91	.90
Duration (*r*)	.98	.98	.97	.87
Duties (fraction)	.95	.94	.84	.87

[6]The BSI (studied later in this research) contained some items that measured attitudes. However, few of them should have been perceived as threatening. There also were some true/false questions that might have encouraged response sets.

Applicants usually overestimated their experience, salary, or job duties, if they disagreed with their previous employers. Cascio (1975) verified 17 items on a biodata inventory using previous employers' reports. The median correlation was .94 between what applicants and employers reported, relatively good agreement. Only two items had correlations of less than .66 (age at first marriage, and kind of high school attended longest). Goldstein (1971) found a much larger proportion of discrepancies than did others. For example, 57% of the time applicants and employers disagreed on how long applicants had been employed, and many disagreed on how often and how much they were paid. In 63 of 94 complete sets of data, they disagreed on two of the four kinds of information collected (i.e., what their job was, how long they had been employed, how much they earned, and why they left). The real state of affairs was not known, and such disagreement could have been coincidental. But Mosel and Cozan (1952) found applicants' responses favored themselves when there were inaccuracies. Applicants (a) made favorable errors when recalling previous jobs, or (b) exaggerated the status of previous employment intentionally.

Shaffer, Saunders, and Owens (1986) pointed out that inaccurate responses might not affect the validity of empirically scored instruments. Respondents' behavior in completing the instrument could be more valid than the content of their responses. The predictive purposes of an inventory might be served adequately if the responses are reliable. However, response accuracy is important if the purpose of the instrument is to gain insight or understanding of applicants' behavior.

Two ways to assess the accuracy of nonverifiable responses are (a) compare responses to the same instrument administered more than once to the same group of applicants, and (b) compare the responses of one or more people who

know each applicant to those of the applicant. Shaffer et al. (1986) reported the only recent study in which both test-retest and interobserver reliability were investigated specifically. They readministered Owens' Biographical Questionnaire (BQ) to subjects who had completed it five years earlier when in school (Owens & Schoenfeldt, 1979). They computed test-retest reliability by correlating responses collected at the first and second administration, and interobserver (actually subject-observer) reliability by correlating subjects' responses with parents' responses to a shortened version of the BQ. This study yielded the following results (among others):

1. Means of many of the more subjective items were significantly different between the two times (less accuracy), but most of the test-retest correlations were substantial.

2. Biodata items were answered accurately, because there were relatively good correlations between parents' and subjects' responses (particularly on the less subjective items).

3. Respondents reported objective information more consistently over time than subjective information. Even moderately subjective information had acceptable levels of reliability.

4. Consistency over time was interpreted as accuracy rather than consistency in distortion, because items judged as having socially desirable responses were answered less consistently over time.

5. Socially desirable responses might decrease the validity of biodata items enough to make it worthwhile to evaluate them, and then eliminate or rewrite them if necessary.

In summary, applicants reported nonverifiable information reliably,

particularly for such a long time period (five years was believed to be long enough to limit the effects of memory on consistency of response).

Crandall (1976) assessed the interobserver reliability of items on an instrument measuring quality of life (e.g., ratings of how much beauty one sees in the world, or how free one is from bother and annoyance).[7] The mean correlation between redundant items was .71. The mean correlation between average ratings of two knowledgeable others (for each respondent) and primary respondents was .31, while the mean correlation between the two knowledgeable others was .31. This was much more reliable than might be expected for research done in the field instead of the laboratory, with extremely subjective items, and knowledgeable others who were not necessarily family members.

In summary, evidence for the reliability of biodata inventories is consistent. Reliability and accuracy appear to be adequate even for nonverifiable items or subjective items.

Researchers found consistently that biodata inventories were valid for predicting many different criteria in a wide range of populations. Asher (1972) reviewed studies published between 1960 and 1970 validating verifiable and historical biodata items (i.e., dealing with events in the respondents' backgrounds that could have been verified). Such items were useful (i.e., their validity coefficients were statistically and practically significant) for predicting criteria such as adolescents' creativity, salespersons' success, and professionals' choices of careers. Of the 31 crossvalidated coefficients he reviewed, 30 were larger than .30, and 17 were larger than .50. He concluded that biodata

---

[7]Asher (1972) referred to these kinds of items as *soft*, "because (a) they cannot be verified, and (b) the options are expressed in abstract value judgments rather than realistic behavior . . ." (p. 263). Hard items are historical and verifiable.

inventories had a dramatically better record than employment interviews for predicting job proficiency.

Asher (1972) summarized three theories about the excellent validity coefficients reported for biodata inventories. First, items (as he defined them) on biodata inventories ask for factual (often verifiable) data, and present less opportunity for applicants to exaggerate or deceive employers (probably less than most employment interviews). Second, only items that are important when crossvalidated are included. Third, biodata inventories usually measure reports of behavior or events similar to the criterion being predicted. They do not measure dimensions hypothesized to cause criterion behavior. For example, grade point average in high school would be used instead of an intelligence test to predict grade point average in college. "In comparison with other predictors as intelligence, aptitude, interest, and personality, biographical items had vastly superior validity" (Asher, 1972; p. 266) for predicting employment criteria. Asher (1972) generalized his conclusions only to biodata items that asked for verifiable and historic data. He made no recommendations about items measuring attitudes, beliefs, or personality characteristics.

Reilly and Chao (1982) reviewed eight alternatives to standardized testing for employment selection. Only biodata inventories and peer evaluation exhibited validity coefficients equal to standardized tests. Of the two, they judged biodata inventories as more practical in a wider variety of situations. Average crossvalidated validity coefficients for various criteria ranged from .32 to .46, with a mean of .35. They reviewed studies of adverse impact and sex differences, and concluded that the "validity and fairness of biodata can be expected to hold for minority and majority groups, but different keys may be needed for males and females" (p. 13). Biodata inventories were cost effective

and easily administered; interviews may be used to collect the same data, but they had much poorer validity. Reilly and Chao (1982) concluded that a "biodata [inventory] would appear to be a recommended alternative to include in a validity study" (p. 53), in spite of some unanswered questions about how they are developed and their potential for adversely affecting minorities or females.

Hunter and Hunter (1984) performed a metaanalysis of studies of many predictors and found biodata inventories were more valid than interest inventories, employment interviews, college grade point average, or reference checks, for predicting a range of criteria. For example, biodata inventories accounted for about seven times as much criterion variance as the employment interview (i.e., 14% vs. 2%). Only two predictors had better support -- ability tests (i.e., some appropriate combination of standardized psychomotor and cognitive abilities instruments) and job samples and work tryouts. The latter two consist of job duties performed temporarily and have nearly perfect point-to-point correspondence with the job. When assessing the economic utility of the various predictors, they concluded that biodata inventories were second only to a composite of ability tests.

Schmitt et al. (1984) found adequate validity coefficients for biodata inventories in their metaanalysis of various predictors and criteria. The mean was .24 for 99 validity coefficients. While researchers usually have attributed differences among validity coefficients to sampling error (cf. Hunter & Hunter, 1984), Schmitt et al. (1984) found that 92% of the variance in the coefficients they summarized remained after correcting for sampling error. Different criteria, different development methods, and different scoring strategies probably resulted in different population (i.e., unobservable true) validity coefficients. In other words, there may be more than one true coefficient of

validity for biodata inventories, because they can be developed and scored many different ways to predict many different criteria.

Biodata Development and Scoring

In spite of impressive validity results, biodata inventories often are criticized for how they are developed and scored (Reilly & Chao, 1982). In most cases, numerous items are created based on intuition, introspection, prior research or theory, or availability. The purpose is to find items that can discriminate among groups differing on some criterion (e.g., succeeded-failed, successful-adequate-unsuccessful, stayed-quit, etc.), and score them to make valid predictions. Methods vary in the way specific items are retained and scored. Hornick, James, and Jones (1977), following Goldberg (1972), grouped procedures for developing and scoring biodata inventories into four classes: empirical, internal, intuitive, and rational. Each is described below. For all methods, items are generated from relevant theory, by intuition or introspection, or by taking items from existing forms (e.g., an application blank).

Empirical. In the empirical procedure, items are selected by some statistical method, such as multiple regression, zero-order correlations, or traditional item analysis. Items are retained and scored if they reach a minimum level of validity upon crossvalidation. The empirical procedure promotes internal heterogeneity, and the resulting scoring key has no construct validity or interpretable structure (unless by serendipity). However, items selected with this method have the lowest colinearity of the four classifications, and perhaps the highest validity and generalizability (particularly for complex composite criteria).

Internal. This classification was so named because items are selected by analysis of the moment or variance-covariance matrix of items. Most commonly used are the factor analysis methods to promote internal consistency. Items selected this way are more homogeneous than those selected with the empirical method, and estimated validity coefficients may shrink less when the items are crossvalidated. But lower coefficients may result because unique variance is discarded and there is greater colinearity among items.

Intuitive. Items are selected in the intuitive method because theory (or postulation, intuition, or experience) indicates that they should predict the criterion of interest. The inventory is neither developed nor scored empirically. It is only as predictive as is the intuition or introspection of the developer. In other words, a researcher using the intuitive method simply uses the incipient inventory and may subject it to a summary statistical analysis (e.g., an overall validity coefficient). Based on the fallibility of expert judgment in making decisions (Meehl, 1954; Reilly & Chao, 1982), this procedure should result in the lowest validity of all the methods. However, Goldberg (1972) and Hornick et al. (1977) did not find lower validity for items developed intuitively.

Rational. The rational method is a combination of the intuitive and internal methods. An intuitively developed inventory is subjected to an analysis of the internal variance of the items. Those with low intercorrelations or inconsistent correlations are dropped.

In practice, the rational method is used in place of the internal method. By definition, the intuitive method does not lend itself to scoring investigation. Therefore, only the rational and empirical methods as described by Hornick et al. (1977) were considered for this study. With the rational method, development and scoring are performed essentially without benefit of criterion

results: the intent is to create item "composites for each construct with high internal consistency" (Hornick et al., 1977; p. 490). However, this could result in an internally consistent instrument that would not predict the criterion of interest (i.e., an invalid inventory with excellent psychometric qualities).

The actual steps in developing an inventory are straightforward. For empirical scoring, a pilot inventory made up of many items is completed by 150 or more subjects (England, 1961).[8] Item responses are usually multiple choice or easily classified, so they can be grouped into four or five categories for each item. Each category of every item is then analyzed to see if it discriminates among subjects in the scoring sample. Weights are assigned to discriminating items, and a composite score may be created for each subject by summing across the selected items. Finally, responses of the crossvalidation sample are scored to estimate the population validity coefficient of the biodata inventory. There are many variations on these steps, and comparisons have been conducted (e.g., Telenson, Alexander, & Barrett, 1983). The rational scoring method differs from the empirical method in that the items' intercorrelations are analyzed instead of the actual responses. As in the empirical method, a composite score is derived and used to predict a criterion. However, the rational method does not involve finding an optimal fit of items to criterion; instead it is concerned with selecting items measuring the same factors (or dimensions). Theoretically, high internal consistency is a sign that the selected items are measuring whatever the developer hoped they would measure. This homogeneous biodata inventory is then validated to see if it predicts the criterion. Responses of the crossvalidation sample are scored to

---

[8]About 25% to 50% of these subjects are reserved for crossvalidating the inventory (England, 1961). The remaining subjects comprise the scoring sample, providing responses for initial item selection and scoring.

estimate the population validity coefficient of the biodata inventory. Thus, both methods require the use of empirical relationships and rational decisions.

Dunnette (1961) discussed criticisms of the empirical method. He noted that a biodata inventory was "a highly effective selection tool" (p. 292), but empirical approaches to prediction completely ignored theoretical issues. He agreed with Toops: Ignoring the "logical linkages and inferred causal relations" (Dunnette, 1961; p. 293) lends nothing to understanding the developmental stages of successful job incumbents, but serves only the function of predicting job performance. This may, perhaps, be one reason why biodata inventories have to be revalidated periodically (cf. Hunter & Hunter, 1984). However, if elements of successful behavior could be specified, then a biodata inventory could be developed and scored to correspond point-to-point to the criterion (Asher, 1972; Pannone, 1984). Of course, that is the goal of rational development and scoring. Hornick et al. (1977) argued that purely empirical methods only hindered understanding and construct validation. But they agreed that the rational approach does not assess construct validity unless multiple criteria of the construct are used, and it discards unique variance that might predict the criterion. In their comparison of empirical and rational methods, there were no significant differences between crossvalidated validity coefficients for predicting two sets of performance ratings from a psychological climate questionnaire. Items selected empirically did not belong to any particular content domains, confirming their expectations.

Following Hornick et al. (1977), Mitchell and Klimoski (1982) compared rational and empirical approaches. The validity coefficient for the empirical biodata inventory shrank substantially when crossvalidated, but remained significantly larger than the crossvalidated validity coefficient for the rational

biodata inventory. They concluded the empirical method should be used for prediction, but tempered that conclusion with three contingencies: First, the practical difference between validity coefficients was slight, and the difference in utility between the two methods will depend on the situation. Second, if empirical scoring procedures are not repeated periodically, the rational biodata inventory might eventually dominate (cf. Hunter & Hunter, 1984). Finally, if items are developed haphazardly, if sample sizes are too small, or if the range of the predictors or criterion are restricted, the empirical method will suffer. However, these last problems also affect the rational method.

Frank (1980) compared an actuarial method (i.e., the traditional method of response categorization and item analysis; cf. Anastasi & Schaefer, 1969; England, 1961) and the general linear model, and found little difference between them in predictive ability. However, the actuarial methods provided better understanding of the interactions (or configurations) among dimensions implied by certain items (in his opinion). The linear model did not provide a scoring scheme that was as interpretable as the actuarial method. However, the difference between the two could have been due to the researcher's experience with the actuarial method.

In summary, empirically based methods to score biodata inventories provided adequate prediction, but did not enhance theory or understanding. Rational methods, usually developed from a priori hypotheses, depended either on empirical definition and confirmation (as in the internal method), or on the verity of intuition. The empirical method is very similar to the internal method. The only difference between them is that the internal method explicitly aggregates items into dimensions. Both methods depend on rational development and empirical selection of items. For example, Anastasi and

Schaefer (1961) were very rational and intuitive in developing the empirically scored inventory they used to study creativity.

Intuitively scoring biodata inventories is a gross empirical procedure: If the inventory is deemed successful and is used, then the entire instrument receives a weight of 1.0; but if it fails, the instrument is weighted 0.0. It is little more than unit weighting all the items, all or none. The intuitive method depends on the accuracy with which the developer can devise items that represent some relevant content area (e.g., Pannone, 1984) or stable personal dimensions. But the literature on decision making indicated that the best way to get insights about how experts use information to make decisions was not by introspection, but by decision modeling (sometimes referred to as policy capturing; see the next section). The intuitive method could provide an attractive, acceptable, face valid alternative to "dust bowl empiricism," if its scoring simulated how the developer (subject matter expert) really used the items to make decisions. In the next section, an overview of the literature on modeling decision behavior is presented to provide the foundation for developing and scoring the BSI to screen applicants for management positions.

Modeling Decision Behavior

Clinical Versus Statistical Decisions

Using statistical instead of clinical methods to combine test scores, weight individual items on inventories and instruments, and even combine clinical judgments to predict various criteria, has received overwhelming support in the literature. For example, Meehl (1954) presented the arguments for and against using clinical judgment, and reviewed available empirical studies. Although he believed clinicians could make unique predictive contributions:

> In spite of the defects and ambiguities present, let me emphasize the brute fact that we have here, depending on one's standards for admission as relevant, from 16 to 20 studies involving a comparison of clinical and actuarial methods, *in all but one of which the predictions made actuarially were either approximately equal or superior to those made by a clinician.*
>
> . . . . . .
>
> In about half the studies, the two methods are equal; in the other half, the clinician is definitely inferior. (p. 119)

He noted in his closing remarks that even when clinical judgments are to be used as predictors, they must be validated so there is evidence they are useful, if not efficient.

In two industrial applications, Dicken and Black (1965) compared the validity of psychometric tests (the Strong Vocation Interest Blank, the MMPI, and the Otis Quick Scoring Mental Ability Test, among others) to clinical interpretations of the same tests. (They compared actuarial and clinical combination of test results, not the predictive validity of tests versus expert judgments.) Individual abilities tests (the MMPI was not studied individually) correlated well with final salary, intelligence, leadership, and likableness. Clinical reports correlated well with final salary and better with individual dimensions. Unfortunately, the researchers did not combine the objective tests statistically to estimate the predictive validity of an optimal or a unit weighted composite. Other research consistently has suggested that such a composite would predict at least as well as clinical interpretations (cf. Meehl, 1954; Sawyer, 1966).

In 1966, Sawyer expanded the competition between clinical and statistical methods to include measurement, or data gathering, as well as prediction. He proposed that the ability of clinicians to gather data influenced the accuracy of their final predictions. Thus, both processes needed to be investigated before judging the predictive ability of clinicians. Meehl (1954) only reviewed studies

in which the same data were combined by the two methods, so he ignored situations in which clinicians might excel because of more subtle means of data collection. Sawyer (1966) categorized 45 empirical studies according to whether data were collected clinically (e.g., by observation or in interviews) or mechanically (e.g., standardized tests, biodata inventories, personnel records, etc.). He noted that these methods are the endpoints of a continuum, and included studies in which both methods were used. Within each of those categories, he further classified studies as to whether data were combined clinically or statistically. The results were consistent with previous studies: the statistical predictions were always as good as, or better than, the clinical predictions, regardless of how the data were collected. Within kinds of prediction, those collection methods which included some mechanical measures were superior to the purely clinical methods. The results were best for methods involving both clinical and mechanical means of data collection. "This suggests that the clinician may be able to contribute most not by direct prediction, but by providing, in objective form, judgments to be combined mechanically" (Sawyer, 1966; p. 193).

Einhorn (1972) directly compared global clinical predictions with statistically combined clinical measurements. Three pathologists examined tumor biopsy slides taken from 93 patients with Hodgkin's disease. Each pathologist rated each slide on nine important dimensions, and estimated how long each patient would survive. For each pathologist, Einhorn (1972) compared the validity of estimated survival time to that of a statistical combination of the nine dimension ratings. Validity coefficients were high for nonlinear combinations of the nine dimension ratings for each pathologist: .359 ($p < .001$), .340 ($p < .001$), .191 ($p < .05$). A statistical combination of the nine means of their ratings also

yielded a sizable correlation (.377, $p < .001$). The highest crossvalidated validity coefficient for predicting actual survival time was for a nonlinear combination of all three pathologists' nine dimension ratings (.396, $p < .001$), but the validity coefficients of their global ratings of survival time clustered around zero. This study affirmed the assertions of Meehl (1954) and Sawyer (1966): Statistical combination offers a substantial increase in predictive validity over clinical combination, even when the data are clinical judgments.

Decision Modeling

Puzzled by the fallibility of expert's decisions, researchers investigated the processes by which decisions were made, to find out how the two methods differed, and why statistical weighting was superior. They compared how experts used information (cues), to how the same information was weighted statistically. Although decision modeling is not a new field of research,[9] the advent of the computer caused dramatic growth in recent years, by making it easier to compute models (Slovic & Lichtenstein, 1971). Now there are many studies of decision behavior modeling. Slovic and Lichtenstein (1971) offered the most comprehensive review to date in their comparison of Bayesian and regression approaches to modeling decisions. Hogarth (1980) provided a good integration of research findings and corresponding theories of judgment (cf. Kahneman, Slovic, & Tversky, 1982).

The various methods of modeling most often have been traced to Brunswick's (e.g., 1955; cf. Petrinovich, 1979) call for the ecologically representative design of experiments to assess how subjects use unreliable or

---

[9]The researchers referred to here were preceded by others, most notably Wallace (1923). However, this overview is limited to those following Meehl (1954), who were motivated by the increasing number of studies comparing clinical and statistical decision making.

fallible environmental cues. The cues must be used probabilistically by subjects, because they are unreliable and inconsistently related to criteria. In practical terms, Brunswick's research paradigm (commonly referred to as the lens model) has been used to compare the relationships between cues in the environment and a criterion to the way subjects use those same cues to predict the same criterion. For example, Dudycha and Naylor (1966) applied the lens model to study decision behavior. They measured (a) the cues available to each subject, (b) the criterion value subjects were to predict, and (c) the predictions each subject made. Specifically, they regressed subjects' predicted criterion values (a two digit number) on the corresponding cues given the subjects (orthogonal pairs of two digit numbers variously related to the criterion). The subjects were given the correct criterion value immediately after each prediction, so they could learn the relationships. The researchers inferred from these data (a) how consistently each subject used the cues across similar problems, (b) how valid the cues were for predicting the true criterion values (i.e., the environmental predictability), (c) how valid subjects' predictions were, and (d) how closely the subjects' cue weights matched the true (environmental) weights. Subjects were able to learn the relationships between cues and criterion with varying accuracy, depending on the strength of the correlations; but even when they had learned the relationships, they failed to use them consistently. This prompted many researchers (e.g., Einhorn, 1972; Goldberg, 1968) to posit that statistical methods were superior because they were more consistent than individuals.

Among the early researchers who used the lens model as a guide to study decision processes were Hoffman (1960), who proposed and described linear and configural, or interactive, models of decision making; and Hammond, Hursch, and

Todd (1964) who applied similar models to study clinical judgments. Hammond et al. (1964) successfully used the lens model to predict how clinical psychologists would use Rorschach psychograms to rate patients' intelligence. Together, these and other researchers presented the initial evidence that (a) multiple regression is an effective tool for describing experts' decisions, (b) the lens model paradigm (as represented by multiple regression) can be as simple or as complex as is needed to describe experts' judgments, and (c) virtually complete descriptions of decisions are possible with simple linear models. However, a model so derived may have little in common with the actual processes or cognitive associations (i.e., the policy) of the subject. Instead, it is one possible process or a parsimonious index of how a subject arrived at a functional response or decision; at most, such a model is an approximation of the decision process. In this paper, the term decision modeling will be used instead of policy capturing, because the latter might imply more to the reader than is justified. Sydiaha (1962), Goldberg (1968), and Roose and Doherty (1976), among others, have addressed this subject cogently. The interested reader is referred to their articles for more disclaimers of the insight gained by modeling.

Goldberg (1968) summarized previous research and concluded that judges may use complex processes to make decisions. But "if one's sole purpose is to reproduce the responses of most clinical judges, then a simple linear model will normally permit the reproduction of 90%-100% of their reliable judgmental variance, probably in most -- if not all -- clinical judgment tasks" (Goldberg, 1968; p. 491). He proposed that the variance explained by adding interaction terms to a linear model of decisions rarely was a sufficient tradeoff for the lost degrees of freedom. Hakel et al. (1970; see the earlier section on making

decisions in the employment interview) found decisions of interviewers could be adequately represented by simple linear models, although the subjects apparently used different strategies.

Goldberg (1968) also addressed the question of whether models of decisions could help train judges to use available information better. He previewed a study in which judges were trained (using feedback) to interpret MMPI profiles. When judges were given the appropriate formula (including weights and cutting scores), they quickly improved their accuracy. However, he was surprised to learn that they eventually reverted to their previous levels of accuracy.

One way researchers attempted to discover the processes judges used was to ask judges how important they felt each cue was in making their decisions. For example, experts might be asked to distribute 100 points among the cues according to how important they felt the cues were (see Cook & Stewart, 1975, for a comparison of seven methods for obtaining subjective weights). Kleinmuntz (1968) even had experts describe their decisions, as they made them, into a tape recorder. The computer program he derived simulated decisions that predicted as well as the best MMPI interpreter, and better than the average interpreters.[10]

Slovic (1969) obtained ratings of growth potential on 128 companies from two stockbrokers. They based their ratings on 11 cues provided by the researcher. He chose stockbrokers as subjects because he considered their work complex, and believed that it called for complex decision processes. The two brokers' models included interaction terms, so they may have used complex strategies. He compared their subjective weights to the weights from their

---

[10]His computer program simulated a hierarchical decision tree, instead of a linear model; although it worked well, it was relatively complex.

models. As has been found in subsequent studies, the judges apparently overemphasized cues unimportant in their models and underemphasized those they relied on most heavily. The brokers' models did not weight all 11 available cues, even though they were considered by one of the brokers the minimum information required to determine growth potential.

Shepard (1964) proposed two reasons why subjective weights consistently differed from those of judges' models: First, judges may have used diverse cues at different times or in different situations, but they associate all the cues with all the situations. Therefore, they assign weights to variables they sometimes used, but did not use in the situation at hand. Second, colinearity among cues will affect their perceived importance and the weights assigned by regression procedures (in both modeling the decisions and estimating optimal weights). So the comparison will depend on the idiosyncracies of the situation and how much the cues and the criterion are intercorrelated.

Brookhouse et al. (1986) tested whether subjects believed it was more socially acceptable to report that they used a maximum number of cues when they made a decision, even though they only used a few. They asked subjects to rate their preference for jobs based on 11 job characteristics (honest ratings), and also to rate them to impress an employment recruiter (faked ratings). Then they were asked to subjectively weight 11 characteristics as they used them to rate the honest and faked preference ratings. Subjects subjectively weighted many more characteristics than they actually used in either condition. Judgments predicted from subjective weights were more closely correlated than those predicted from the decision models, because the honest subjective weights were nearly the same as the faked subjective weights.

They concluded "social desirability response bias [was] a greater factor in subjective weighting than in policy capturing" (p. 317).

Einhorn (1970) proposed that judges used information in ways that were neither linear nor compensatory. For example, a judge best described by a compensatory model would recommend employment for applicants deficient on one or more variables of interest, if the remaining variables were high enough to compensate for such deficiency. But a judge best described by a conjunctive model would recommend employment only if applicants met minimum requirements on all variables of interest. A disjunctively oriented judge would recommend employment if applicants excelled on one of the variables (e.g., football players would be hired if they performed any one function very well, without regard for other tasks).

Einhorn (1970) provided ways to transform cues to create disjunctive and conjunctive models,[11] and showed noncompensatory models had larger crossvalidated coefficients than compensatory models for predicting some decisions. It is possible that Slovic's (1969) brokers were using one of the noncompensatory models. This would explain why some of the cues the brokers perceived as important received negligible weights in their models.

Einhorn (1971) posited that conjunctive or disjunctive models may be easier for judges to use, even though they are more complex mathematically. For example, the conjunctive model implies a search for negative information and the disjunctive model implies a search for positive information (see the previous section on interviewing for a discussion of this strategy). Judges insisted that they used complex configural processes, so Einhorn (1971) investigated whether

---

[11]Transforming variables in a model does not consume degrees of freedom, unlike adding interaction terms to simple linear models. The transformed variables are used in simple linear models to estimate nonlinear relationships.

models depended on the kind and quantity of information presented. What he found is important for several reasons: First, the linear, conjunctive, and disjunctive models all were predictive of different judges when double crossvalidated; so judges' strategies (not just weights or importance) were inferred. Second, the nonlinear noncompensatory model appeared to be superior to the linear model, although this was a function of the particular task employed, thus showing that judges used complex strategies, and not necessarily compensatory ones. Third, the two tasks he used in his studies were associated with different kinds of models. His first study involved job choice, and the conjunctive model was superior. This was consistent with the notion that a job will be attractive only if all of its inducements exceed some minimum level. On the other hand, the second study involved selecting candidates for graduate school, and different judges were described best by different models. The linear model was not clearly superior nor inferior to the other models, thus tempering Goldberg's (1968) assertions.

Ashton (1974) investigated the decisions of 63 internal auditors. Their decisions were modeled well by linear models. Interactions were not important, probably because there were few cues, and they were straightforward financial indicators. The auditors used cues consistently, and they estimated subjective weights close to those of their models. Ashton (1974) attributed this consistency and insight to the professional training auditors receive and the kinds of cues they used to make judgments. Similarly, Norman (1976) found linear models adequately described how undergraduate students rated hypothetical job applicants. They based their ratings on length of work experience, score on an abilities test, and an interview rating.

Bootstrapping

Dawes (1971) was among the first to apply the derived model of a judge to a prediction task. In what is termed bootstrapping, he applied the model of a graduate admissions committee at the University of Oregon to screen applications to the Department of Psychology. His findings supported three principles of human decision making he proposed. Framed in the context of his research, they were the following:

1. "A simple linear combination of the criteria the admissions committee considers will do a better job of predicting performance in graduate school than will the admissions committee itself" (p. 181).

2. ". . . behavior of the admissions committee studied can be simulated by a linear combination of the criteria it considers" (p. 182).

3. ". . . the results of the simulation may be more predictive of the outcome than is the [committee]" (p. 182).

He based the first principle on research cited by Meehl (1954), Goldberg (1968), Sawyer (1966), and others who consistently found that predictors (including clinicians' judgments) could be combined better by statistical than by clinical (or judgmental, intuitive, introspective, empathic, subjective, etc.) means. The second principle referred to the accuracy with which decision behavior can be modeled and predicted, if the important cues are available. The last principle was an integration of the first two principles: Judges' statistical models can improve the consistency of their decisions. The last principle did not mean that predictions bootstrapped from a judge's model are equivalent to a statistical combination of predictors. Instead, it proposed that synthetic predictions are similar to those of the judge, but better. In this most important facet of his study, Dawes (1971) was able to combine four cues from

the graduate school application that would have allowed him to reduce the pool of applicants by 55% without excluding any the committee subsequently invited to the university. Again, the model of the committee was not necessarily the best way to predict the applicants' performance, but it was the best way to apply the committee's inferred selection policy.

Bootstrapping is a way human decisions can be simulated in an efficient, consistent manner, eliminating intermittent errors of judgment. In other words, "such decisions may be less capricious and more valid than those made by the decision maker relying on his own intuitions" (Dawes, 1971; p. 187). In the example above, it could have saved much of the time committee members spent reading and screening applications. They could have been presented with a smaller set of applications, sorted in predicted order of attractiveness, divided into subgroups of interest (e.g., minority, majority, male, female, etc.), a set chosen according to the committee's (more consistent) simulated policy.

Dawes and Corrigan (1974) were interested in why models of judges were more valid than the judges themselves. They noticed that (a) the kinds of predictors (cues) studied usually can be rescaled to have approximately linear relationships with the criterion (i.e., conditionally monotone relationships), (b) the error in predictors tends to increase linearity, and (c) weights that vary from the optimum (e.g., population regression weights) do not affect the correlation between criterion and predicted criterion in practical terms (i.e., linear models are robust with respect to errors in weighting). Furthermore, they believed judges engaged in prediction tasks (especially experts) had reasonable ideas of the direction, and occasionally the magnitude, of relationships among predictors and criterion. Consequently, judges' intuitive

weights were little more than random weights, correctly signed and normally distributed. They hypothesized

1. Intuitive decisions were valid if they emphasized variables that covaried nontrivially with the criterion.

2. Bootstrapped models of intuitive judgments applied judges' (essentially random) weights consistently.

3. Unit weights were equal to the average of all random (or all judges') weights.

4. Judges' models and random weights probably were less valid than models with unit or regression weights.

When they tested their hypotheses, judges' intuitive weights were least valid, followed by judges' bootstrapped linear models and random normal weights, then unit weights, and finally, optimal or regression weights. The differences between unit and optimal weights were negligible in most circumstances (see Schmidt, 1971, for more on unit weighting), and they concluded that all one needs to do is figure out which predictors are important, and add them up (i.e., use unit weights).

Why use anything other than optimal weights? When Dawes (1979) extended the arguments presented by Dawes and Corrigan (1974), he proposed that "people -- especially the experts in a field -- are much better at selecting and coding information than they are at integrating it" (p. 573). He justified bootstrapping as a way to select variables and estimate the direction (and perhaps the magnitude) of relationships, especially when criteria were unavailable (e.g., when using an instrument in a new situation) or with poorly operationalized variables (e.g., interpersonal skills, motivation to work, organizational fit, growth potential, etc.). In these cases, expert judges'

subjective weights would probably be farther from optimal weights (e.g., because of social desirability; Brookhouse et al., 1986) than would the weights of their modeled decisions (Hoffman, 1960; Einhorn, 1972). Also, optimal weights would be estimated from the relationship between predictors and proximal or irrelevant criteria: There is no guarantee that these criteria would be adequate to infer distal or ultimate relationships (for a discussion of the dimensions of criteria, see Brogden & Taylor, 1950). However, experts' judgments might predict more closely the relationship between predictors and a distal (or relevant, but poorly operationalized) criterion.

Dawes (1979) responded to some of the more frequent objections to replacing human judgments with linear models (convincing arguments are found on pp. 578-581). As for the criticisms that linear models only predict, for example, 16% of the variance in future faculty ratings, he replied

> Statistical prediction, because it includes the specification (usually a low correlation coefficient) of exactly how poorly we can predict, bluntly strikes us with the fact that life is not all that predictable. Unsystematic clinical prediction (or "postdiction"), in contrast, allows us the comforting illusion that life is in fact predictable and that we can predict it. (p. 580)

In other words, 16% of the variance in faculty ratings is better than none; but the fallacy of the complaint is that 100% of the variance is predictable. He addressed the charge that bootstrapping linear models is unethical because it reduces, for example, graduate school applicants to a GPA and a GRE score:

> Do we really believe that we can do a better or a fairer job by a 10-minute folder evaluation or a half-hour interview than is done by these two mere numbers? Such cognitive conceit . . . is unethical, especially given the fact of no evidence whatsoever indicating that we do a better job than does the linear equation. (And even making exceptions must be done with extreme care if it is to be ethical, for if we admit someone with a low linear score on the basis that he or she has some special talent, we are automatically rejecting someone with a higher score, who might well have had an equally impressive talent . . .). (p. 581)

In summary, the decisions of experts or judges can be predicted with consistency if some of the cues they use can be coded. Even if none of the cues can be coded, experts' ratings of the dimensions of a prediction problem can be used to predict their holistic or global judgments. Applying the resulting models (perhaps with unit weights) to the prediction problem (a) increases the validity of the decision, (b) provides better weights than asking the judges what cues they used, (c) ordinarily is much less expensive than using judges, and (d) can estimate which variables are important in situations where there are many potential predictors and inadequate criteria. Decision modeling can also be used to build a composite criterion, by regressing judges' ratings of some global criterion on potential criteria.

Some applications of decision modeling to problems in industrial and organizational psychology are presented next.

Decision Modeling in Industrial and Organizational Psychology

Few applications of decision modeling that dealt with problems in industrial and organizational psychology were found in the literature. Of those, even fewer dealt with bootstrapping (i.e., applying the judges' model to the decision).

Roose and Doherty (1976) directly tested some of the suggestions made by Dawes and Corrigan (1974). They created profiles with 64 cues extracted from the personnel files of 360 life insurance salesmen (200 in a scoring sample, and 160 in a crossvalidation sample). Sixteen agency managers rated each profile on whether the applicant described would be successful on the company's first year criterion. The judges' dichotomous ratings were converted to a ten point scale based on the confidence they expressed in each rating. There were very consistent models for each judge, but wide variations in criterion predictions among judges for a given profile. The judges' models included between five and

nine cues (39 different cues were used at least once), and intercorrelations among judges ranged from .02 to .61, averaging .39. However, intercorrelations among the predicted criterion scores from the judges' linear models ranged from .11 to .83, averaging .55.[12] They described these as "error free" measures of interjudge agreement. This agreement among predicted values illustrated the importance of assessing agreement by comparing outcomes instead of weights (cf. Brookhouse et al., 1986), and the robustness of linear models with respect to different weights and different colinear cues. When judges' models were bootstrapped, the results were equivocal: Some judges' predictions improved while others decreased; there was no practical mean difference across judges. But unit weights applied to the selected variables for each judge improved overall validity substantially. The largest bootstrapped validity coefficient was seen for a unit weighted composite of all judges (cf. Einhorn, 1972; Dawes & Corrigan, 1974). These results supported previous findings, but with less contrived cues, and with many more cues than had been studied previously. Of importance for scoring the BSI used herein was that judges were used to select cues (Einhorn & Hogarth, 1975) and those cues then performed best when unit weighted. Slovic, Fischhoff, and Lichtenstein (1977) pointed out that experts (e.g., psychologists, gamblers, bankers, stock brokers, graduate students, etc.) are content experts, not experts in normative decision making, so they would be expected to know which predictors are important, but not necessarily how to combine them to predict a given criterion. Roose and Doherty (1976) recommended to the sponsoring company that optimal regression weights be used to influence which cues agency managers used to hire salespeople, to modify

---

[12]The range and mean for the intercorrelations of the linear predictions were not reported, but were computed from Table 3 (p. 241; Roose & Doherty, 1976), using Fisher's transformation.

and unify hiring decisions. But previous studies (cf. Goldberg, 1968) suggested the managers would revert to their previous decision styles after training. Also, optimal weights were selected using a single proximal criterion, ignoring longer term or important, albeit inadequately operationalized, criteria. The criticisms of empirically scored biodata inventories applied to this study as well. Perhaps using a unit weighted composite of all (or the most consistent, experienced, or successful) agency managers would produce an enduring scoring key (cf. Hunter & Hunter, 1984), valid across a broader range of applicants, criteria, and agencies. The key would be empirically derived from the intuition of experts, and might come closer to producing a robust, internally consistent inventory, combining the best of empirical and rational methods. The authors noted how similar were their tasks, goals, and results to those of researchers studying employment interviewing. Yet they did not suggest that what they had created was a valid way to screen applicants for their company; in fact, in a few short steps they would have had a substitute for the agency level interview. Instead, they described their study as a way to clarify company policy, and a way to decide how much standardized test information agency managers should have before interviewing an applicant.

Libby (1976) obtained dichotomous judgments from 46 bank loan officers regarding which of 60 companies were likely to have defaulted on loans. Judges were supplied with a number of financial ratios, and their decisions were compared to bootstrapped decisions of their own models. The loan officers outperformed their own models ($p < .01$) on the same sample. Also the composite opinions of the judges (i.e., the sum of the dichotomous ratings for each company) outperformed the composite model of the judges. However, Goldberg (1976) reanalyzed Libby's (1976) data after transforming the

dichotomous criterion (default-no default) to a six point scale by including each judge's stated level of confidence in each rating. He transformed the cues (financial ratios, substantially skewed) to more symmetric distributions. The percentage of models outperforming their judges increased from 23% to 72%. He compared the results of this reanalysis to those of a previous unrelated study (Goldberg, 1970) and found them quite similar.

Zimmer (1981) replicated Libby's (1976) research using 30 Australian loan officers. Judges' crossvalidated models outperformed the judges themselves, but the differences were not significant. Zimmer (1981) and Libby (1976) both found loan officers made consistent predictions and had accurate insights into the way they weighted cues. These probably were signs that (a) the judges were experienced at actually making these decisions, (b) they were interested in the task (Libby, 1976), (c) the cues were important for the task (Goldberg, 1976), and (d) the criterion (loan default) was clear and had provided good feedback to experienced loan officers. Instead of concluding that the models had failed because they did not significantly improve predictions, Zimmer (1981) concluded that the equation of the most accurate loan officer (or better yet, the composite or unit weighted equations) could screen loan applications inexpensively, especially for new loan officers or those with poor accuracy.

Hobson, Mendel, and Gibson (1981) modeled the performance appraisal decisions of 19 faculty members and their department chairperson. The results were consistent with previous studies of decision modeling: The subjects varied in consistency, and had different rating models. They concluded (a) they were able to adequately model the raters, (b) a supervisor's model could rate

subordinates more consistently and efficiently than could the supervisor,[13] and (c) a bootstrapped model could be used as feedback to help subordinates learn which behaviors facilitated rewards from a supervisor. Unfortunately, they did not investigate bootstrapping in this study.

Zedeck, Tziner, and Middlestadt (1983) modeled ten interviewers' ratings of 412 applicants to an officer training school. The interviewers were all consistent (81% to 93% of the variance in the overall decision explained by the nine dimensions, not crossvalidated), but they seemed to use different strategies to predict success. The researchers inferred the importance of dimensions by the size of the zero-order correlations between cues and predictions. They labeled that the decision strategy of the interviewers (a questionable method). The overall success of the interviewers was not significant, but that may have been due to the criteria used -- ratings on 19 trait dimensions and an overall performance dimension after six and twelve weeks at school (i.e., proximal training performance). They concluded that interviewers used different strategies, weighted different dimensions, and were not equally valid: They needed a good preinterview screening instrument.

Dougherty, Ebert, and Callender (1986) modeled the decisions of three employment interviewers who rated 120 actual applicants during real interviews. The judges each interviewed 40 applicants and listened to tape recordings of the others' interviews. They rated applicants on eight dimensions covering interpersonal skills, work motivation, intelligence, and traditional employment information usually collected on biodata inventories or application blanks. The researchers wanted to find out how consistent the interviewers were, whether

---

[13]Supervisors might use their own bootstrapped policies as a starting point for each performance appraisal.

training made any difference in ratings, and how their validity compared to that of their bootstrapped models. The interviewers' overall ratings were predicted consistently by linear equations both before and after training (all pre- and posttraining equations had squared multiple correlations between .80 and .90). They were less consistent after training, which the researchers attributed to more nonlinear processes.[14] The bootstrapped models (pre- and posttraining) for two of the three interviewers were significantly better than their actual ratings across ten validity coefficients (for ten criteria). The third interviewer's predictive validity was good, and was not bested by the derived model. When Interviewer 3's model was bootstrapped on Interviewer 2's ratings of individual dimensions, validity improved significantly; however, the opposite happened for Interviewer 1. So Interviewer 2 and Interviewer 3 may have used similar strategies, different from Interviewer 1, or Interviewer 2 may have been so far from the optimum that any weights would have worked better. The bootstrapped posttraining models were better than all three raters. Aggregate validity for all three interviewers was not significant, even though Interviewer 3 was significantly valid. In sum, they showed that "bootstrapping improved on selected aspects of even a good interviewer's decision process" (p. 14). They proposed interview modeling is one area where differential (instead of unit) weighting is indicated. They pointed out the fallibility of the criteria used for validation, and suggested research in three areas: (a) using the strategy of an accurate interviewer to train others, (b) using the bootstrapped model of a good interviewer to predict for other interviewers, and (c) letting interviewers see

---

[14]Training emphasized unique aspects of the interview. Consequently, two of the three the interviewers used test scores and application blank information less after training.

predictions from their own bootstrapped models before making employment decisions.

Hamilton and Dickinson (1987) bootstrapped ten job incumbents' ratings of 100 hypothetical workers' profiles, and obtained estimates used in generating J-coefficients (i.e., synthetic validity coefficients). Specifically, they regressed the performance levels ascribed to the profiles by each judge, on the levels of each of six job elements represented in the profiles. They inferred the elements important for job performance from the regression weights. The ten judges were consistent (the multiple correlation between the job elements and performance was .93), and their models were similar enough to use a composite model (the multiple correlation dropped to .91 for the composite model). They used inferences from bootstrapped models to estimate (a) the population regression coefficients for job dimensions predicting job performance, (b) the intercorrelations among job elements, and (c) the correlations between elements and job performance. This was an example of how to bootstrap experts' judgments in a situation where the true relationships among predictors and criterion are unknown. Actually, these researchers had concurrent validity information available to check this method of generating J-coefficients. Bootstrapping (and two other methods) adequately reproduced the pattern of concurrent validities. However, the procedure was designed for situations where standard validation is not feasible. For example, it could be used when sample sizes are too small for a standard validity study, with new or redesigned jobs, or when the criterion is inadequate.

Research has shown the best way to interview is to (a) restrict inferences to applicants' social skills and motivation to work (and perhaps intelligence); (b) force interviewers to use consistently all the relevant information applicants provide; (c) train interviewers to ignore irrelevant information, stereotypes, prototypes, and nonverbal cues; and (d) provide maximum structure for the interview based on information relevant to job performance. Some researchers suggested interviews should be replaced by written instruments, and most suggested using written instruments to collect most of the information (e.g., Denton, 1964; Dunnette, 1961; Ulrich & Trumbo, 1965; Wagner, 1949). Many suggested information gathered in interviews should be combined statistically (e.g., Dawes, 1979; Dougherty et al., 1986; Wagner, 1949).

Biodata inventories were shown to collect information reliably and accurately. They also collected information better and less expensively than did interviews. Even subjective information was collected with good reliability over time. The methods for developing and scoring biodata inventories were either atheoretical, or depended too much on the introspection or empathy of developers. It was suggested that an empirical way to use experts' judgments might provide an acceptable, valid way to combine rational and empirical methods.

Statistical combination of predictors (including clinical judgments) was shown superior to clinical combination. Researchers found both linear and compensatory decision processes, and nonlinear and noncompensatory decision processes could be modeled adequately with linear regression models. More recent studies showed that these same models could be bootstrapped to improve the consistency, reliability, and predictive validity of experts' decisions. In

situations where there are many potential predictors, poorly operationalized criteria, or no criteria, bootstrapping provides a way to estimate the direction (and perhaps the magnitude) of predictor-criterion relationships. Adequate predictions were made by estimating the relevance and sign of potential predictors from models of decisions, and then unit weighting the important predictors in the indicated direction. Bootstrapping was shown important in studies of industrial and organizational psychology.

In the study reported in the next chapter, a biodata inventory scored by modeling the decisions of two judges (interviewers) was investigated. The inventory was developed to replace employment interviews for applicants to a managerial position, because interviews have been shown consistently to have poor reliability, low validity, and high costs. Interviews are often used for such positions, because managerial success is poorly operationalized, and the facets (and therefore standardized predictors) of job performance are not established. The instrument followed the structure of an interview designed for managers: Items were generated by a management psychologist on the basis of a job analysis, so they would be relevant to job performance, and interesting and familiar to the judges (cf. Latham et al., 1980; Libby, 1976), one of whom was the developer. The goal of this research was to (a) investigate whether interviewers could make consistent employment decisions using a written interview instrument, (b) study the relationships among various dimensions rated on the basis of the written instrument, (c) model the decisions of the interviewers, (d) use the models to select important variables from the hundreds of codable items on the instrument, (e) bootstrap the unit weighted decision models to predict success, and (f) compare the models of the interviewers.

Since many performance measures were available from the sponsor of this research, predictive validity was compared among the models, the interviewers, and optimal empirical models. However, these comparisons were ancillary to the rest of the research, because there were so many dimensions represented by the criteria, and little evidence to suggest whether any were related to the ultimate criterion of "managerial success."

This research was not simply a validation study. Instead, it was undertaken to demonstrate a theoretically sound process to overcome some of the deficiencies inherent in employment interviewing, and to propose a novel way to develop and score biodata inventories in the absence of a strong, well defined criterion.

CHAPTER III

METHOD

The data used in this research were collected in a corporation located in the southeastern United States. It had approximately 450 retail stores and 20,000 employees. This study was part of a concurrent validation, so applicants were managers employed by the corporation, as described in the next section.

Subjects

Applicants

Applicants were store managers who completed the BSI while employed by the corporation. The available manager population (N = 321) was sent the BSI. Summary demographic data were compiled from the managers' original applications, completed when they applied for employment with the corporation or when their store was purchased by the corporation. Their mean age was approximately 35.25 years, approximately 95% were male, tenure with the company averaged 6.16 years, and 7% were minorities.

Raters

Two judges provided ratings of each respondent's probability of success as a store manager. Rater 1 was a male, licensed psychologist with a PhD in school psychology and postgraduate training in industrial and organizational psychology. He had worked for seven years in a regional mental health clinic, and had been employed for two years as a psychologist specializing in executive and managerial employment interviewing. Rater 1 was the principal designer of the BSI. He was considered an expert interviewer in this study, but he was not presumed to be more of an expert than other professionals in his field.

Rater 2 was a male employee of the corporation with one year of experience as a recruiter. Rater 2 had been a store manager for the corporation for five years, and had worked for 25 years as a store manager in the industry. Rater 2 was promoted to recruiter by the corporation because of his broad experience in the industry. During his tenure as a recruiter, he received on-the-job training from another, more experienced, company recruiter, but he had not received training from Rater 1 in interviewing or assessment techniques. He had not attended any recruiting or interviewing classes or seminars.

Coder 1 was a female nonexempt employee of the corporation's Department of Human Resources and was responsible for entering information from BSIs into a computer database, using a coding key. She had less than one year of experience with the corporation at the beginning of the research, but she had been a nurse for 16 years. She previously had entered data obtained from the managers' original employment applications, and was familiar with the task. Following completion of the coding process, she was promoted within the Department of Human Resources to an exempt position in recruiting.

Instruments

Biodata Screening Instrument (BSI)

The BSI was developed by Rater 1 with input from the author, after a comprehensive job analysis was conducted. The BSI was designed to collect the same kinds of data as Rater 1 collected by interviewing. Applicants were asked to report past events and experiences on the BSI, and to answer questions about knowledge and attitudes relevant to the job. Items were created to collect behavioral data about applicants' (a) intellectual effectiveness, (b)

emotional characteristics, (c) interpersonal skills, (d) insight into self and others, (e) planning and organizational skills, (f) physical ability, (g) personal effectiveness, and (h) overall fit to the organization and job. Names assigned to dimensions were not intended to represent traits or constructs, but were convenient labels familiar to Rater 1; they allowed easy reference to groups of knowledge, skills, abilities, and attitudes which Rater 1 believed were relevant to the job.

The BSI is included in the appendix, but information identifying the corporation has been removed. It has four sections, the first of which was composed of open-ended items concerning the following areas: employment history, education, health history, preference for geographic location, personal goals and philosophy, and personal history. How these open-ended items were coded depended on the item. For example, an item asking for the qualities applicants admired in other people was coded as the number of responses fitting a specific category. An item asking which courses applicants liked most in high school was coded as the number of courses listed in each of several areas (e.g., mathematics, psychology, etc.).

The second section was a 100 item personality inventory using a true-false format without an explicit undecided option. It was developed by Rater 1 specifically for the BSI, and was exploratory. There were items intended to measure faking and social desirability in responses, and items about social skills and motivation to work. Responses of *true* were coded as 1, *false* as 0, and missing responses were noted.

The third section was a set of six multiple choice questions about management philosophy and situations. For example,

Another employee is talking to you about his last trip. The conversation is boring. It would be best to:

a) continue to listen politely
b) listen but act disinterested
c) tell him in a straightforward manner that the subject does not interest you
d) continue to look at your watch in an impatient manner.

There were no a priori correct or incorrect responses; instead, the responses were coded as categorical variables.

The final section of the BSI was composed of 12 esoteric questions specific to the industry. They could be answered with simple research (e.g., an encyclopedia, a dictionary, etc.), so they were included as power questions requiring an increment in knowledge, experience, or effort. The total number answered correctly was coded as a single item, instead of 12 separate items.

The coding key was the result of intuitive, rational processes, as are most coding keys. But it did not depend on "snooping" the data for ideas about coding, because it was developed before the BSI was administered.

Weighted Application Blank (WAB)

The applications originally submitted by managers when they were employed were coded and analyzed. The resulting WAB was validated before the BSI: The WAB and BSI projects were both started at the same time, but the WAB was validated before the BSI was administered. Completed WABs were available for the population of managers, and the WAB and the BSI were developed independently, so the WAB provided an alternative screening device with which to compare the BSI. The WAB also provided the demographic data on which managers completing the BSI were compared to those not completing the BSI. The validity coefficient for the WAB predicting a composite criterion[15] was

[15]The composite criterion is described later in this chapter.

found in previous research to be .294, with a 90% confidence interval between .133 and .402.

Rating Scales

A one page rating form, included in the appendix, was created to collect the raters' judgments on eight dimensions of interest: intellectual effectiveness (INTEFF), emotional characteristics (EMOTCHAR), interpersonal skills (INTSKILL), insight into self and others (INSIGHT), planning and organizational skills (ORGSKILL), physical ability (HEALTH), personal effectiveness (PERSEFF), and overall fit to the organization and job (ORGFIT). Two global decision ratings were also included: an overall hiring decision (HIRE), and an estimate of potential for future advancement (PROMOTE).

The eight dimensions were assessed using scales with eight numeric anchors and six verbal anchors: 0 -- *no data/can't tell*, 1 -- *obviously deficit* [*sic*], 2 -- *below average*, 3 -- (no anchor), 4 -- *average*, 5 -- (no anchor), 6 -- *above average*, 7 -- *exceptional*. The scales were also represented as vertical number lines, and raters were to record their judgments in boxes above the number lines.

The HIRE scale was represented as a horizontal number line with three numeric and verbal anchors: 1 -- *no hire*, 4 -- *undecided/reserve judgment*, 7 -- *definite hire*. There was a box to the left of the number line, where the raters were to record their judgments. The PROMOTE scale was identical to the HIRE scale, except that five anchors were used: 1 -- *little or no potential*, 3 -- *somewhat unlikely*, 4 -- *undecided/reserve judgment*, 5 -- *likely*, 7 -- *highly likely*. Unfortunately, this last scale was reproduced on the rating form with equal distances between anchors, even though they were not separated by equal numeric distances. The form was already in use before this

discrepancy was noted, and the scale was used as printed so bias caused by inconsistent intervals would be constant throughout the research.

Criteria

In Chapter II it was proposed that when a criterion is inadequate, scoring an inventory with decision models should be helpful. A relevant criterion of how the (concurrent) applicants in this research performed was elusive, so scoring the BSI with decision models was appropriate. But some measure of how different scoring methods performed was desired, so the corporation provided financial data and manager ratings for 429 stores. These criteria were not proved relevant, unbiased, or even under the control of store managers, but they were the only criteria available.

Financial

All financial data were standardized and reported as z-scores because they were confidential and proprietary to the corporation. The z-scores were based on means and standard deviations from the population of stores, so they were representative of managers' relative positions among peers. Some of the financial data were averaged across two departments reporting to the managers (Department 1 and Department 2).

Total sales (TOTSALES). TOTSALES was the dollar value of all goods and services sold through the store. This included goods and services not under the direct control of the store manager. When the author noted the opportunity bias of TOTSALES as a criterion (cf. Brogden & Taylor, 1950), the corporation's representatives responded that the managers at larger stores were better managers, otherwise they would not have been to those stores. Thus, TOTSALES might be a measure of advancement within the corporation.

Effective scheduling variance (ESV). ESV was a measure of how efficiently managers scheduled personnel given the ebb and flow of personnel requirements throughout the workweek. This measure was based on deviations of projected requirements from actual personnel needs. Larger ESV is associated with overscheduling, a costly practice. Mean ESV for Departments 1 and 2 was used. Table III-1 shows the correlation between these two departments.

Items per direct hour (ITEMS). ITEMS measured the number of discrete items sold at a store, divided by the number of paid hours of direct labor required to sell those items. This was the cost of labor associated with selling a piece of merchandise -- a cost purportedly influenced by store managers. Higher values represented lower relative direct labor costs and better store operation. Mean ITEMS for Departments 1 and 2 was used. Table III-1 shows the correlation between these two departments.

Wage percentage (WAGE). WAGE was a measure of the wages paid for direct labor as a percentage of total sales. Higher values were associated with higher relative labor costs. However, labor rates depended on the local labor market, and the corporation had policies constraining excessive wages. Mean WAGE for Departments 1 and 2 was used. Table III-1 shows the correlation between these two departments.

Product shrink (SHRINK). SHRINK was the difference between the dollar value of goods purchased for resale and the dollar value of goods available for resale (i.e., goods sold or in inventory). SHRINK resulted from many causes -- virtually all of them the responsibility of store managers: employee theft; spoiled goods; goods accepted but not received; goods damaged before, after, or during stocking; and customer returns. A negative value was the result of recovering items for resale and occurred infrequently. SHRINK was reported as

Table III-1. Correlations Between the Same Criteria Reported for Two Departments

Criterion	Pearson Correlation Coefficient	Number of Observations
Effective scheduling variance	.443	412
Items per direct hour	.524	414
Wage percentage	.508	426

a percentage of total sales, and was not related to TOTSALES (r = -.040, n = 423).

Gross profit after shrink (GPS). GPS was profit after subtracting the cost of shrink, for the entire store. This included profit from operations within the store not reporting to store managers. GPS was reported as a percentage of TOTSALES, and was not related to TOTSALES (r = -.019, n = 423).

Controllable expenses (CONTEXP). CONTEXP included all costs of doing business attributable to store managers, such as utilities, wages, advertising, and so forth. It did not include costs associated with purchasing, realty, or corporate expenses. CONTEXP was reported as a percentage of total sales, and was related to sales (r = -.689, n = 423).

Employee Turnover. Full-time turnover (FULLTURN) and part-time turnover (PARTTURN) were available as percentages of total nonexempt staff. These two criteria were not combined because they correlated poorly (r = .219, n = 350).

Ratings

Employee attitude scores (EMPATT). EMPATT was measured by an annual confidential survey of employees about working conditions in stores. Employees of each store were asked about their manager's work habits, personnel practices, and how satisfied they were with their manager. EMPATT indicated the relative number of negative responses to the survey, and was obtained from store managers' personnel records.

Performance scores (PERFSCOR). PERFSCOR was an annual supervisory rating, with exempt and nonexempt employees measured on different scales. If

a manager's most recent PERFSCOR was rated on the nonexempt scale, it was deleted.[16]

Store audits (AUDIT). Internal auditors conducted unannounced audits of stores at various intervals. They assessed the operating state of stores on such attributes as appearance, staffing, cleanliness, and financial accuracy. AUDIT resulted from this process.

The intercorrelations among EMPATT, PERFSCOR, and AUDIT are shown in Table III-2. They were very low, and indicated either low true intercorrelations or poor reliability.

Reliability of Criteria

Criteria were available for (a) the first quarter of 1986 and (b) all of 1986, so estimates of reliability over time were computed. However, only 307 stores and 271 managers were included in both sets of criteria, and only 148 managers were at the same stores at the beginning and end of 1986. Reliability coefficients probably were positively biased, because first-quarter performance contributed to year-end results, and some ratings were collected at intervals greater than one year. Table III-3 shows the correlations between first-quarter and year-end data matched on (a) stores, (b) managers, and (c) stores and managers.

---

[16]Converting scores from nonexempt to exempt scales would have presumed conditions that could not be tested. For example, would an applicant's relative position among nonexempt peers remain constant among exempt personnel? Are distribution properties the same on a 100 point scale and a five point scale?

Table III-2. Intercorrelations Among Employee Attitude Scores, Performance Scores, and Store Audit Scores[1]

Criterion	Employee Attitude Score	Performance Score
Performance Score	-0.176	
Audit Score	-0.051	0.108

[1]All coefficients were based on 277 observations.

Table III-3. Reliability Coefficients for Criteria Computed from First-quarter and Year-end Data: Matched on Managers, Stores, or Managers and Stores[1]

Criterion	Matched on Managers	Matched on Stores	Matched on Stores and Managers
Total sales	.573 (263)	.804 (305)	.865 (147)
ESV[2] (Department 1)	.233 (255)	.248 (299)	.265 (144)
ESV (Department 2)	.149 (255)	.078 (299)	.273 (144)
ESV (Mean)	.232 (255)	.205 (299)	.337 (144)
ITEMS[3] (Department 1)	.196 (256)	.305 (299)	.410 (144)
ITEMS (Department 2)	.394 (256)	.619 (299)	.671 (144)
ITEMS (Mean)	.216 (256)	.382 (299)	.489 (144)
WAGE[4] (Department 1)	.368 (263)	.401 (305)	.481 (147)
WAGE (Department 2)	.460 (263)	.634 (305)	.686 (147)
WAGE (Mean)	.414 (263)	.549 (305)	.614 (147)
Product shrink	.139 (261)	.141 (303)	.203 (146)
Gross profit after shrink	.131 (261)	.281 (303)	.439 (146)
Controllable expenses	.344 (261)	.622 (303)	.666 (146)
Full-time turnover	.067 (218)	.107 (265)	.074 (125)
Part-time turnover	.080 (247)	.158 (300)	.090 (144)
Employee attitude score	.409 (175)	.520 (217)	.581 (115)
Performance score	.302 (220)	.088 (225)	.167 (123)
Store audit	.048 (242)	.240 (285)	.118 (142)

[1]There were 271 stores and 305 managers with data at both times, and 148 managers with the same stores across both time periods. Number of observations for each coefficient is shown in parentheses.

[2]Effective scheduling variance.

[3]Items per direct hour.

[4]Wages as a percentage of total sales.

Composite Criterion

A composite criterion was prepared by modeling the decisions of the Vice President of Operations for the corporation, because he was responsible for, and familiar with, the performance of store managers. He rated 50 randomly sampled (unidentified) managers on the basis of the criteria (cues) listed above (i.e., TOTSALES, ESV for Departments 1 and 2, ITEMS for Departments 1 and 2, etc.). The rating scale was anchored by the following descriptions: 1 -- *a least desirable manager*, 4 -- *satisfactory or competent manager*, 7 -- *my ideal manager*. He did not have to assign integers, and he was instructed to use as much of the scale as possible. His ratings were standardized and regressed on the cues. The criteria which accounted for most of the variance in his ratings were selected using stepwise multiple regression, and regression weights were estimated using the following modified double crossvalidation procedure.

The executive's 50 ratings and corresponding cues were standardized and randomly divided into two equal samples of 25 ratings. In each sample, ratings were regressed on cues, and the resulting regression weights were used to predict ratings in the sample in which they were *not* derived. The mean correlation between those ratings and the predicted ratings across samples estimated the crossvalidated prediction coefficient, and the mean of each pair of regression weights estimated the crossvalidated beta weights (Hunter & Hunter, 1984). Each sample also was unit-weighted,[17] and the mean correlation coefficient between ratings and unit-weighted cues provided an estimate of the unit-weighted prediction coefficient. This double crossvalidation was performed 500 times, and regression weights and prediction coefficients were saved each

---

[17]The direction of each unit weight was determined by the direction of the corresponding optimal (ordinary least squares) regression weight.

time. The medians of the 500 sets of mean regression weights estimated the weights for combining the criteria into a composite. The medians of the 500 sets of mean prediction coefficients estimated the strength of the relationship between the executive's ratings and the cues selected (cf. Lunneborg, 1985), and the median difference between the squared prediction coefficients indicated whether unit or optimal weights should be used.

Cues selected by stepwise multiple regression were excluded from the composite criterion if the range of the middle 95% of their 500 regression weights included 0. The cues included in the composite criterion and their weights are shown in Table III-4. They were SHRINK, CONTEXP, EMPATT, GPS, WAGE (for Department 1 only), AUDIT, and ITEMS (for Department 1 only). Unit weights virtually always predicted more of the variance in the executive's ratings: The median difference between coefficients of determination was -.125 (the negative sign indicates superiority for unit weights), with a 95% confidence interval between -.339 and .032. The median unit-weighted prediction coefficient for the executive's ratings was .914, with a 95% confidence interval between .905 and .925.[18] Therefore, the seven criteria selected predicted about 83.6% of the variance in the executive's ratings (the 95% confidence interval was between 81.9% and 85.5%). This unit-weighted composite was one of the criteria used later for validation.

---

[18]Nonparametric confidence intervals were derived by sorting the statistic of interest, and selecting the two points cutting off the tails of the distribution. In this instance, the endpoints were estimated by taking the 13th and 488th sorted estimates from 500 double crossvalidations.

Table III-4. Crossvalidated[1] Model Statistics: A Corporate Executive's Ratings of Managers Predicted from Seven Performance Cues (n = 50)[2]

Statistic	Lower 95% Bound	Median	Upper 95% Bound	Z-score[3]
REGRESSION WEIGHTS				
Items per direct hour[4]	.102	.215	.360	3.31
Wage percentage[4]	-.366	-.264	-.149	-4.95
Store audit score	.061	.126	.195	3.90
Controllable expenses	-.367	-.299	-.239	-8.71
Product shrink	-.455	-.347	-.245	-6.59
Gross profit after shrink	.144	.237	.355	4.53
Employee attitude score	-.268	-.201	-.143	-6.15
PREDICTION COEFFICIENTS				
Crossvalidated coefficient	.835	.891	.921	20.21
Unit-weighted coefficient	.905	.914	.925	97.56
Difference between squared coefficients	-.339	-.125	.032	-1.37

[1]Double crossvalidations using randomly allocated samples were performed 500 times.

[2]Twelve performance criteria were used as cues, and seven were selected for this analysis using stepwise multiple regression.

[3]Mean estimates divided by standard deviations (n-1) of estimates. Coefficients were first transformed using Fisher's *r* to Z transformation (Hays, 1981).

[4]Measured on Department 1 only.

Data Collection

When used, randomization was achieved with a computer-generated set of uniformly distributed pseudorandom numbers.

Applicant data. Store managers who had held that position during the first two quarters of 1986 were sent the BSI as part of a concurrent validation study. They were informed in a cover letter that the corporation's Department of Human Resources was involved in a management selection study. It stated that results would be considered confidential and would be reviewed only by authorized personnel, and that participation was important to the corporation. They were not explicitly informed that the BSI would not be used for administrative, promotional, compensation, or disciplinary purposes. The cover letter was dated June 5, 1986 with a requested due date of June 30, 1986. By August 30, 1986, 168 complete or partially complete BSIs were returned to the Department of Human Resources.

Coding. Coder 1 coded 168 BSIs by February, 1987 and sent the data to the author for analysis. The actual coding key is being used by the corporation, so it is not reproduced here. Generally, it directed the coding of cues such as (a) the number of different responses to specific open-ended items; (b) the number of those responses judged positive or negative; (c) the comprehensiveness, neatness, and style of responses; 4) the number of jobs held; and 5) the nature of club or organization memberships. True-false and multiple choice items were simply transcribed to the database. Including identification information, the coded BSIs consisted of 245 cues per applicant.

Dimension ratings. The raters were given the BSIs in an identical random sequence. The sequence was noted for subsequent analysis of practice effects.

Each rater judged each applicant on each dimension. No instructions were given to the raters about how to rate applicants. But each rater knew the purpose of the validation study was to predict subsequent performance with these scales, and they knew their performance would be observed and evaluated. The BSI was long, and there were many applicants, so the raters were given as much time as they needed to complete the ratings. They both completed the task by October, 1987.

Data Analysis

Applicants' written responses (cues), judges' ratings of those responses, and the corporation's criteria were the three data sets available for this study. Descriptive statistics were obtained for all three, and relationships among cues, judges' ratings, and criteria were investigated.

Missing data. Missing cues were replaced with the means of the corresponding variables when practical, a procedure recommended by Cohen and Cohen (1983). Also, the total number of missing answers on the true-false section were included as a separate variable in all analyses. Observations were deleted if any criteria or ratings were missing.

Descriptive statistics. The coder's error rate was assessed by verifying a sample of all cues entered. It was estimated within broad bounds, because she had to get data directly from BSIs instead of from prepared coding sheets. Descriptive statistics for ratings were compared across applicants, and reliability was inferred. Descriptive statistics for criteria were compared between respondents and nonrespondents.

Computational methods. The essential question in the hypotheses tested concerned differences between or among various estimated relationships. Three ways such differences can be assessed are 1) intuitive judgment of practical

differences among coefficients, 2) analytical methods relying on normal theory assumptions, and 3) percentile approximations of normal theory tests inferred from distributions of estimates obtained by repeatedly analyzing random samples of data. The last of these is known as *bootstrapping* in statistical literature, but was referred to here as *systematic resampling* to avoid confusion with the process of bootstrapping decision models.

This study used a single sample, but comparisons of correlation coefficients usually require independent samples (cf. Hays, 1981). So systematic resampling often was used to test hypotheses and parameter estimates. Systematic resampling was used to estimate the variability of a statistic of interest for the population of samples with the same number of observations as the research data. This was accomplished by drawing many samples *with replacement*[19] from the available observations, and computing the statistic of interest for each sample. The results were saved from all samples, and rank ordered to create a distribution for the statistic. The resulting distribution was an estimate of the sampling distribution of the statistic in the population of all samples of that size (for complete descriptions, see Efron, 1982, 1987; Efron & Gong, 1983; Lunneborg, 1985).

In this research, the size of all systematic samples equaled the number of observations available. Observations were resampled 500 times (Efron, 1987, estimated as few as 25 replications are sufficient for some purposes), and distributions of the 500 parameter estimates were used to determine point

---

[19]The key in any procedure which uses systematic resampling is to sample with replacement. For example, a sample of 50 drawn from an ad hoc population of 50 probably will exclude some observations and include multiple copies of others. By drawing many such samples, one can estimate the distribution of a statistic for the population of samples. This procedure was used to compute results that generalized to other similar samples.

estimates and confidence intervals. For example, a systematically resampled correlation coefficient was estimated as the median of the accumulated distribution; 95% confidence intervals were estimated as the endpoints of the middle 95% of the accumulated distribution (the 488th and 13th sorted estimates).

When there was more than one potential predictor, the modified double crossvalidation method described above for estimating weights for the composite criterion was used: Standardized variables were selected with stepwise multiple regression, and crossvalidated and unit-weighted statistics were estimated 500 times with random division of the observations into two equal samples each time. The variables selected were assigned unit weights unless differential weights were justified. Correlations between predicted and actual dependent variables estimated multiple correlation coefficients, because relative predictions were more important than point estimates. Those correlations were squared to estimate coefficients of multiple determination. When means, medians, or standard deviations for correlation coefficients were required, the coefficients were first normalized with Fisher's *r* to Z transformation (Hays, 1981).

Hypotheses and Research Questions

The hypotheses investigated in Chapter IV addressed three sets of related questions:

1. Could HIRE and PROMOTE be modeled from the eight dimension ratings? If so, were the models linear and compensatory? Were unit weights as accurate as optimal and crossvalidated weights?

2. Could raters' eight dimension and two global ratings be modeled from cues (i.e., BSI responses) or principal components of cues? If so, would unit

weights outperform crossvalidated weights? Would individual cues predict the ratings better than a few principal components? Would interrater reliability be better for bootstrapped ratings than for actual ratings?

3. Given good approximations of HIRE and PROMOTE by dimension ratings (Hypothesis 1, above), could HIRE and PROMOTE be modeled better using (a) bootstrapped dimension ratings, (b) actual dimension ratings, or (c) BSI cues? Would cues important in models of relevant dimension ratings be different from those important in models of HIRE and PROMOTE derived directly from cues?

Because the criteria provided had doubtful reliability and relevance, the purpose in this research was to use modeled decisions instead of criterion-related validity to create a scoring procedure; however, the criteria provided the opportunity to answer three interesting research questions (Questions 1, 2, and 3) in Chapter IV:

1. How valid were (a) ratings, and (b) bootstrapped ratings for predicting the available criteria?

2. How valid were the BSI cues for predicting the criteria?

3. How did utility compare among (a) a hypothetical interviewer, (b) the WAB, and (c) the BSI cues for predicting the composite criterion described previously?

CHAPTER IV

RESULTS

Coding Accuracy

Of the 238 cues coded on the BSI, 140 (59%) were entered without coding, and 98 (41%) required simple coding or transformation. Intercoder reliability was not assessed because there was only one coder. Coding was not complex, so a sample of the 39,508 coded items (238 cues for each of 166 respondents) were verified by the author. A random sample of 202 entries was required to estimate the error rate within bounds of 3% with an a priori estimate of 5% (Scheaffer, Mendenhall, & Ott, 1979; equation 4.19). The sampling frame was constrained to a random subset of 25 BSIs, to reduce the number of separate BSIs inspected (cf. Schmitt et al., 1984). Eight errors were found, so the population error rate was estimated between 1.3% and 6.6%, $p < .05$.

Demographics

One hundred sixty-six applicants responded with complete information, a 52% return rate, but at least three of the managers were not correctly identified because of duplicate identification numbers. The following means are demographics for managers who responded and those who did not:

Group	Age	Percent Male	Years Tenure	Percent White
Responded	36.06	.929	7.55	.965
Did not respond	33.87	.960	7.89	.911

The MANOVA simultaneous test of differences between means was significant, $F(4, 337) = 2.864, p = .0234$, so univariate effects were tested. The results showed that responding managers were significantly older than managers who did not respond (36.09 years old vs. 33.87 years old), $F(1, 340) = 7.62, p < .01$. So results or conclusions reported here might depend to some extent on the age of applicants, making generalization to the population of potential applicants tentative.

Rating Scale Characteristics

Unfortunately, the raters did not understand the instructions about rating the BSIs in random order. They rated them in an undocumented order and did not consistently date the rating forms, so it was not possible to assess practice effects in this research. However, no obvious systematic bias was revealed in any of the subsequent analyses.

Rater 1 rated more applicants than did Rater 2. Also, Rater 1 and Rater 2 both rated more than 166 BSIs, the number of BSIs coded, because some BSIs arrived after coding was completed and were not included in the computer data set.

Summary Statistics

Table IV-1 shows the means, standard deviations, and squared multiple correlations (SMCs) of each scale on the judgment rating forms for Rater 1 and Rater 2. SMCs are indicators of the amount of colinearity among variables (Neter, Wasserman, & Kutner, 1985), expressed as the amount of variance in one variable predicted by a linear combination of all other variables. In this case, the SMC for each dimension rating is the amount of variance accounted for by the other seven dimension ratings, but not the global ratings (HIRE and

Table IV-1. Rating Scale Characteristics for Global and Dimension Ratings by Two Raters

Rating Scale	N	Mean	Standard Deviation	Minimum	Maximum	SMC[1]
Rater 1						
INTEFF	184	4.768	.892	0.0[2]	7.0	.256
EMOTCHAR	184	3.978	.716	2.0	6.0	.577
INTSKILL	184	4.049	.758	3.0	6.0	.401
INSIGHT	184	3.938	.718	2.0	6.0	.391
ORGSKILL	184	4.462	.852	2.5	6.5	.630
HEALTH	184	4.157	1.138	2.0	7.0	.137
PERSEFF	184	4.195	.716	2.0	6.0	.721
ORGFIT	184	4.167	.994	2.0	6.8	.806
HIRE	184	4.408	1.246	1.0	7.0	.578
PROMOTE	184	3.883	.889	2.0	7.0	.490
Rater 2						
INTEFF	170	4.321	.693	3.0	6.0	.366
EMOTCHAR	170	4.750	.642	3.0	6.0	.590
INTSKILL	170	4.712	.636	3.0	6.0	.506
INSIGHT	170	4.737	.723	3.0	7.0	.527
ORGSKILL	169	4.331	.582	3.0	6.0	.545
HEALTH	169	4.429	.601	3.0	6.0	.486
PERSEFF	169	4.675	.613	4.0	6.0	.753
ORGFIT	166	4.681	.633	4.0	6.0	.780
HIRE	170	4.832	.827	1.0	7.0	.506
PROMOTE	170	3.574	.788	1.0	6.0	.481

[1]Squared multiple correlation. This was computed using all ratings except HIRE and PROMOTE.

[2]This value was converted to missing, so there were 183 observations available for computing SMCs.

PROMOTE). The SMC for each global rating was computed from the eight dimension ratings, thus representing the variance linearly predictable by the dimension ratings.

Rater 1 appeared to use more of the scale than did Rater 2, as indicated by the somewhat larger standard deviations and wider ranges. Also, Rater 1 may have discriminated more among the dimensions than did Rater 2, as evidenced by relatively low SMCs for INTEFF, INTSKILL, INSIGHT, and HEALTH. The SMCs for Raters 1 and 2 are compared in Figure IV-1.

Reliability

Table IV-2 shows the reliability of ratings estimated by intraclass correlation coefficients. The first set estimates the average reliability of each scale across raters. Between-rater variance was excluded from the error term because Rater 1 and Rater 2 comprised the entire population of raters, and each applicant was rated by both raters (Ebel, 1951). The second set of estimates is comprised of Pearson product-moment correlation coefficients (r) between raters. As shown, the intraclass estimates ranged from .231 to .545, and the rs were between .153 and .429. These estimates indicated poor interrater reliability, lower than might be expected for a structured interview. The intraclass correlations across the ten scales, within raters, were .878 and .909 for Raters 1 and 2, respectively. Error variance due to scale differences was excluded from the error term because the scales were used simultaneously (Ebel, 1951).[20] These high intrarater reliabilities indicated that most of the variance within raters was accounted for by systematic differences in ratings among

---

[20]Ebel (1951) noted that between-rater (scale) variance should be removed from the error term when ranks or Z-scores are used, when averages of all raters (scales) are used for each subject, or when using unit weights.

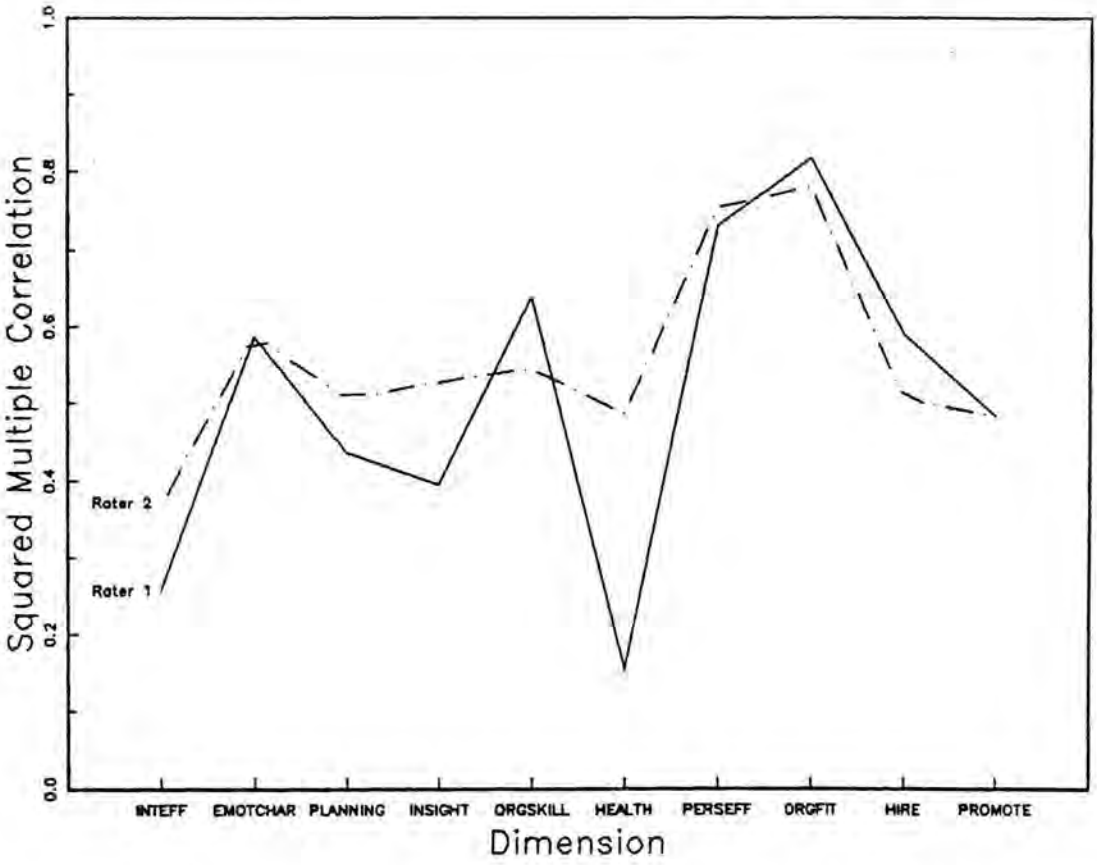

Figure IV-1. Squared multiple correlations (SMCs) for global and dimension ratings for two raters. All SMCs were computed without using either global ratings (i.e., HIRE and PROMOTE).

Table IV-2. Intraclass Correlation and Pearson Product-moment Correlation Estimates of Interrater Reliability for Two Raters

Rating Scale	N	Intraclass Correlation	Pearson Correlation
INTEFF	156	.483	.331
EMOTCHAR	156	.231	.153
INTSKILL	156	.233	.159
INSIGHT	156	.295	.178
ORGSKILL	155	.341	.228
HEALTH	155	.285	.219
PERSEFF	155	.253	.169
ORGFIT	152	.442	.331
HIRE	156	.545	.429
PROMOTE	156	.450	.297

applicants. However, they may have been due to poor discrimination among the scales, or rater halo.

Criteria

Individual Criteria

Figure IV-2 shows criterion data for managers who responded and those who did not. The MANOVA simultaneous test of differences between corresponding means of criteria for respondents and nonrespondents was not significant, $F(15, 223) = .802, p = .6748$. So the means of respondents and nonrespondents were approximately the same on the criteria.

Composite Criterion

The difference between means of the (unit-weighted) composite criterion for respondents and nonrespondents was not significant $F(1, 237) = .062, p = .804$. So respondents' and nonrespondents' means were approximately equal.

Hypothesis Tests

Hypothesis 1

The first hypothesis was that HIRE and PROMOTE could be modeled for each rater from his eight dimension ratings. Stepwise regression was used to select dimension ratings relevant to each global rating. The modified double crossvalidation procedure was used to estimate weights and shrunken prediction coefficients for each model equation. The results are shown in Table IV-3. The prediction coefficients are significantly different from zero because their lower 95% bounds are greater than zero. Therefore, significant proportions of variance in raters' global ratings were predicted by two to five dimension

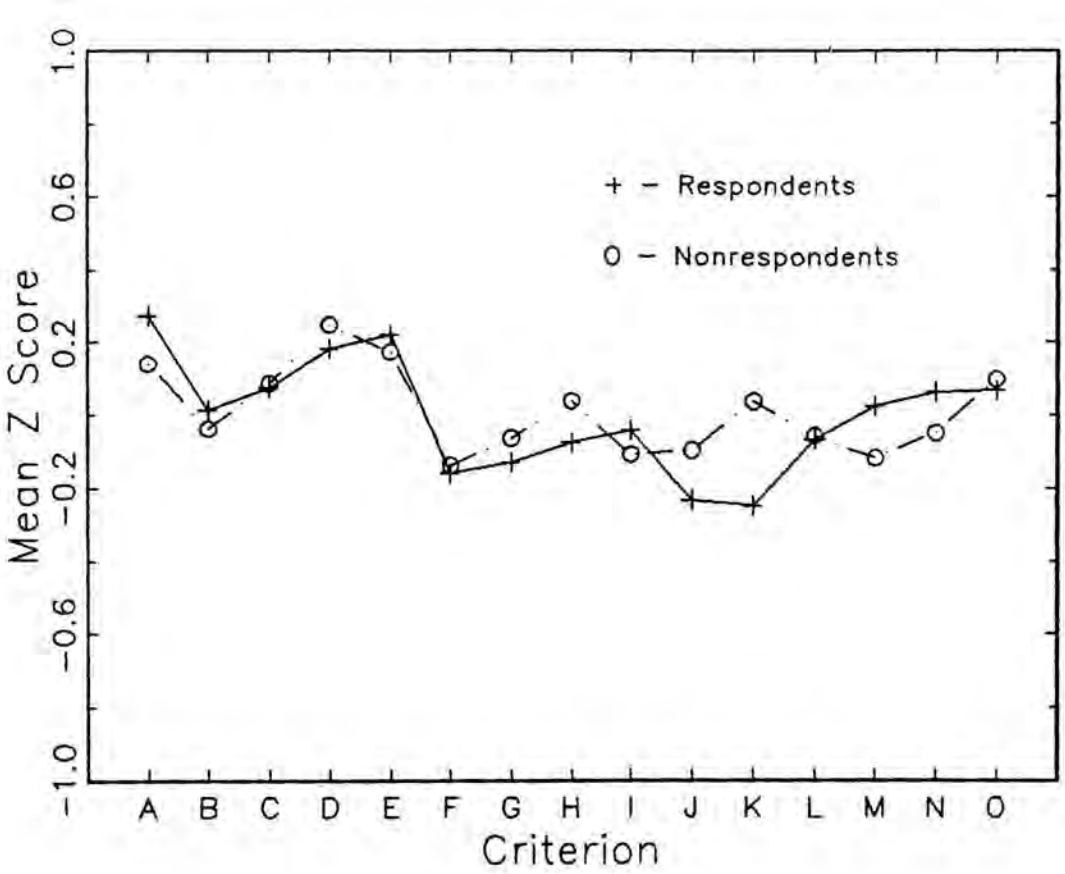

Figure IV-2. Means of standardized criteria for managers who responded and those who did not. Criteria are indicated as follows: A = total sales; B, C = effective scheduling variance for two departments; D, E = items per direct hour for two departments; F, G = wages as a percent of total sales for two departments; H = product shrink; I = gross profit after shrink; J = controllable expenses; K, L = full- and part-time turnover, respectively; M = employee attitude score; N = performance score; O = audit score.

Table IV-3. Crossvalidated Model Coefficients for Global Ratings Predicted from Selected Dimension Ratings[1]

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
			RATER 1			
HIRE	5	184	.758	.671	.729	.753
PROMOTE	3	183	.691	.570	.669	.740
			RATER 2			
HIRE	4	165	.706	.447	.667	.784
PROMOTE	2	169	.684	.593	.680	.742

[1]Two equal size samples were randomly allocated 500 times. Each time, the model of interest was double crossvalidated.

[2]All crossvalidated coefficients are zero-order correlations between predicted and actual ratings estimated from 500 iterations.

[3]Number of predictors in the model.

[4]Optimal or ordinary least squares multiple correlation coefficient.

ratings using simple linear models. The medians of all distributions were well above .60, and the lower 95% confidence bounds of three of the four distributions were above .50. On average, the crossvalidated prediction coefficients were greater than .65.

Hypothesis 2

This hypothesis concerned the form of the model that could best describe the relationship between dimension ratings and each global rating. Previous research (see Chapter II) indicated that simple linear models were sufficient to describe most decisions. However, Einhorn (1970) presented transformations that allow linear models to describe nonlinear conjunctive or nonlinear disjunctive processes. It was hypothesized that linear compensatory models would be as good as either of Einhorn's (1970) nonlinear models. To test this hypothesis, the stepwise regression procedure was performed on data transformed as suggested by Einhorn (1970). The results were compared to the output from stepwise regressions computed during the test of Hypothesis 1. Results for all three models are shown in Table IV-4. In only one instance (Rater 1, HIRE) was a nonlinear model superior to a compensatory model. In that instance, the increment in the multiple correlation was less than .01 (i.e., one percent of the total variance). In the other cases, using the nonlinear models resulted in increased error variance ranging from 1.4% to 18.7% of the total variance. Although the ratings of Rater 1 were predicted well by the conjunctive model, the simple linear model performed at least as well: Rater 1 may have used some conjunctive processes, but his ratings were modeled well by linear compensatory models.

Table IV-4. Comparison of Stepwise Regression Results using Linear and Nonlinear Models to Predict Global Ratings from Dimension Ratings by Two Raters

Rating Scale	N	Number of Predictors	Squared Multiple Correlation
	LINEAR COMPENSATORY		
Rater 1			
HIRE	183	5	.574
PROMOTE	183	3	.477
Rater 2			
HIRE	165	4	.499
PROMOTE	165	2	.469
	NONLINEAR CONJUNCTIVE		
Rater 1			
HIRE	183	4	.583
PROMOTE	183	3	.463
Rater 2			
HIRE	165	3	.333
PROMOTE	165	2	.388
	NONLINEAR DISJUNCTIVE		
Rater 1			
HIRE	183	4	.431
PROMOTE	183	2	.405
Rater 2			
HIRE	165	4	.312
PROMOTE	165	2	.378

Hypothesis 3

This hypothesis concerned the issue of whether to use unit or absolute (crossvalidated or optimal) weights for modeling and prediction. Previous researchers (e.g., Schmidt, 1971; Dawes & Corrigan, 1974) documented well the case for unit weighting, so it was hypothesized that unit weights consistently would perform as well as crossvalidated weights when modeling global ratings from dimension ratings.[21] To test this hypothesis, unit weights were applied to the crossvalidation samples randomly allocated during the test of Hypothesis 1. Mean prediction coefficients using unit and crossvalidated weights and the difference between the squared coefficients were computed for each of the 500 iterations. The results are presented in Table IV-5. The 95% confidence intervals of the coefficients overlapped, suggesting that neither unit nor absolute weights were clearly superior for prediction in these particular circumstances. In addition, the 95% confidence interval of the difference between the squared coefficients was centered approximately about zero. Therefore, neither weighting scheme was superior.

Hypothesis 4

This hypothesis was that the reliability of the global ratings would improve when bootstrapped from relevant dimension ratings. For each rater, HIRE and PROMOTE were bootstrapped by summing the standardized relevant dimension ratings. Then intraclass and Pearson correlation coefficients were computed between raters for each bootstrapped global rating:

---

[21]Unit weights were not expected to perform consistently better, because there were only eight potential predictors, and well over 100 observations.

Table IV-5. Comparison Between Crossvalidated and Unit-weighted Model Coefficients for Global Ratings Predicted from Selected Dimension Ratings[1]

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
			RATER 1			
HIRE						
Crossvalidated	5	184	.758	.671	.729	.753
Unit-weighted				.713	.724	.731
Difference				-.069	.009	.041
PROMOTE						
Crossvalidated	3	183	.691	.570	.669	.740
Unit-weighted				.594	.666	.724
Difference				-.080	.008	.052
			RATER 2			
HIRE						
Crossvalidated	4	165	.706	.447	.667	.784
Unit-weighted				.653	.722	.781
Difference				-.253	-.084	.026
PROMOTE						
Crossvalidated	2	169	.684	.593	.680	.742
Unit-weighted				.613	.686	.747
Difference				-.067	-.004	.007

[1]Two equal size samples were randomly allocated 500 times. Each time, the model of interest was double crossvalidated, unit-weighted, and the results were compared.

[2]All unit-weighted and crossvalidated coefficients are zero-order correlations between predicted and actual ratings estimated from 500 iterations. The difference coefficient is the squared crossvalidated coefficient minus the squared unit-weighted statistic for each of 500 iterations: A negative difference coefficient indicates the superiority of unit-weights.

[3]Number of predictors in the model.

[4]Optimal or ordinary least squares multiple correlation coefficient.

Rating	Bootstrapped Observations	Intraclass Correlation	Pearson Correlation
HIRE	151	.505	.342
PROMOTE	154	.546	.385

Compared with the reliability results presented earlier, bootstrapping increased both coefficients of reliability substantially for PROMOTE, and decreased them moderately for HIRE. Although the mean reliability appeared to increase, none of the coefficients was above the .70 level often cited as adequate (Latham et al., 1980).

Hypothesis 5

It was proposed that each of the dimension and global ratings could be modeled directly from the cues available to the raters on the BSIs. The following procedure was followed for each rating by each rater:

1. In order to reduce the computational burden, BSI cues exhibiting little or no relationship with the rating were excluded from later steps. The zero-order correlation coefficient between each of the 262 cues[22] and the dimension rating was computed. A cue was included in the pool of potential predictors if its correlation with the rating was significant at the .10 level (two-tailed).

2. Cues from the BSI were selected from the pool of potential predictors using a stepwise regression procedure. All variables which remained significant at the .05 level were retained.

---

[22]There were 262 cues because some of the cues originally coded were transformed to multiple indicator variables. All categories of indicator variables were always represented: Dummy coding was used instead of contrast or effects coding so that the intercept did not represent a group or category.

3. Weights, prediction coefficients, and a comparison between the performance of unit and crossvalidated weights were computed using the modified double crossvalidation procedure described earlier, with one further modification: When a weight was inestimable after observations were randomly allocated to two samples, a generalized least squares regression was used, and the inestimable weight was assigned a weight of 0 for computations during that iteration. This often occurred because some response frequencies were small for some cues. Such cues were not dropped summarily from the analysis because some of them may have been important to the raters because of, or in spite of, the rarity of certain responses. In an actual selection study, if crossvalidation results (i.e., the distribution of estimates) showed that a parameter was generally inestimable, unstable, or insignificant then it would not be considered in future selection decisions; however, since models were not estimated more than once in this research, all BSI cues selected by the stepwise procedure were used to estimate predictive statistics.

4. Weights were estimated by the median of the distribution of 500 double crossvalidated estimates.

Optimal multiple correlation coefficients (i.e., before crossvalidation), crossvalidated prediction coefficients, and the number of BSI cues included in each model are shown in Table IV-6. All confidence intervals were greater than 0, and most upper bounds exceeded .60. Shrinkage, or the difference between the optimal multiple correlation coefficient and the median crossvalidated prediction coefficient, was substantial for virtually all models. In all instances except one (Rater 2, EMOTCHAR), the upper bounds of the 95% confidence intervals were less than the corresponding optimal coefficient. Therefore, the crossvalidated prediction coefficients were significantly lower

Table IV-6. Crossvalidated[1] and Optimal Model Coefficients for Global and Dimension Ratings Predicted from Selected Biodata Screening Instrument Cues

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
			RATER 1			
INTEFF	13	147	.796	.405	.647	.755
EMOTCHAR	12	147	.694	.325	.543	.668
INTSKILL	13	147	.730	.443	.605	.696
INSIGHT	14	147	.791	.445	.631	.746
ORGSKILL	10	147	.587	.167	.417	.564
HEALTH	13	147	.815	.573	.721	.808
PERSEFF	13	147	.733	.389	.571	.694
ORGFIT	13	147	.712	.350	.565	.688
HIRE	13	147	.740	.390	.566	.695
PROMOTE	11	147	.716	.329	.573	.693
			RATER 2			
INTEFF	12	131	.768	.367	.581	.719
EMOTCHAR	7	131	.575	.232	.453	.652
INTSKILL	10	131	.663	.194	.507	.650
INSIGHT	10	131	.649	.245	.466	.612
ORGSKILL	10	130	.693	.400	.538	.622
HEALTH	9	130	.662	.422	.535	.613
PERSEFF	8	130	.582	.337	.462	.525
ORGFIT	11	127	.636	.150	.395	.577
HIRE	13	131	.726	.253	.464	.625
PROMOTE	11	131	.698	.198	.447	.624

[1]Two equal size samples were randomly allocated 500 times. Each time, the model of interest was double crossvalidated and the results were accumulated.

[2]All crossvalidated coefficients are zero-order correlations between predicted and actual ratings estimated from 500 iterations.

[3]Number of predictors in the model.

[4]Optimal or ordinary least squares multiple correlation coefficient.

than the point estimates provided by the corresponding optimal multiple correlation coefficients ($p < .05$).

Hypothesis 6

Unlike Hypothesis 3, for this hypothesis it was stated that unit weights would perform significantly better than would crossvalidated weights when modeling the ten (dimension and global) ratings from the BSI cues. This was hypothesized because there was a large number of potential predictors compared to the number of observations, particularly when the number of observations was halved during double crossvalidation. As in Hypothesis 4, unit weights were applied simultaneously with the 500 double crossvalidations used to test Hypothesis 5. During each iteration, the mean unit-weighted prediction coefficient was computed on the same two samples used to estimate the mean crossvalidated prediction coefficient. The median and 95% confidence intervals for the unit-weighted prediction coefficients and the difference between squared crossvalidated and unit-weighted coefficients are shown for each rater in Table IV-7.

In every instance but one (Rater 2, HEALTH), the 95% confidence interval for the unit weighted model included the optimal multiple correlation coefficient. So unit-weighted models of these raters' decisions did not shrink significantly from the optimal ordinary least squares solution. This is in contrast to the crossvalidated prediction coefficients (see Hypothesis 5, above) which were significantly different from the optimal coefficients. This contrast is confirmed by the results presented in Table IV-7, where in every instance the unit-weighted models outperformed the crossvalidated models ($p < .05$).

Table IV-7. Comparisons Among Crossvalidated[1] and Unit-Weighted Model Coefficients for Global and Dimension Ratings Predicted from Selected Biodata Screening Instrument Cues

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
			RATER 1			
INTEFF						
Unit-weighted	13	147	.796	.701	.763	.818
Difference				-.451	-.215	-.057
EMOTCHAR						
Unit-weighted	12	147	.694	.603	.687	.746
Difference				-.338	-.201	-.081
INTSKILL						
Unit-weighted	13	147	.730	.662	.714	.760
Difference				-.339	-.180	-.055
INSIGHT						
Unit-weighted	14	147	.791	.671	.747	.805
Difference				-.388	-.203	-.028
ORGSKILL						
Unit-weighted	10	147	.587	.476	.578	.649
Difference				-.297	-.162	-.053
HEALTH						
Unit-weighted	13	147	.815	.729	.789	.837
Difference				-.331	-.156	-.011
PERSEFF						
Unit-weighted	13	147	.733	.650	.717	.774
Difference				-.366	-.220	-.081
ORGFIT						
Unit-weighted	13	147	.712	.628	.705	.765
Difference				-.358	-.209	-.079
HIRE						
Unit-weighted	13	147	.740	.655	.729	.783
Difference				-.385	-.247	-.109
PROMOTE						
Unit-weighted	11	147	.716	.641	.763	.711
Difference				-.393	-.210	-.063
			RATER 2			
INTEFF						
Unit-weighted	12	131	.768	.679	.749	.797
Difference				-.449	-.269	-.087
EMOTCHAR						
Unit-weighted	7	131	.575	.496	.575	.652
Difference				-.247	-.127	-.033

Table IV-7 (Continued)

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
INTSKILL						
Unit-weighted	10	131	.663	.570	.659	.727
Difference				-.350	-.196	-.077
INSIGHT						
Unit-weighted	10	131	.649	.560	.645	.709
Difference				-.363	-.209	-.081
ORGSKILL						
Unit-weighted	10	130	.693	.667	.685	.700
Difference				-.330	-.205	-.097
HEALTH						
Unit-weighted	9	130	.662	.633	.648	.662
Difference				-.262	-.151	-.054
PERSEFF						
Unit-weighted	8	130	.582	.545	.571	.586
Difference				-.223	-.121	-.052
ORGFIT						
Unit-weighted	11	127	.636	.524	.633	.709
Difference				-.384	-.254	-.125
HIRE						
Unit-weighted	13	131	.726	.617	.713	.768
Difference				-.459	-.325	-.153
PROMOTE						
Unit-weighted	11	131	.698	.603	.683	.746
Difference				-.451	-.288	-.108

[1]Two equal size samples were randomly allocated 500 times. Each time, the model of interest was double crossvalidated and unit-weighted.

[2]All coefficients (except optimal) are zero-order correlations between predicted and actual ratings estimated from 500 iterations. The difference coefficient is the squared crossvalidated coefficient minus the squared unit-weighted statistic for each of 500 iterations: A negative difference coefficient indicates the superiority of unit-weights.

[3]Number of predictors in the model.

[4]Optimal or ordinary least squares multiple correlation coefficient.

Hypothesis 7

It was hypothesized that models of ratings predicted by BSI cues would outperform models of the same ratings predicted by several principal components of the BSI cues. This hypothesis was prompted by discussions in the literature on rational methods for scoring biodata inventories, in which it was argued that internal variance methods discard unique variance as error (see Chapter II for a discussion of biodata scoring methods). But clinicians cite attention to such unique variance as their advantage over statistical methods (cf. Meehl, 1954). Also, it was postulated (but not tested here) that the two raters would use relatively few cues to make their decisions, instead of trying to infer logical or rational underlying structure or factors. For a rater to perform the latter process with the profusion of items on the BSI would be very complex and difficult (if not impossible).[23]

This hypothesis was tested by performing a principal components analysis on the 262 BSI cues, and using the first 82 components, which all had eigenvalues greater than 1.0. These principal components accounted for 86.4% of the total variance, probably most of the common variance. The most predictive of these 82 principal components were selected using stepwise regression for each model. For each rating by each rater, two ratings were bootstrapped -- from the selected principal components and from the BSI cues selected previously --

---

[23]Evidence not presented in this research indicated that Rater 1 actually used a written form he developed rationally to rate some or all of the BSIs. However, properly attending to all of the cues presented in all of the items would have required a computerized process.

using unit weights.[24] The difference between the squared prediction coefficients (i.e., predicting the corresponding actual ratings) for these two bootstrapped ratings was systematically resampled 500 times. The results are shown in Table IV-8. In virtually every case, the median difference favored the ratings bootstrapped from the BSI cues. In most instances the upper 95% confidence bound was less than 0, indicating that the ratings bootstrapped from BSI cues were significantly better predictors than those bootstrapped from the principal components. Comparison of Tables IV-7 and IV-8 also reveals that (a) the optimal multiple correlation coefficients for models using BSI cues were consistently larger than those for models using principal components, and (b) the optimal multiple correlation coefficients for models using principal components were always below the median estimates (and near or below the lower 95% confidence limits) for the unit-weighted models using BSI cues. There did not appear to be any reason to further investigate models using principal components.

Hypothesis 8

This hypothesis was included to test whether reliability coefficients were larger for ratings bootstrapped from BSI cues than for the actual ratings. Table IV-9 shows the intraclass correlations between raters for the actual ratings and the ratings bootstrapped from the BSI cues.

It was expected that all the bootstrapped ratings would be more reliable than the actual ratings, but only certain ratings were improved. Seven of the ten ratings were more reliable after bootstrapping, and three were less reliable.

---

[24]Given the results of the previous hypothesis tests, unit weights were presumed to be as good as, or superior to, crossvalidated weights when using principal components. So principal components were assigned unit weights corresponding to the stepwise regression results.

Table IV-8. Comparisons Among Unit-Weighted Model Coefficients for Ratings Predicted from Selected Principal Components and from Biodata Screening Instrument Cues[1]

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
			RATER 1			
INTEFF						
PCA	12	147	.661	.544	.630	.714
Difference				-.400	-.243	-.068
EMOTCHAR						
PCA	12	147	.603	.461	.594	.694
Difference				-.298	-.146	.012
INTSKILL						
PCA	13	147	.670	.545	.649	.714
Difference				-.272	-.111	.049
INSIGHT						
PCA	16	147	.697	.555	.671	.764
Difference				-.279	-.146	-.004
ORGSKILL						
PCA	8	147	.491	.348	.482	.605
Difference				-.258	-.103	.080
HEALTH						
PCA	13	147	.744	.627	.721	.791
Difference				-.325	-.162	.054
PERSEFF						
PCA	11	147	.629	.515	.617	.705
Difference				-.288	-.161	-.034
ORGFIT						
PCA	9	147	.564	.431	.547	.639
Difference				-.386	-.232	-.082
HIRE						
PCA	8	147	.628	.491	.603	.708
Difference				-.369	-.210	-.024
PROMOTE						
PCA	7	147	.663	.489	.595	.691
Difference				-.316	-.180	-.057
			RATER 2			
INTEFF						
PCA	6	131	.581	.442	.546	.637
Difference				-.477	-.303	-.143
EMOTCHAR						
PCA	6	131	.524	.371	.520	.637
Difference				-.231	-.067	.095

Table IV-8 (Continued)

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
INTSKILL						
PCA	5	131	.478	.320	.482	.597
Difference				-.378	-.229	-.072
INSIGHT						
PCA	7	131	.502	.376	.499	.603
Difference				-.352	-.183	.007
ORGSKILL						
PCA	9	130	.630	.439	.598	.719
Difference				-.335	-.131	.062
HEALTH						
PCA	12	130	.682	.540	.664	.757
Difference				-.135	.026	.206
PERSEFF						
PCA	5	130	.450	.306	.451	.565
Difference				-.300	-.137	.063
ORGFIT						
PCA	8	127	.559	.457	.556	.640
Difference				-.258	-.102	.039
HIRE						
PCA	5	131	.478	.290	.466	.627
Difference				-.495	-.314	-.134
PROMOTE						
PCA	7	131	.561	.397	.534	.642
Difference				-.382	-.202	-.030

[1]Two equal size samples were randomly allocated 500 times. Each time, the model of interest was unit-weighted from principal components.

[2]PCA coefficients are zero-order correlations between predicted and actual ratings estimated from 500 iterations. The difference coefficient is the difference between squared coefficients corresponding to principal components and BSI cues for each of 500 iterations: A negative difference coefficient indicates the superiority of the BSI cues.

[3]Number of predictors in the model.

[4]Optimal or ordinary least squares multiple correlation coefficient.

Table IV-9. Intraclass Correlation Estimates of Interrater Reliability for Actual Ratings and Ratings Bootstrapped from Biodata Screening Instrument Cues for Two Raters

Rating Scale	N[1]	Original Intraclass Correlation	Bootstrapped Intraclass Correlation
INTEFF	125	.483	.571
EMOTCHAR	125	.231	.361
INTSKILL	125	.233	.152
INSIGHT	125	.295	.197
ORGSKILL	124	.341	.405
HEALTH	124	.285	.452
PERSEFF	124	.253	.189
ORGFIT	121	.442	.648
HIRE	125	.545	.668
PROMOTE	125	.450	.554

[1]The number of observations is reported for bootstrapped intraclass correlations only.

The three ratings which were less reliable when bootstrapped also were among the least reliable ratings when not bootstrapped. So reliability coefficients for most ratings were improved by bootstrapping, and overall interrater reliability was improved.

Hypothesis 9

Hypothesis 9 concerned how well the global ratings could be modeled by (a) the actual dimension ratings, (b) the bootstrapped dimension ratings, and (c) the BSI cues. It was hypothesized that the BSI cues could be used to model the global ratings as accurately as could either the actual or bootstrapped dimension ratings (thus eliminating the need for actual ratings of BSIs). Table IV-10 shows double crossvalidated coefficients for models of the global variables predicted from bootstrapped dimension ratings. The dimensions included were those selected during the test of Hypothesis 1 (i.e., no new stepwise regressions were run). Also shown are coefficients computed earlier during the tests of Hypotheses 1 and 5, which corresponded to global ratings modeled using dimension ratings and BSI cues, respectively.

It is apparent from the results that the coefficients for models bootstrapped from dimension ratings and models bootstrapped from BSI cues were identical for practical purposes.[25] Also apparent is that the models bootstrapped from the bootstrapped dimension ratings were less predictive of the global ratings than were the other models. For Rater 1, the upper 95% confidence bounds of the coefficients corresponding to the bootstrapped dimension ratings are below the median coefficients for either of the other two models, and are

---

[25]This result does not, however, indicate that the models will predict the same, only that they account for about the same amount of variance in the global ratings. It is possible that they predict different aspects of the ratings with approximately equal validity.

Table IV-10. Comparisons Among Prediction Coefficients of Unit-weighted Models of Global Ratings Bootstrapped from Selected Dimension Ratings, Bootstrapped Dimension Ratings, and Biodata Screening Instrument (BSI) Cues[1]

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
MODELS USING BOOTSTRAPPED DIMENSION RATINGS						
Rater 1						
HIRE	5	147	.658	.522	.621	.699
PROMOTE	3	147	.596	.493	.584	.659
Rater 2						
HIRE	4	126	.458	.435	.458	.490
PROMOTE	2	130	.562	.542	.563	.583
MODELS USING ACTUAL DIMENSION RATINGS						
Rater 1						
HIRE	5	184	.758	.713	.724	.731
PROMOTE	3	183	.691	.594	.666	.724
Rater 2						
HIRE	4	165	.706	.653	.722	.781
PROMOTE	2	169	.685	.613	.686	.747
MODELS USING BSI CUES						
Rater 1						
HIRE	13	147	.740	.655	.729	.783
PROMOTE	11	147	.716	.641	.711	.763
Rater 2						
HIRE	13	131	.723	.617	.713	.768
PROMOTE	11	131	.698	.603	.683	.746

[1]Two equal size samples were randomly allocated 500 times. Each time, the model of interest was unit-weighted, and the results were accumulated to create distributions of 500 parameter estimates.

[2]All coefficients (except optimal) are zero-order correlations between predicted and actual ratings estimated from 500 iterations.

[3]Number of predictors in the model.

[4]Ordinary least squares multiple correlation coefficient.

quite near the lower 95% confidence bounds for the other two models. The upper 95% confidence bounds for the coefficients corresponding to the bootstrapped dimension ratings of Rater 2 are less than the lower 95% confidence bounds for the other two models. So the models bootstrapped from the bootstrapped dimension ratings of Rater 2 were significantly less predictive of the global ratings of Rater 2, $p < .05$ (two-tailed). Across all models, the global ratings bootstrapped from the BSI cues were as predictive of the global ratings as were the raters' own dimension ratings.

Hypothesis 10

It was hypothesized that approximately the same cues would be important for modeling the eight dimension ratings as would be important for modeling the two global ratings. This was based on the postulation that raters would pay attention to a relatively limited number of cues on the BSI. This was not presumed to be a test of which cues actually were used by the raters in making their decisions, but of whether all of the decision models would be comprised of a small number of powerful cues. This was tested by noting how many of the cues which were included in models of each global rating were also important in individual dimension ratings.

For Rater 1, 9 of the 13 cues used to model HIRE also were used to model one or more of the dimension ratings used to model HIRE. Only 4 of the 11 cues used to model PROMOTE were also important for modeling the dimension ratings most relevant to PROMOTE. For Rater 2, 6 of the 13 cues were used to model both HIRE and the dimension ratings most related to HIRE. But just 2 of the 11 cues used to model PROMOTE for Rater 2 were part of the models of relevant dimension ratings. The total number of cues used for modeling all ratings and how frequently cues were used are presented in Table IV-11. These

Table IV-11. Comparisons of Use of Biodata Screening Instrument Cues to Model Global and Dimension Ratings for Two Raters

Statistic	Rater 1	Rater 2
Total cues used	125	101
Total unique cues	71	60
Cues used one time	44	36
Cues used two times	11	17
Cues used three times	10	1
Cues used four times	3	2
Cues used five times	1	4
Cues used six times	2	0

results are not consistent with the hypothesis that only a limited set of cues could be used to model all of the ratings. It appears that ratings were the result of related, but different decisions.

Research Questions

Question 1

The first question of interest was whether the global ratings were valid for predicting the many criteria collected by the corporation. This was investigated by systematically resampling the Pearson product-moment correlations between each criterion and each potential predictor (i.e., the actual global ratings and the global ratings bootstrapped from the BSI cues). The results are presented in Table IV-12, and show that few of the criteria were predicted well by any of the ratings. The upper and lower 95% bounds of the distributions of coefficients were computed; however, to save space, only the medians of the distributions are shown. The relationships were not always in directions consistent with the nature of the criteria and dimensions involved. Also, given the large number of correlations computed, even significant correlations may be spurious.

Question 2

Since biodata inventories have been shown consistently to be valid predictors of organizational criteria, the validity of the BSI for predicting the available criteria was investigated. The same steps were used to develop a prediction equation as were used previously to create decision models: Stepwise regression was used to select potential predictors, and modified double crossvalidation was used to estimate parameters and compute confidence intervals. The results are shown in Table IV-13, for all of the available

Table IV-12. Systematically Resampled Validity Coefficients[1] for Global Ratings by Two Raters Predicting Organizational Criteria

Criterion	HIRE	PROMOTE	Bootstrapped HIRE	Bootstrapped PROMOTE
		RATER 1		
ITEMS	-.026	-.203[2]	-.061	-.131
WAGE	-.022	.051	-.035	.040
ESV	-.087	-.247[3]	-.128	-.262[3]
FULLTURN	.060	.023	.112	.034
PARTTURN	.057	.031	-.034	-.065
TOTSALES	.102	-.158	.090	-.134[2]
CONTEXP	-.131	.094	-.178[3]	.070
SHRINK	.098	.090	.024	.069
GPS	-.034	-.054	.051	.030
PERFSCOR	.123	-.082	.142	-.035
AUDIT	.052	.113	.139	.001
EMPATT	-.106	.010	-.115	-.018
COMPOSIT	.019	-.079	.097	-.062
		RATER 2		
ITEMS	.048	.154	.106	.148
WAGE	.053	.029	.123	.079
ESV	-.081	-.194[3]	-.130	-.084
FULLTURN	.083	.128	.220[3]	.119
PARTTURN	.159[2]	.073	.159	.136
TOTSALES	-.041	-.159	-.107	-.229[3]
CONTEXP	.032	.063	.091	.170
SHRINK	.134	.124	.259[3]	.015
GPS	-.069	-.132	-.165	-.084
PERFSCOR	.037	.014	-.063	-.061
AUDIT	-.012	.029	.025	.164
EMPATT	-.016	.030	.125	.070
COMPOSIT	-.039	-.055	-.140	.002

[1]Pearson correlation coefficients were computed on 500 samples drawn at random with replacement. Median coefficients are reported here. The sample sizes were 92 for Rater 1 and 80 for Rater 2.

[2]$p < .10$

[3]$p < .05$

Table IV-13. Crossvalidated, Unit-weighted, and Optimal Regression Coefficients for Criteria Predicted from Selected Biodata Screening Instrument Cues[1]

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
ITEMS						
Crossvalidated	10	144	.596	.237	.391	.512
Unit-weighted				.541	.578	.602
Difference				-.282	-.189	-.083
WAGE						
Crossvalidated	15	149	.691	.274	.472	.616
Unit-weighted				.585	.655	.707
Difference				-.347	-.225	-.084
ESV						
Crossvalidated	12	143	.696	.356	.525	.652
Unit-weighted				.615	.691	.748
Difference				-.350	-.224	-.093
FULLTURN						
Crossvalidated	10	122	.675	.455	.572	.641
Unit-weighted				.618	.640	.654
Difference				-.215	-.096	-.006
PARTTURN						
Crossvalidated	12	141	.682	.159	.420	.611
Unit-weighted				.558	.651	.723
Difference				-.406	-.272	-.098
TOTSALES						
Crossvalidated	11	149	.623	.140	.359	.535
Unit-weighted				.501	.604	.682
Difference				-.380	-.242	-.094
CONTEXP						
Crossvalidated	15	147	.694	.133	.463	.645
Unit-weighted				.650	.691	.760
Difference				-.474	-.282	-.133
SHRINK						
Crossvalidated	10	147	.585	.130	.386	.546
Unit-weighted				.476	.574	.653
Difference				-.306	-.186	-.063

Table IV-13 (Continued)

Coefficient[2]	p[3]	N	Optimal[4]	Lower 95% Bound	Median	Upper 95% Bound
GPS						
Crossvalidated	10	147	.678	.145	.393	.552
Unit-weighted				.501	.714	.626
Difference				-.412	-.251	-.055
PERFSCOR						
Crossvalidated	8	141	.615	.261	.463	.613
Unit-weighted				.519	.597	.670
Difference				-.275	-.150	-.021
AUDIT						
Crossvalidated	13	140	.705	.390	.525	.628
Unit-weighted				.660	.685	.700
Difference				-.346	-.217	-.092
EMPATT						
Crossvalidated	12	131	.676	.292	.510	.665
Unit-weighted				.589	.665	.737
Difference				-.333	-.206	-.064
COMPOSITE						
Crossvalidated	11	124	.709	.447	.563	.638
Unit-weighted				.676	.690	.704
Difference				-.305	-.186	-.089

[1]Two equal size samples were randomly allocated 500 times. Each time, the model of interest was double crossvalidated, unit-weighted, and the results were compared.

[2]All unit-weighted and crossvalidated coefficients are zero-order correlations between predicted and actual criteria estimated from 500 iterations. The difference coefficient is the squared crossvalidated coefficient minus the squared unit-weighted statistic for each of 500 iterations: A negative difference coefficient indicates the superiority of unit-weights.

[3]Number of predictors in the model.

[4]Optimal or ordinary least squares multiple correlation coefficient.

criteria. As can be seen, the BSI was approximately as predictive of individual criteria as it was of raters' decisions.

The criterion validity of statistical combinations of the eight dimension ratings was investigated next because statistical combinations of the BSI items were highly valid. It was thought that the raters may have measured the dimensions but were unable to combine them to predict these criteria. The results are shown in Table IV-14, and they indicate that statistical combinations of the ratings were valid for some of the criteria.

Question 3

This research question concerned the relative usefulness of the BSI, the Weighted Application Blank (WAB) described earlier, and a hypothetical interviewer for predicting the composite criterion.[26] Three utility estimates were computed for each test: one each for the median estimate and the upper and lower confidence limits. The validity of the BSI was estimated by the coefficients computed for Question 2, above. The validity of the hypothetical interviewer was estimated by coefficients computed for HIRE bootstrapped from the data for Rater 1 in Question 1, above. Costs and selection statistics were obtained from the corporation's Department of Human Resources:

1. The standard deviation of the criterion was estimated at 20% of output (Schmidt & Hunter, 1983). This allowed a generic measure of utility, instead of a dollar value.[27]

---

[26]The WAB was validated using earlier composite criterion estimates.

[27]Another standard deviation suggested by those authors was 40% of mean salary. Using this formula, the standard deviation of the criterion would be estimated at $12,000.00.

Table IV-14. Unit-weighted Coefficients for Criteria Predicted from Selected Actual and Bootstrapped Dimension Ratings[1]

Coefficient[2]	p[3]	N	Lower 95% Bound	Median	Upper 95% Bound
		RATER 1			
WAGE					
Actual	3	134	.071	.215	.360
Bootstrapped			.127	.272	.426
Difference			-.117	-.027	.052
ESV					
Actual	2	128	.024	.182	.330
Bootstrapped			.129	.262	.398
Difference			-.119	-.032	.041
FULLTURN					
Actual	2	109	.021	.241	.427
Bootstrapped			-.124	.085	.282
Difference			-.008	.042	.140
TOTSALES					
Actual	2	134	.087	.253	.420
Bootstrapped			-.028	.123	.263
Difference			-.013	.047	.150
CONTEXP					
Actual	1	132	.028	.180	.325
Bootstrapped			.005	.150	.305
Difference			-.032	.008	.059
SHRINK					
Actual	3	132	.202	.326	.424
Bootstrapped			.061	.212	.376
Difference			-.068	.057	.162
GPS					
Actual	2	132	-.110	.057	.300
Bootstrapped			.053	.184	.378
Difference			-.084	-.024	.006
PERFSCOR					
Actual	1	128	.086	.240	.405
Bootstrapped			-.094	.122	.300
Difference			-.013	.040	.113
AUDIT					
Actual	3	126	.057	.297	.495
Bootstrapped			-.009	.225	.456
Difference			-.045	.028	.115
EMPATT					
Actual	3	116	-.013	.145	.329
Bootstrapped			.183	.346	.492
Difference			-.221	-.090	.013

Table IV-14 (Continued)

Coefficient[2]	p[3]	N	Lower 95% Bound	Median	Upper 95% Bound
COMPOSITE					
Actual	2	110	.013	.211	.381
Bootstrapped			.033	.206	.369
Difference			-.080	.001	.095
		RATER 2			
ITEMS					
Actual	1	115	-.021	.167	.302
Bootstrapped			-.088	.041	.147
Difference			-.005	.024	.081
WAGE					
Actual	2	118	.163	.321	.447
Bootstrapped			-.153	.022	.185
Difference			.009	.097	.193
ESV					
Actual	1	115	-.020	.169	.358
Bootstrapped			-.004	.171	.327
Difference			-.058	.000	.069
FULLTURN					
Actual	1	96	.118	.306	.479
Bootstrapped			.055	.232	.401
Difference			-.047	.036	.141
PARTTURN					
Actual	3	112	.097	.253	.408
Bootstrapped			.010	.179	.334
Difference			-.039	.027	.115
TOTSALES					
Actual	2	116	.003	.200	.382
Bootstrapped			-.032	.170	.360
Difference			-.060	.007	.089
CONTEXP					
Actual	1	117	-.056	.134	.298
Bootstrapped			-.116	.046	.182
Difference			-.015	.012	.073
SHRINK					
Actual	2	117	.116	.281	.429
Bootstrapped			-.019	.158	.312
Difference			-.055	.055	.173
GPS					
Actual	1	118	.063	.199	.343
Bootstrapped			.009	.167	.291
Difference			-.028	.011	.075

Table IV-14 (Continued)

Coefficient[2]	p[3]	N	Lower 95% Bound	Median	Upper 95% Bound
PERFSCOR					
Actual	3	114	.082	.267	.419
Bootstrapped			.014	.226	.384
Difference			-.073	.022	.131
EMPATT					
Actual	2	103	-.096	.092	.253
Bootstrapped			.099	.284	.429
Difference			-.177	-.064	.022
COMPOSITE					
Actual	1	98	-.092	.091	.238
Bootstrapped			-.037	.185	.366
Difference			-.105	-.020	.015

[1]Random samples were drawn with replacement 500 times. Each time, the model of interest was correlated with the criterion and the results were compared.

[2]Coefficients are zero-order correlations between predicted and actual criteria estimated from 500 iterations. The difference coefficient is the squared coefficient for the actual dimension rating minus the squared coefficient for the bootstrapped ratings for each of 500 iterations: A negative difference coefficient indicates the superiority of bootstrapped ratings.

[3]Number of predictors in the model.

2. The total number of applicants, the number hired, and the selection ratio were estimated at 3500, 310, and .09, respectively.

3. The average increase in the standardized criterion with a selection ratio of .09 was computed for each criterion using data on all stores and managers, and was estimated at 1.6297.

A utility equation presented in Arnold, Rauschenberger, Soubel, and Guion (1982) was used to estimate the average utility increase over random selection per selected applicant for each test. Using this equation, the utility of a test is the product of the validity coefficient, the average increase of the standardized criterion, and the standard deviation of the criterion. Although not appropriate in the context of a generic utility analysis, the cost of testing may be computed by dividing the cost of administration per applicant by the selection ratio. This cost may then be subtracted from the utility when a dollar value is used for the standard deviation of the criterion. Usually cost is insignificant; however, when interviewing is involved, cost may be important. Results are presented below, but it must be emphasized that these results are comparative only, because the premises on which they were based were tenuous at best. The following results are scaled in percentage increases in output expected when using each of the tests:

Test	Lower Bound	Median	Upper Bound
Weighted Application	4.3%	9.6%	13.1%
BSI	22.0%	22.5%	22.9%
Interviewer	-3.6%	3.2%	10.4%

CHAPTER V

DISCUSSION AND CONCLUSIONS

Discussion

The estimated accuracy with which the data were coded, reported in Chapter IV at about 96%, was within the expected range. Although keypunch operators usually are expected to err only a fraction of one percent of the time, coding data for this study required more than keypunch skills: The coder was required to read each BSI, interpret the applicant's handwriting, and make judgments about the content. She was required to categorize answers; judge humor, grammar, and conciseness; and discriminate among positive and negative motives, goals, and explanations, for example. There was considerable room for error in this task even though the scoring key contained explicit instructions. The risk of increased coding errors (incurred when deviating from the relatively errorless process of keypunching multiple choice responses or machine-scoring optically scanned answer sheets) was taken because open-ended items were expected to provide a better sample of applicants' behavior, skills, and traits (Kahn & Cannell, 1957). Fortunately, the error rate for the latter was reasonably low.

Analysis of the comparison of the two groups of managers (those who responded and those who did not) suggests that, for those comparisons, the two groups were similar. The slight age difference between respondents and nonrespondents may have indicated a difference in some underlying tendency related to age (e.g., maturity, or attitudes toward the corporation), or it may have been incidental to this research. In either case, inferences about samples of managers other than the one investigated here must be made with at least

one caveat: The results reported in this research may be different for any sample with a different mean age. Obviously, inferences made about other populations of managers probably are unwarranted because it is unlikely that organizational conditions, manager characteristics, and so forth, would be the same.

Practice effects for raters were not assessed because the prescribed sequence in which the BSIs were to be rated was not followed. However, the raters thought they were rating BSIs in the proper sequence, and they did not order them explicitly according to some attribute[28]. Unfortunately, there is no record of the order in which they were actually rated. Rater 2 did not date any of his rating sheets, so his order effects were not investigated further. But Rater 1 dated some of his ratings, and the few that were completed very early (e.g., August, 1986) appeared to have a wider range of ratings and many more noninteger ratings; ratings apparently completed later (e.g., fall, 1987) were generally integers with a narrower range. This order effect was not tested for significance because there were only a few for which the date could be established, and there was no indication of how they were ordered within each date. Also, even though Rater 1 commented that he changed his rating method during the research[29], the results presented in Chapter IV showed that much of the variance in his decisions was predicted, regardless of changes he made. Although it would have been interesting to test the hypothesis that

---

[28]It appears that the coder assigned the random numbers (which were also used to match managers with criteria) and the raters applied the same or another random number list to randomize the random order.

[29]He did not specify how or why he changed his rating method.

different models would have fit different sequences of ratings, there was no evidence in the literature to support that hypothesis.

As noted in Chapter IV, the ratings of Rater 1 had larger standard deviations and wider ranges than did those of Rater 2. Also, the squared multiple correlations (SMCs) for Rater 1 appeared to be more variable on many of the dimensions than did the SMCs for Rater 2, and the models developed from the BSI cues to describe the ratings of Rater 1 used more predictors than did the models for Rater 2. Taken together, these results implied that Rater 1 may have discriminated more among the eight dimension ratings and two global ratings than did Rater 2; however, these results were not incontrovertible. Differences were expected because Rater 1 was the principal designer of the BSI, and he may have developed the BSI in such a way as to identify individuals similar to, or liked by, himself, but less attractive to Rater 2 (Orpen, 1984). It is also possible that important data were not as salient to Rater 2 as they were to Rater 1, because of the design of the BSI (Schuh, 1973).

The squared multiple correlations (SMCs) generally were better (i.e., lower) than was expected. Higher SMCs indicate more colinearity among ratings; more colinearity suggests less discrimination among scales and the presence of rater halo. However, even if halo were not the cause, more colinearity is associated with redundant variables and less stable regression weights. Neter et al. (1985) recommended corrective measures (e.g., using the correlation transformation or ridge regression) when SMCs exceed about .50 and approach .90. With the exception of PERSEFF and ORGFIT, both raters' SMCs were near or less than .50. The correlation transformation was used for all computations, even though the SMCs were moderate.

Interrater reliability was poor, particularly for ratings based on such structured data. Low reliability was expected for some of the poorly defined or ambiguous dimensions (e.g., EMOTCHAR and PERSEFF). Although not high in absolute terms, the reliability coefficients for most of the ratings (e.g., INTEFF and HIRE) were consistent with those summarized in Arvey and Campion (1982), and were improved when bootstrapped from the BSI cues. Reliability coefficients were particularly low for the interpersonal dimensions which was unexpected because interpersonal skills and sociability were dimensions purportedly measured well in employment interviews. It is possible that nonverbal cues were more important for assessing social skills than has been shown previously; however, models for the social dimensions were consistent, even though the raters did not appear to measure similar dimensions. The extremely high intrarater reliability coefficients (.878 and .909) indicated that systematic differences between applicants accounted for most of the variance in ratings, while a relatively small amount was due to indeterminacy and error. This result was probably due in part to rater halo, as well as consistency and accuracy.

The criterion battery and the composite criterion means were essentially the same for respondents and nonrespondents. This was important for two reasons: First, it meant that simply completing the BSI was not predictive of organizational performance. Second, it suggested that respondents and nonrespondents were equal on sanctioned organizational criteria. If only good or poor managers were included in the research, results would not be viewed as generalizable; however, generalizations from sample to population are more likely valid when there are no differences between the two.

Also of note, although incidental to the purpose of the research, was the use of modified double crossvalidations to estimate upper and lower confidence bounds for various estimates of parameters. This procedure was designed using Ayabe's (1985) idea to double-crossvalidate the model of interest until the change in the mean squared multiple correlation coefficient diminishes sufficiently, and Efron's (1982) bootstrap method of systematic resampling to create nonparametric distributions of estimates. The reasons the modified double crossvalidation was used in this research were (a) it provided a "feel" for the data by repeatedly sampling the available observations, (b) it used all of the data instead of summary statistics to test complex hypotheses (e.g., whether unit or crossvalidated weights were more predictive), (c) the distribution of estimates it provided could be inspected free of most assumptions, and (d) it was a relatively simple means to estimate the variability of complex parameter estimates without computing (or relying on the assumptions of) analytic formulae. Systematic resampling of univariate statistics was used for the same reasons. Note that variables which were not significant after 500 iterations, were left in the model and the results reported. But prediction models would be fine-tuned if used in actual selection. Also note that the double crossvalidation procedure yields mean prediction coefficients for both unit and optimal analyses, so the range of coefficients was somewhat restricted.

The test of the first hypothesis showed that global ratings could be predicted (modeled) from knowledge of individual dimension ratings. The model coefficients were large and statistically significant. This result was expected for either of two reasons: (a) because rater halo allowed virtually any rating to be predicted from any other ratings, or (b) because the process of rating individual dimensions caused the raters to consider the utility of the different

dimensions for assessment (Belec & Rowe, 1983; Arvey & Campion, 1982). If the former were the reason for accurate models, then the models should have been composed of one or two dimension ratings predicting both HIRE and PROMOTE in similar fashion. If the latter were the reason, the models should have included different dimension ratings for HIRE and PROMOTE, perhaps more than two per model. Further investigation revealed that, for Rater 1, HIRE and PROMOTE were modeled by a total of eight dimension ratings, and six of them were unique to a model. Both models of the global ratings had at least one variable that was unique to that model. For Rater 2, HIRE and PROMOTE were modeled by a total of six dimension ratings, and the models had only one dimension rating in common. Given the moderate SMCs and the variety of dimension ratings involved in models of the global ratings, it was unlikely that the accuracy of the models was due solely to rater halo. Also, the models were dissimilar between raters, an indication that perhaps they considered different dimensions important; or perhaps they had different conceptions of what defined the dimensions. Since the models using the BSI cues did not include approximately the same cues across ratings, the global ratings were probably measuring different dimensions.

The test of the second hypothesis revealed that the disjunctive model clearly performed worse when attempting to model global ratings from dimension ratings than did the linear compensatory model. Thus, neither rater would rate an applicant favorably with a single overwhelming dimension rating (unlike a football recruiter who would offer a position to someone who only could punt). So these raters probably would rate an intelligent applicant low if he or she had poor interpersonal skills, poor organizational skills, and so forth. Conversely, the decisions of Rater 1 appeared to be modeled adequately with

the conjunctive model: He might be inclined to rate favorably an applicant with just adequate attributes on all dimensions, perhaps more so than an applicant with some favorable and some unfavorable attributes -- even if the sums of the ratings were similar for each applicant. In spite of the good fit of the nonlinear conjunctive model for Rater 1, it offered no practical increment in predictability over the linear compensatory model. Therefore, linear compensatory models adequately predicted all global ratings from dimension ratings, and were used in subsequent analyses. Interactive models were not fit as part of this research because there was little in the literature to recommend such measures, particularly since the simple linear models performed well without crossproducts or other nonlinear terms. It was felt that fitting more complex models would risk overfitting and sample-specific results without providing much more prediction. These decisions were supported by the conclusions of other researchers who found little practical use for nonlinear models (e.g., Dawes, 1971; Dawes & Corrigan, 1974; Goldberg, 1968; Hammond et al., 1964; Hogarth, 1980; Hoffman, 1960; Sydiaha, 1959, 1962).

The next hypothesis test, comparing unit and crossvalidated prediction weights for modeling global ratings from dimension ratings, found no practical difference between the two weighting schemes. This result was expected because, as Schmidt (1971) argued, the superiority of regression or unit weights depends essentially on sample size and the number of predictors in the model[30].

The test of the fourth hypothesis found that bootstrapping raters' decisions did not make them more reliable than the original decisions. However, in this case the ratings were being bootstrapped from other ratings. So the reliability

---

[30]The discussion by Dawes and Corrigan (1974) about the superiority of unit weights was presented in the context of applying unit weights to clinical judgments to predict distal criteria, not to predict other ratings by the same rater.

of the ratings used in the model would constrain the reliability of the bootstrapped ratings, making such bootstrapping appear fruitless. But the ratings might benefit from the smoothing or averaging effect of the combination of other ratings. Unit weights were used in the bootstrap models, so the models were actually averaging the ratings. On reflection, the results were consistent with this discussion: The reliability of HIRE (the more reliable the two ratings) decreased moderately, while the reliability of PROMOTE increased. The global ratings nearly were equal after bootstrapping, an effect analogous to averaging. The bootstrapped reliability coefficients were less than .70; however, the intraclass correlations of .50 to .55 were better than has been found in many studies of the interview, and indicated some consistency between raters with diverse backgrounds and points of view.

The results of the tests of Hypotheses 1, 2, and 3 established relationships among the dimension and global ratings, and suggested that more was at work than simple rater halo. Also affirmed was the equivalence of unit weights in modeling the global ratings. It was shown that modeling global ratings from component ratings was helpful only to the extent that the components were more reliable, more valid, or more easily obtained than the global ratings.

The next analyses were directed to testing the most important hypothesis in this research: that accurate models of raters' decisions could be gleaned from a substantial array of ecologically valid[31] cues. The results were to be used to generalize previous modeling research which used contrived cues and orthogonal designs, and to estimate how consistently raters used cues to make decisions in

---

[31]At least, more valid than most previous research related to the employment interview.

a realistic task[32]. Again, it is important to note that the cues included in the models were not necessarily the same models the raters used; the cues included in the models were those which were best related to the raters' decisions, not necessarily those to which the raters attended.

The results from the test of Hypothesis 5 were highly significant, and confirmed that these raters' decisions could be modeled well from the basic set of stimuli. The lower bound of the confidence intervals for crossvalidated prediction coefficients was greater than zero for every model. These coefficients indicated how consistently raters used cues to make their decisions. In every model but one, the median coefficient was larger than .40. The models shrank considerably when crossvalidated, and the ratio of observations to predictors was about 12:1 for Rater 1 and 13:1 for Rater 2 (compared to 46:1 and 56:1 for the dimension-based models). So unit weights were expected to provide better indications of the raters' consistency (Hypothesis 6). The results showed that unit weights predicted up to 45% more total variance in raters' decisions than did the crossvalidated coefficients, and all differences were significant ($p < .05$). Also, the median coefficient approached or exceeded the optimal coefficient in most cases. More important than confirming the hypothesis was finding that unit weights (again, simply summing selected standardized scores) accounted for more than 50% of the variance in many ratings, and each raters' consistency was shown to be significantly high -- both practically and statistically. It was possible to compare the consistency of the raters using these results by noting which confidence intervals did not overlap. Rater 1 was expected to be more consistent than Rater 2 because he chose the

---

[32]BSIs would be read instead of standard applications (although the application would still be completed) for preinterview screening if a scoring model were not successfully created.

rating dimensions (Orpen, 1984) and created corresponding items. In fact, Rater 1 was significantly more consistent than Rater 2 in rating HEALTH and PERSEFF, but Rater 2 was more consistent in rating ORGSKILL. However, from simple inspection of the unit-weighted coefficients, one could say that Rater 1 was slightly more consistent than Rater 2 overall.

The logical next step was to see whether principal components would be more useful than individual cues for modeling raters' decisions, because the two raters appeared to be very consistent in rating the BSIs on the ten scales. If principal components predicted ratings better than the cues did, then one might conclude that (a) raters inferred some statistically reproducible structure from the BSI and consistently used it to rate applicants, (b) raters used cues which were represented well by consistency in the way the BSIs were completed and coded, or (c) there was underlying structure to the BSI (and how it was completed and coded) that corresponded to raters' inferences about applicants. However, the test of Hypothesis 7 revealed that, in 19 of the 20 models, the principal components model did not correspond as well to the ratings as did the model using actual BSI cues. In 10 out of 20 models, the principal components model was significantly less predictive of the ratings ($p < .05$).

The results of this test call into question the process of rationally scoring biodata inventories: If an internal variance method (e.g., principal components analysis) were used to select or confirm items important to the developer of the inventory, then (these results suggest) the items so selected might not predict the score the developer would give after reading actual responses. This is consistent with findings in the literature about modeling, which indicated that introspective weighting (similar to rationally developing and scoring a biodata inventory, as Rater 1 did with the BSI) resulted in different decisions than did

modeling (e.g., Brookhouse et al., 1986; Shepard, 1964). One disadvantage of using cues instead of principal components is that specific responses might be rare in the population, and coincidental in the sample. However, it is unlikely that a rare response would contribute enough variance in ratings to remain in a model after 500 crossvalidations. But if it did, it would be important to include it in the model in case the same response occurred in other samples[33].

The hypothesis tested in the next analysis was that bootstrapping ratings from BSI cues would improve interrater reliability. Unlike the results for Hypothesis 4, the intraclass reliability coefficients for the ratings improved significantly when bootstrapped from the BSI cues. But because some of the reliability coefficients decreased, these results lent support to an earlier supposition: Low reliability might indicate that some of dimensions were perceived or defined differently by the raters. The three reliability coefficients which were lower for bootstrapped ratings corresponded to ratings of INTSKILL, INSIGHT, and PERSEFF, three dimensions on which raters disagreed about which cues implied the dimension. However, by the same logic, the coefficient for EMOTCHAR probably should have decreased, unless the raters had similar theories about that dimension. It is possible that dimensions other than those with the highest reliability coefficients (a) were not appropriately measured in this setting (although the researchers cited in Chapter II frequently mentioned interpersonal skills as one of the dimensions measured well by interviewer content), or (b) were undefined by these raters (i.e., one or both raters had no clear, consistent idea about what these dimensions were). However, the latter is unlikely because high consistency was indicated by the models for these

---

[33]Aamodt and Pierce (1987) and Telenson, Alexander, and Barrett (1983) discussed methods for analyzing and scoring rare responses to inventory items.

dimensions. An interesting finding not reported in Chapter IV was that INTSKILL and PERSEFF were included in models of global ratings by Rater 1, while INSIGHT was part of the model of HIRE for Rater 2; none of these dimension ratings appeared to have been used consistently by both raters, unless they were redundant with dimension ratings already in the models. The most reliable bootstrapped ratings were those one might expect to be defined similarly by the two raters: HIRE, ORGFIT, INTEFF, and PROMOTE all had reliability coefficients above .50, and the first three have been mentioned in the literature as appropriately measured in the interview (which is what the BSI purportedly simulated). These results generally were consistent with other findings reported in the interviewing literature.

The test of Hypothesis 9 revealed that using BSI cues to predict global ratings accounted for as much variance in the global ratings as predicting them from dimension ratings. In spite of discussions of rater halo, this was an impressive result, one that was not expected, given that bootstrapped dimensions did not predict as well as did the cues. However, "accounting for as much variance" does not mean "accounting for the same variance," so the correlation between the relevant unit-weighted dimensions and the relevant unit-weighted BSI cues were computed for HIRE and PROMOTE for each rater. The results are presented below:

Rater	Bootstrapped HIRE	Bootstrapped PROMOTE
Rater 1	.662 (n = 134)	.570 (n = 134)
Rater 2	.601 (n = 115)	.480 (n = 119)

The last hypothesis was not tested in the traditional statistical sense, but instead was simply a compilation of data. It was hypothesized that the raters used only a few BSI cues to rate applicants. This was expected because researchers in decision-making have often concluded that a few variables explain subjects' decisions adequately. But such research was usually conducted with multiple trials of a single decision, and may be conducive to relying on a few variables. In the present study, multiple trials of ten decisions were made. But raters still were expected to attend to only a few BSI cues, because of memory limitations, perceptual biases, and fatigue. Before further discussion, it should be noted that the cues compiled were those included in the decision models, and were not necessarily the cues to which raters attended. Also, raters may have attended to a few cues inconsistently, even though their decisions were modeled consistently by other variables. The data showed that many variables were significant in the ten decision models for each rater, and many were significant more than once. The models for both raters used an average of 11 cues per model, more than the infamous rule of thumb for memory limitations of "seven plus or minus two" variables. However, it is interesting that just 20 cues were included 70 times in the 20 models of both raters, accounting for one third of the predictors (a mean of 3.5 shared cues per model). One cue was used a total of ten times between the raters[34]. With caution, one might infer that the raters used many cues to rate applicants, but they relied on some more than others. So the evidence appears to reject the hypothesis, except that actual cue use was not determined, and any conclusion is tenuous and arguable.

---

[34]Judging from the results, it was used five or six times for Rater 1 and four or five times for Rater 2.

An interesting finding not reported in Chapter IV, but discovered while compiling the data for Hypothesis 10, concerned the kind of items included in the models. Of the 20 cues included in the models of both raters, the most often used cue was inferred from an open-ended item about what qualities the applicant admired in other people. In fact, nine of the 20 cues were inferred from open-ended items, eight were directly coded from the exploratory true-false items, two were categories checked by the applicant (the traditional biodata item), and one was a fill-in-the-blank item ("How many . . . ?"). The results supported the decision to use open-ended items to collect interview-type data, because the responses were apparently related to the decisions of the raters.

The first research question yielded results that were somewhat expected, but disappointing nevertheless. The global ratings were virtually unrelated to any of the criteria. Although some tentative conclusions could be drawn, it was felt that having seven (of 104) correlations significant at the .05 level (and only ten significant at the .10 level) indicated that there was no general relationship between these criteria and the ratings.

As presented in Chapter III, the criteria collected by the corporation were marginally, if at all, relevant to managerial performance. It is not disputed that these criteria (perhaps with the exception of performance ratings) are vital to the operation of the corporation; but the combination of type and importance to the organization makes them suspect as indicators of managerial performance. For example, most of the criteria are used by management as indicators of store operations; they are quickly attended to if they deviate from established norms (e.g., turnover or product shrink), or else there are programmed operations to assure that they do not deviate from optimal levels (e.g., wages, profits, or

expenses). So the absence of validity might be due simply to the lack of meaningful variance in the criteria.

The second research question was asked to determine how well the BSI cues could predict the criteria. The results showed that the BSI had excellent unit-weighted validity for all of the criteria: The lower 95% confidence bound exceeded .50 for every criterion measure. Specifically, BSI scores empirically derived accounted for between 46% and 50% of the variance in the composite criterion, an enormous amount relative to reported coefficients for other selection methods regardless of criterion. The combined results of the analyses of Questions 1 and 2 might indicate that the BSI gathered information very well, but raters were not able to use that information effectively (cf. Wagner, 1949). In light of this inference, the eight dimension ratings were used to predict the criteria to see if statistical combination of the component ratings would have validity (cf. Einhorn, 1972). Some of the criteria were not predicted because no ratings met the entrance criterion for the stepwise procedure. But for others, combinations of dimension ratings produced significant crossvalidated and unit-weighted validity coefficients. This was an indication that these two raters had gathered information from the BSI, but they were unable to combine that information to predict the organizational criteria. However, these organizational criteria may not be relevant to the ultimate performance of the managers.

The final question was asked to determine whether the BSI was useful for predicting managerial performance on the organizational criteria. The standard deviation of the criterion was measured by percentage output instead of dollars, so the costs of the various selection procedures could not be subtracted easily from the utility estimates. When costs of assessment were ignored, the utility

of each procedure was directly related to its validity, so the empirically scored BSI afforded the greatest potential savings for the corporation. If relative costs were included in the utility analyses, the BSI probably would remain the most useful because of its very high validity. However, the utility of the employment interview would decrease substantially if its costs were included. In economic terms, the BSI would result in a savings of more than $13,000 per manager per year if it replaced random selection, and more than $11,000 per manager per year if it replaced interviewing as the preferred method of selecting entry-level managers. Even the WAB (a two-page application blank) would result in a savings of about $5,000 per manager per year if used in place of employment interviewing.

Qualifications and Limitations

Before presenting the conclusions of this research, some limitations and qualifications need to be expressed:

1. In this research only two raters were used, so results and conclusions may not generalize to more than a small percentage of raters. However, important findings have come from studies using few raters (e.g., Dougherty et al., 1986), and Brunswick's call for ecologically sound research included multiple trials using a single subject at a time.

2. The subjects used were managers in a concurrent validation study, not applicants. It was not known (a) if real applicants would complete the BSIs with similar responses, (b) if real applicants would even complete such a demanding task, or (c) if raters would be able to infer job-related information from applicants who had not previously worked in the corporation or industry. Also, the mean age differed between respondents and nonrespondents, so these

results may not be applicable to younger applicants, especially applicants who are in their twenties (the age at which most are hired).

3. The instrument studied in this research was modeled after a structured employment interview, but conclusions about employment interviewing drawn from this research should be accepted with caution.

4. Only 262 variables were derived from the BSI data. While these variables were treated as the full array of environmental cues available to raters, many more could have been derived. For example, Anastasi and Schaefer (e.g., 1969) derived over 2000 variables in their studies of creativity in adolescents.

5. Double crossvalidation provided only estimates of mean crossvalidated validity, so the results reported here may be somewhat optimistic. However, the variances of the unit-weighted coefficients computed in this study were relatively small compared to the associated coefficients, so the confidence intervals corresponding to unit weights probably were not too biased.

6. The variance within ratings was not very large, so some relationships may have been restricted. Also, no interactive models using dimension ratings were tested (although nonlinear models were) for fear of inferring spurious decision processes from the data. Configural models may have been appropriate for modeling some global variables from dimension ratings, but they were not fitted.

7. It was not practical to include interaction terms in models using BSI cues, because there were so many cues. The linear prediction coefficients were adequate to model most, or all, ratings.

8. Most important to qualifying the results of the research questions were suspicions about the relevance of the criteria for assessing the performance of

the managers. While some of the validity coefficients were quite high, other criteria may have been more relevant to these managerial positions: The criteria used may not have been under the control of, or influenced by, the managers. For example, criterion values for managers may have been determined by store assignments. While that would be one kind of performance appraisal (i.e., being rewarded or punished by assignment to a particular store), more immediate, behaviorally-based criteria would have been preferable. Also, the reliability of the criteria over time was poor, and some of the criteria were more reliable when coefficients were computed for stores than for managers. This last result supported the argument that the criteria were more dependent on the store and corporate management than on the actions of the manager.

Conclusions

The most important conclusion drawn from this research was that the decisions of raters in a field study could be modeled successfully from a wide array of ecologically valid stimuli, proving that Brunswick's lens model paradigm has utility for both practitioners and researchers. This study was far more complex than modeling impacts of a set of two-digit numbers on the prediction of a third number (Dudycha & Naylor, 1966), or predicting the decisions of a graduate admissions committee from applicants' grade point averages and scores on the Graduate Record Examination (Dawes, 1971). Multiple regression was found to be effective for describing raters' decisions, and simple linear regression provided sufficient approximations of actual ratings: The results of this research showed that the conclusions of Hammond et al. (1964) and

Hoffman (1960) generalized to the selection of applicants by experts[35] for complex jobs. If one were to find a truly expert employment interviewer or management assessor, then this process probably would yield even better results: Better decisions imply more consistent use of environmental cues, which is the best situation in which to model decisions. The results presented here have shown that it is possible to model experts effectively so that their models can be used to select applicants. It remains to be shown that experts can validly identify potentially successful applicants.

Identifying experts who can identify good applicants is dependent on developing proper criteria with which to measure job performance. A significant finding in this research was that raters were unable to make valid global hiring decisions. This was more than a confirmation of Meehl's (1954) or Sawyer's (1966) conclusions that statistical prediction is better than clinical prediction, because the interrater reliability of the ratings and the consistency with which the raters apparently used BSI cues implied that they were indeed measuring some characteristics of the applicants or their behavior (another conclusion based on this research). Also, the dimension ratings provided good validity for many of the criteria, when the ratings were combined statistically. It is possible that prototypes used by the raters to judge applicants were inaccurate, or were accurate but did not correspond to these particular criteria (Rowe, 1984). The argument was presented that the criteria probably were irrelevant to managerial performance, although they may have been related to the managers' organizational status or standing. Therefore, a second conclusion from this research is that the problem with the validity of employment

---

[35]These raters were experts, when compared to the college students used in many interviewing studies.

interviews lies not entirely with interviewers or interviews, but also with the criteria used. Further evidence to support this conclusion can be found in Chapter II, where it was noted that:

1. Comparisons between the dimensions measured well in the interview and the criteria used in this study leave few expectations of high validity coefficients for predicting the criteria from the dimensions.

2. Researchers have found that job information can increase interviewers' validity by allowing the interview to conform to job requirements (Arvey & Campion, 1982; Orpen, 1985).

3. Latham et al. (1980), Latham and Saari (1984), Arvey et al. (1987), and others who have reported good validity coefficients for interviewers have, without exception, used behavioral or verifiable criteria which could be tied closely to the structured interview.

It is likely that the raters used in this research would have performed well in the above studies, and the decisions of interviewers in the cited studies could have been modeled successfully using written versions of their interviews (or perhaps the BSI). Furthermore, if criteria corresponding more closely to the dimensions measured in the interview were available for this research, the validity coefficients probably would have been higher. This did not affect the primary outcome of this research (which was about modeling), however, because the criteria probably had little effect on how well raters were modeled. Their global ratings were directed to predicting organizational success as they understood it, and they may have been valid at their task.

Another conclusion from this research is that instead of using the intuitive, internal, or rational methods to score biodata inventories, the decisions of the developer (or some other individual considered an expert in predicting what the

inventory purports to measure) should be modeled using a sample of subjects. It was found in this study that, at least for principal components analysis, the structure identified did not necessarily correspond to the impressions formed by raters. Also, the process of intuitive scoring (part of the rational scoring procedure) capitalizes on subjective weighting by the developer -- a process found to be ineffective for weighting predictors (cf. Dawes, 1971; Dawes & Corrigan, 1974).

The results from this study were consistent with the conclusions of Denton (1964), Palacios et al. (1966), and Walsh (1967), that substantially similar information can be collected on questionnaires and in interviews. This was concluded because there was moderate interrater reliability exhibited, the dimension ratings were valid for predicting both global ratings and criteria, and the topics of items on the BSI were essentially the same as the topics Rater 1 used in his interviews. These results also confirmed (a) Einhorn's (1972) finding that component ratings can be more valid than global ratings when statistically combined; (b) Schmitt's (1976) conclusion that raters have trouble inferring relationships among cues to predict future performance; and (c) the conclusion of Slovic et al. (1977) that professional raters may be content experts, but usually not decision-making experts.

Open-ended items required slightly more care in analyzing, but they appeared to have good utility for selection. Their most impressive aspect was that multiple inferences could be coded from each one, and they can be recoded to represent other variables for future studies.

The BSI was developed successfully using the structure of an interview. The results of the empirical validation indicated validity beyond that usually associated with biodata inventories, which generally have been acknowledged to

be among the most valid predictors for employee selection. This is important because it suggests that using theory to develop a biodata inventory is possible, worth the small effort, and does not require a complex theory. In addition to exhibiting considerable validity, the BSI gathered information which might be used to understand why some managers succeed with the corporation. This could provide important information with which to develop or refine models of managerial success.

The excellent validity of the BSI when scored empirically, is accompanied by its excellent validity for modeling raters' decisions. It appears that biodata inventories may be valid for almost any prediction task for which there are items on which respondents can be classified. So the BSI scoring key could be changed to predict a wide range of criteria. It is the conclusion of the author that the empirical validity of the BSI probably is due to its usefulness for identifying categories of applicants; applicants who were differentiated on the criteria were also differentiated by one of the virtually unlimited possible combinations of BSI cues. However, it is also the conclusion of the author that the BSI was useful for modeling raters' decisions because the items included in models corresponded closely to items the raters actually used to make their decisions. The alternative is that BSI scores simply differentiated approximately the same respondents as did the ratings. While there was no conclusive evidence of this difference, research is suggested later that might shed light on these two roles of the BSI -- as categorizing tool and stimulus.

An instrument, such as the BSI, may not measure well some dimensions (e.g., interpersonal skills). The face-to-face interview may be necessary to adequately measure some dimensions (e.g., interpersonal skills), according to some researchers and authors. Instruments, when scored with decision models,

are probably quite dependent on the skills and abilities of the individual rater, as was evidenced by the differences between raters in models, coefficients, and interpretation of dimensions. So measuring dimensions may simply be a function of who is modeled.

An alternative use for the BSI, is to use the completed inventory as the source of information for those making employment decisions (e.g., recruiters). It could be read instead of, or in addition to, an employment application. As reported here, raters seemed to gather a broad range of data from reading the BSI. The cost would be less than for face-to-face interviewing (particularly for long distance recruiting), and it would approximate a highly constrained and structured interview. The BSI could be used when selecting for a few positions, or for a very small company. A rating form should be used that corresponds to the job analysis results and whatever performance criteria would be used ultimately to evaluate job incumbents.

It was shown in this research that the BSI, or another similarly developed instrument, could be used in applications where a company wanted to predict the decisions of an interviewer or recruiter without regard to operationalized criteria. For example, if there were an interviewer retiring or leaving who had been perceived as particularly helpful, a procedure similar to the one presented here could be used to model global or component ratings and used inexpensively to make consistent decisions. Or if the ratings desired were of emotional characteristics (or some other dimension) as perceived by a particular person, written tests could be administered and scored less expensively and more quickly than individual interviews could take place. A company might have use for one of the more controversial tests, such as an honesty test, but may feel that the existing paper-and-pencil tests are not attractive. Because such a

criterion is elusive, at best, or dangerous, at worst, there might be some use in modeling the decisions of a clinician who had special interests or training in identifying dishonesty (or dangerousness, or whatever the dimension of interest were) to make consistent decisions for inventories administered to groups. This would also allow traditional tests of adverse impact, as a check on the fairness of such a procedure.

It was also concluded, in light of the large number of cues successfully used in this research, that modeling has great potential for studies of traditional topics in industrial and organizational psychology. Leadership, organizational decision-making, satisfaction, communication, compensation and benefits, work and leisure -- virtually any research which can benefit from insight into intuitive processes would benefit from modeling the decisions of subjects, particularly if that research has depended traditionally on subjective reports of choice or preference (e.g., performance appraisal). Also, a very inexpensive, quick procedure for combining multiple criteria into a composite was presented. This procedure could be used with any number of people to find convergence of opinion. Such convergence might be of benefit when analyzing an organization's goals, training needs, or problems.

Items presented late in the BSI were frequently significant in models of raters. This could mean that the raters did not make decisions early in the process (cf. Sydiaha, 1959; Webster, 1964), that raters skipped around the BSI to get to the information they needed, or the significant cues were not part of the decision process but correlated well with the ratings. The second is the most appealing conclusion, and points to a real advantage to using written records for selection: They can be used as required, and can be referred to after the applicant is gone (if he or she were present at all).

Discussing specific information that predicted raters' decisions would have breached the agreement between the corporation and the author by revealing potential predictors, so specific cues or kinds of cues were not discussed. However, the most often included cue was one the coder was told to enter as negative information contained in an open-ended question. This could lead to the same conclusion reached by early researchers: that the interview is (in part, at least) a search for negative information. The item also appeared later in the BSI, so it also supports the conclusion that raters did not make their decision with information presented early in the BSI.

In some of the models, coded variables regarding the appearance of the completed BSI (and other attributes not considered content) were significant. This may have been the written equivalent to nonverbal cues. Using a written instrument allowed these cues to be coded, and they became part of the predictor battery instead of introducing bias: a distinct advantage of the BSI.

The results of this research also supported Hunter and Hunter's (1984) finding that biodata inventories account for about seven times as much variance in criteria as do employment interviews. This research was a good test of that conclusion, because the data were identical for each method (although tempered somewhat by the arguments and conclusions about criteria presented above).

Wagner's ultimate question of whether subjective integration of much information is better than statistical integration of less information must be answered in the negative, after reviewing this study and those that have preceded it.

Finally, the BSI (especially if scored with decision models) might not require revalidation as frequently as recommended by some researchers (e.g., Dunnette, 1961; Hunter & Hunter, 1984) because it was constructed more

purposefully and scored less serendipitously than those traditionally reported, and it contains open-ended items which might exhibit more stable characteristics than do more constrained items.

Suggestions for Future Research

A study might be conducted with two different written interview forms like the BSI, completed by a single sample of applicants and scored by a single group of raters on identical dimension ratings in random order. If both instruments resulted in similar dimension ratings within applicants and raters, then further conclusions could be made concerning what interviewers measure in the written interview. Similarly, if applicants completed the BSI, then were interviewed using a structured interview, interviewers' outcomes could be modeled from the BSI cues, principal components or factors of the BSI, and subjectively defined BSI dimensions. This would lend valuable insight into the interview process as defined by various statistical and clinical indicators.

The same data used in this study could also be used to study similarity effects, by having the two raters complete the BSI. Their bootstrapped dimension ratings could be used as covariates in analyzing the data. It would be interesting to see if the raters' global variables were related to their own projected attitudes and experiences.

Raters could be asked to rate randomly selected applicants on global variables and others on dimensions only. This would allow a test of the hypothesis that making many component ratings change raters' decisions about applicants (Belec & Rowe, 1983).

Finally, a study of multiple criteria using behavioral, psychological, financial, and personal ratings of the applicants used in this study would allow

researchers to specify which criteria raters were rating, if such criteria exist. Then the validity and utility of various selection and performance appraisal methods could be better measured.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aamodt, M. G., & Pierce, W. L. (1987). Comparison of the rare response and vertical percent methods for scoring the biographical information blank. *Educational and Psychological Measurement, 47*, 505-511.

Anastasi, A., & Schaefer, C. E. (1969). Biographical correlates of artistic and literary creativity in adolescent girls. *Journal of Applied Psychology, 53*, 267-273.

Anderson, C. W. (1960). The relation between speaking times and decision in the employment interview. *Journal of Applied Psychology, 44*, 267-268.

Arnold, J. D., Rauschenberger, J. M., Soubel, W. G., & Guion, R. M. (1982). Validation and utility of a strength test for selecting steelworkers. *Journal of Applied Psychology, 67*, 588-604.

Arvey, R. D. (1979). Unfair discrimination in the employment interview: Legal and psychological aspects. *Psychological Bulletin, 86*, 736-765.

Arvey, R. D., & Campion, J. E. (1982). The employment interview: A summary and review of recent research. *Personnel Psychology, 35*, 281-322.

Arvey, R. D., Miller, H. E., Gould, R., & Burch, P. (1987). Interview validity for selecting sales clerks. *Personnel Psychology, 40*, 1-12.

Asher, J. J. (1972). The biographical item: Can it be improved? *Personnel Psychology, 27*, 519-533.

Ashton, R. H. (1974). Cue utilization and expert judgments: A comparison of independent auditors with other judges. *Journal of Applied Psychology, 59*, 437-444.

Ayabe, C. R. (1985). Multicrossvalidation and the jackknife in the estimation of shrinkage of the multiple coefficient of correlation. *Educational and Psychological Measurement, 45*, 445-451.

Baron, R. A. (1983). "Sweet smell of success"? The impact of pleasant artificial scents on evaluations of job applicants. *Journal of Applied Psychology, 68*, 709-713.

Baron, R. A. (1986). Self-presentation in job interviews: When there can be "too much of a good thing". *Journal of Applied Social Psychology, 16*, 16-28.

Beehr, T. A., & Gilmore, D. C. (1982). Applicant attractiveness as a perceived job-relevant variable in selection of management trainees. *Academy of Management Journal, 25*, 607-617.

Belec, B. E., & Rowe, P. M. (1983). Temporal placement of information, expectancy, causal attributions, and overall final judgments in employment decision making. *Canadian Journal of Behavioral Science, 15*, 106-120.

Bennett, E. M., Allport, R. A., & Goldstein, A. C. (1954). Communication through limited-response questioning. *Public Opinion Quarterly, 18*, 303-308.

Bonneau, L. R. (1957). An interview for selecting teachers. *Dissertation Abstracts, 17*, 537-538.

Brogden, H. E., & Taylor, E. K. (1950). The theory and classification of criterion bias. *Educational and Psychological Measurement, 10*, 159-186.

Brookhouse, K. J., Guion, R. M., & Doherty, M. E. (1986). Social desirability response bias as one source of the discrepancy between subjective weights and regression weights. *Organizational Behavior and Human Decision Processes, 37*, 316-328.

Brunswick, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*, 193-217.

Carlson, R. E. (1967). Selection interview decisions: The effect of interviewer experience relative quota situation, and applicant sample on interview decisions. *Personnel Psychology, 20*, 269-280.

Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. *Journal of Applied Psychology, 60*, 767-769.

Cascio, W. F. (1982). *Applied psychology in personnel management* (2nd ed.). Reston, VA: Reston.

Cascio, W. F., & Ramos, R. A. (1986). Development and application of a new method for assessing job performance in behavioral/economic terms. *Journal of Applied Psychology, 71*, 20-28.

Cash, T. F., Gillen, B., & Burns, D. S. (1977). Sexism and "beautyism" in personnel consultant decision making. *Journal of Applied Psychology, 62*, 301-307.

Cicchetti, D. V., Showalter, D., & Tryer, P. J. (1985). The effect of number of rating scale categories on levels of interrater reliability: A monte carlo investigation. *Applied Psychological Measurement, 9*, 31-36.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cook, R. L., & Stewart, T. R. (1975). A comparison of seven methods for obtaining subjective descriptions of judgmental policy. *Organizational Behavior and Human Performance, 13*, 31-45.

Crandall, R. (1976). Validation of self-report measures using ratings by others. *Sociological Methods and Research, 4*, 380-400.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley.

Dalessio, A., & Imada, A. S. (1984). Relationships between interview selection decisions and perceptions of applicant similarity to an ideal employee and self: A field study. *Human Relations, 37*, 67-80.

Dawes, R. M. (1971). A case study of graduate admissions: Application of three principles of human decision making. *American Psychologist, 26*, 180-188.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.

Dawes, R. M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Denton, J. C. (1964). The validation of interview-type data. *Personnel Psychology, 17*, 281-287.

Dicken, C. F., & Black, J. D. (1965). Predictive validity of psychometric evaluations of supervisors. *Journal of Applied Psychology, 49*, 34-47.

Dipboye, R. L., Fontenelle, G. A., & Garner, K. (1984). Effects of previewing the application on interview process and outcomes. *Journal of Applied Psychology, 69*, 118-128.

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

Dipboye, R. L., Stramler, C. S., & Fontenelle, G. A. (1984). The effects of the application on recall of information from the interview. *Academy of Management Journal, 27*, 561-575.

Dougherty, T. W., Ebert, R. J., & Callender, J. C. (1986). Policy capturing in the employment interview. *Journal of Applied Psychology, 71*, 9-15.

Dudycha, L. W., & Naylor, J. C. (1966). Characteristics of the human inference process in complex choice behavior situations. *Organizational Behavior and Human Performance, 1*, 110-128.

Dunnette, M. D. (1961). Personnel management. *Annual Review of Psychology, 13*, 285-314.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407-424.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans* (CBMS-NSF 38). Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82,* 171-185.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37,* 36-48.

Einhorn, H. J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin, 73,* 221-230.

Einhorn, H. J. (1971). Use of nonlinear, noncompensatory models as a function of task and amount of information. *Organizational Behavior and Human Performance, 6,* 1-27.

Einhorn, H. J. (1972). Expert measurement and mechanical combination. *Organizational Behavior and Human Performance, 7,* 86-106.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision making. *Organizational Behavior and Human Performance, 13,* 171-192.

England, G. W. (1961). *Development and use of weighted application blanks.* Dubuque, IA: W. C. Brown.

Fear, R. A. (1984). *The evaluation interview* (3rd ed.). New York: McGraw-Hill.

Fear, R. A., & Ross, J. F. (1983). *Jobs, dollars, and EEO.* New York: McGraw-Hill.

Forsythe, S., Drake, M. F., & Cox, C. E. (1985). Influence of applicant's dress on interviewer's selection decisions. *Journal of Applied Psychology, 70,* 374-378.

Frank, B. A. (1980). A comparison of an actuarial and a linear model for predicting organizational behavior. *Applied Psychological Measurement, 4,* 171-181.

Ghiselli, E. E. (1966). The validity of a personnel interview. *Personnel Psychology, 19,* 389-394.

Gibson, J. J. (1941). A critical review of the concept of set in contemporary experimental psychology. *Psychological Bulletin, 38,* 781-817.

Gifford, R., Ng, C. F., & Wilkinson, M. (1985). Nonverbal cues in the employment interview: Links between applicant qualities and interviewer judgments. *Journal of Applied Psychology, 70,* 729-736.

Giles, W. F., & Feild, H. S. (1982). Accuracy of interviewers' perceptions of the importance of intrinsic and extrinsic job characteristics to male and female applicants. *Academy of Management Journal, 25,* 148-157.

Gilmore, D. C., Beehr, T. A., & Love, K. G. (1986). Effects of applicant sex, applicant physical attractiveness, type of rater and type of job on interview decisions. *Journal of Occupational Psychology, 59,* 103-109.

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23*, 483-496.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence, for a method of improving on clinical inferences. *Psychological Bulletin, 73*, 422-432.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs, 72*(Pt. 2)

Goldberg, L. R. (1976). Man versus model of man: Just how conflicting is that evidence? *Organizational Behavior and Human Performance, 16*, 13-22.

Goldstein, I. L. (1971). The application blank: How honest are the responses? *Journal of Applied Psychology, 55*, 491-492.

Gorman, C. D., Clover, W. H., & Doherty, M. E. (1978). Can we learn anything about interviewing real people from "interviews" of paper people? Two studies of the external validity of a paradigm. *Organizational Behavior and Human Performance, 22*, 165-192.

Grant, D. L., & Bray, D. W. (1969). Contribution of the interview to assessment of management potential. *Journal of Applied Psychology, 53*, 24-34.

Hakel, M. D., Dobmeyer, T. W., & Dunnette, M. D. (1970). Relative importance of three content dimensions in overall suitability ratings of job applicants' resumes. *Journal of Applied Psychology, 54*, 65-71.

Half, R. (1985). *Robert Half on hiring*. New York: Crown.

Hamilton, J. W., & Dickinson, T. L. (1987). Comparison of several procedures for generating J-coefficients. *Journal of Applied Psychology, 72*, 49-54.

Hammond, K. R., Hursch, C. J., & Todd, F. J. (1964). Analyzing the components of clinical inference. *Psychological Review, 71*, 438-456.

Hays, W. L. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart, and Winston.

Heilman, M. E. (1980). The impact of situational factors on personnel decisions concerning women: Varying the sex composition of the applicant pool. *Organizational Behavior and Human Performance, 26*, 386-396.

Heilman, M. E., & Saruwatari, L. R. (1979). When beauty is beastly: The effects of appearance and sex on evaluations of job applicants for managerial an nonmanagerial jobs. *Organizational Behavior and Human Performance, 23*, 360-372.

Hobson, C. J., Mendel, R. M., & Gibson, F. W. (1981). Clarifying performance appraisal criteria. *Organizational Behavior and Human Performance, 28*, 164-188.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological Bulletin, 57*, 116-131.

Hogarth, R. M. (1980). *Judgement and choice: The psychology of decision.* Chichester, England: Wiley.

Hornick, C. W., James, L. R., & Jones, A. P. (1977). Empirical item keying versus a rational approach to analyzing a psychological climate questionnaire. *Applied Psychological Measurement, 1*, 489-500.

Hunter, J. E., & Hunter, R. F. (1984). The validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Jackson, D. N., Peacock, A. C., & Holden, R. R. (1982). Professional interviewers' trait inferential structures for diverse occupational groups. *Organizational Behavior and Human Performance, 29*, 1-20.

Jackson, D. N., Peacock, A. C., & Smith, J. P. (1980). Impressions of personality in the employment interview. *Journal of Personality and Social Psychology, 39*, 294-307.

James, S. P., Campbell, I. M., & Lovegrove, S. A. (1984). Personality differentiation in a police-selection interview. *Journal of Applied Psychology, 69*, 129-134.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*, 577-580.

Kahn, R. L., & Cannell, C. F. (1957). *The dynamics of interviewing.* New York: Wiley.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.) (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge: Cambridge University Press.

Kalin, R., & Rayko, D. S. (1978). Discrimination in evaluative judgments against foreign-accented job candidates. *Psychological Reports, 43*, 1203-1209.

Keating, E., Paterson, D. G., & Stone, C. H. (1950). Validity of work histories obtained by interview. *Journal of Applied Psychology, 34*, 6-11.

Kleinmuntz, B. (1968). The processing of clinical information by man and machine. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp.149-186). New York: Wiley.

Klimoski, R. J., & Strickland, W. J. (1977). Assessment centers--valid or merely prescient. *Personnel Psychology, 30*, 353-361.

Landy, F. J., & Trumbo, D. A. (1980). *Psychology of work behavior.* Homewood, IL: Dorsey Press.

Langdale, J. A., & Weitz, J. (1973). Estimating the influence of job information on interviewer agreement. *Journal of Applied Psychology, 57*, 23-27.

Latham, G. P., & Saari, L. M. (1984). Do people do what they say? Further studies on the situational interview. *Journal of Applied Psychology, 69*, 569-573.

Latham, G. P., Saari, L. M., Pursell, E. D., & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422-427.

Libby, R. (1976). Man versus model of man: Some conflicting evidence. *Organizational Behavior and Human Performance, 16*, 1-12.

Lunneborg, C. E. (1985). Estimating the correlation coefficient: The bootstrap approach. *Psychological Bulletin, 98*, 209-215.

Mayfield, E. C. (1964). The selection interview--a re-evaluation of published research. *Personnel Psychology, 17*, 239-260.

McDonald, T., & Hakel, M. D. (1985). Effects of applicant race, sex, suitability, and answers on interviewer's questioning strategy and ratings. *Personnel Psychology, 38*, 321-334.

Meehl, P. E. (1954). *Clinical versus statistical prediction*. Minneapolis: University of Minnesota.

Metzner, H., & Mann, F. (1952). A limited comparison of two methods of data collection: The fixed alternative questionnaire and the open-ended interview. *American Sociological Review, 17*, 486-491.

Mitchell, T. W., & Klimoski, R. J. (1982). Is it rational to be empirical? A test of methods for scoring biographical data. *Journal of Applied Psychology, 67*, 411-418.

Mosel, J. L., & Cozan, L. W. (1952). The accuracy of application blank work histories. *Journal of Applied Psychology, 36*, 365-369.

Mullins, T. W. (1978). *Racial attitudes and the selection interview: A factoral experiment*. Unpublished doctoral dissertation, University of Houston.

Mullins, T. W. (1982). Interviewer decisions as a function of applicant race, applicant quality and interviewer prejudice. *Personnel Psychology, 35*, 163-174.

Neter, J., Wasserman, W., & Kutner, M. H. (1985). *Applied linear statistical models* (2nd ed.). Homewood, Il: Richard D. Irwin.

Norman, K. L. (1976). Weight and value in an information integration model: Subjective rating of job applicants. *Organizational Behavior and Human Performance, 16*, 193-204.

Nunnally, J. C. (1970). *Introduction to psychological measurement*. New York: McGraw-Hill.

O'Brien, R. G., & Kaiser, M. K. (1984). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97*, 316-333.

Orpen, C. (1984). Attitude similarity, attraction, and decision-making in the employment interview. *Journal of Psychology, 117*, 111-120.

Orpen, C. (1985). Patterned behavior description interviews versus unstructured interviews: A comparative validity study. *Journal of Applied Psychology, 70*, 774-776.

Owens, W. A. (1984). *A generalized classification of persons--box within a box*. Athens: University of Georgia, Institute for Behavioral Research.

Owens, W. A., & Schoenfeldt, L. F. (1979). Toward a classification of persons [Monograph]. *Journal of Applied Psychology, 63*, 569-607.

Palacios, M. H., Newberry, L. A., & Bootzin, R. (1966). Predictive validity of the interview. *Journal of Applied Psychology, 50*, 67-72.

Pannone, R. D. (1984). Predicting test performance: A content valid approach to screening applicants. *Personnel Psychology, 37*, 507-514.

Parsons, C. K., & Liden, R. C. (1984). Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology, 69*, 557-568.

Peacock, A. C. (1982). Dynamics of inferential judgments in the employment interview. *Dissertation Abstracts International, 43*, 1650B.

Petrinovich, L. (1979). Probabilistic functionalism: A conception of research method. *American Psychologist, 34*, 373-390.

Phelps, R. H., & Shanteau, J. (1978). Livestock judges: How much information can an expert use? *Organizational Behavior and Human Performance, 21*, 209-219.

Powell, G. N. (1986). Effects of expectancy-confirmation processes and applicants' qualifications on recruiters' evaluations. *Psychological Reports, 58*, 1003-1010.

Rasmussen, K. G. (1984). Nonverbal behavior, verbal behavior, resume credentials, and selection interview outcomes. *Journal of Applied Psychology, 69*, 551-556.

Reid, P. T., Kleiman, L. S., & Travis, C. B. (1986). Attribution and sex differences in the employment interview. *Journal of Social Psychology, 126*, 205-212.

Reilly, R. R., & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62.

Rohrer, Hibler, & Replogle (1965). *Managers for tomorrow*. New York: New American Library.

Roose, J. E., & Doherty, M. E. (1976). Judgment theory applied to the selection of life insurance salesmen. *Organizational Behavior and Human Performance, 16*, 231-249.

Rosen, B., & Jerdee, T. H. (1976). The influence of age stereotypes on managerial decisions. *Journal of Applied Psychology, 61*, 428-432.

Rothstein, M. G. (1984). Predicting job performance criteria from interview-based impressions of personality. *Dissertation Abstracts International, 44*, 3232B.

Rothstein, M., & Jackson, D. N. (1980). Decision-making in the employment interview: An experimental approach. *Journal of Applied Psychology, 65*, 271-283.

Rowe, P. M. (1963). Individual differences in selection decisions. *Journal of Applied Psychology, 47*, 304-307.

Rowe, P. M. (1984). Decision processes in personnel selection. Special issue: Social psychology applied to social issues in Canada. *Canadian Journal of Behavioral Science, 16*, 326-337.

Sackett, P. R. (1982). The interviewer as hypothesis tester: The effects of impression of an applicant on subsequent interviewer behavior. *Personnel Psychology, 35*, 789-304.

Sawyer, J. (1966). Measurement *and* prediction, clinical *and* statistical. *Psychological Bulletin, 3*, 178-200.

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31*, 699-714.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29*, 79-101.

Schmitt, N., Gooding, R. Z., Noe, R. A., & Kirsch, M. (1984). Metaanalyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Shaffer, G. S., Saunders, V., & Owens, W. A. (1986). Additional evidence for the accuracy of biographical data: Long-term retest and observer ratings. *Personnel Psychology, 39*, 791-809.

Sheaffer, R. L., Mendenhall, W., & Ott, L. (1979). *Elementary survey sampling*. North Scituate, MA: Duxbury.

Shepard, R. N. (1964). On subjectively optimum selection among multiattribute alternatives. In M. W. Shelly II & G. L. Bryan (Eds.), *Human judgments and optimality* (pp. 257-281). New York: Wiley.

Siess, T. F., & Jackson, D. N (1970). Vocational interests and personality: An empirical integration. *Journal of Counseling Psychology, 17*, 27-35.

Slovic, P. (1969). Analyzing the expert judge: A descriptive study of a stockbroker's decision processes. *Journal of Applied Psychology, 53*, 255-263.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649-744.

Slovic, P., & Lichtenstein, S. (1973). Comparison of Bayesian and regression approaches to the study of information processing in judgment. In L. Rappaport & D. A. Summers (Eds.), *Human judgment and social interaction* (pp. 16-108). New York: Holt, Rinehart, & Winston.

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology, 28*, 1-39.

Snyder, M., & Swann, W. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology, 36*, 1202-1212.

Springbett, B. M. (1958). Factors affecting the final decision in the employment interview. *Canadian Journal of Psychology, 12*, 13-22.

Sterrett, J. H. (1978). The job interview: Body language and perceptions of potential effectiveness. *Journal of Applied Psychology, 63*, 388-390.

Sydiaha, D. (1959). On the equivalence of clinical and statistical methods. *Journal of Applied Psychology, 43*, 395-401.

Sydiaha, D. (1961). Bales' interaction process analysis of personnel selection interviews. *Journal of Applied Psychology, 45*, 393-401.

Sydiaha, D. (1962). Interviewer consistency in the use of empathic models in personnel selection. *Journal of Applied Psychology, 46*, 344-349.

Telenson, P. A., Alexander, R. A., & Barrett, G. V. (1983). Scoring the biographical information blank: A comparison of three weighting techniques. *Applied Psychological Measurement, 7*, 73-80.

Tinsley, H. E. A., & Weiss, D. J. (1973). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology, 22*, 358-376.

Tucker, D. H., & Rowe, P. M. (1977). Consulting the application form prior to the interview: An essential step in the selection process. *Journal of Applied Psychology, 62*, 283-287.

Tucker, D. H., & Rowe, P. M. (1979). Relationship between expectancy, causal attributions, and final hiring decisions in the employment interview. *Journal of Applied Psychology, 64*, 27-34.

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*, 100-116.

Wagner, R. (1949). The employment interview: A critical summary. *Personnel Psychology, 2*, 17-46.

Wallace, H. A. (1923). What is in the corn judge's mind? *Journal of Agronomy, 15*, 300-304.

Walsh, W. B. (1967). The validity of self-report. *Journal of Counseling Psychology, 14*, 18-23.

Webster, E. C. (1964). *Decision making in the employment interview*. Montreal: Eagle.

Weiss, D. J., & Dawis, R. V. (1960). An objective validation of factual interview data. *Journal of Applied Psychology, 44*, 381-385.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372-376.

Wiener, Y., & Schneiderman, M. L. (1974). Use of job information as a criterion in employment decisions of interviewers. *Journal of Applied Psychology, 59*, 699-704.

Wright, O. R. (1969). Summary of research on the selection interview since 1964. *Personnel Psychology, 22*, 391-413.

Zedeck, S., Tziner, A., & Middlestadt, S. E. (1983). Interviewer validity and reliability: An individual analysis approach. *Personnel Psychology, 36*, 355-370.

Zimmer, I. (1981). A comparison of the prediction accuracy of loan officers and their linear-additive models. *Organizational Behavior and Human Performance, 27*, 69-74.

APPENDIX

Applicant Name _____

Date Completed _____

Employment Decision Form

	Anchor Scale	Intellectual Effectiveness	Emotional Characteristics	Interpersonal Skills	Insight Into Self and Others	Planning and Organizational Skills	Physical Ability	Personal Effectiveness	Overall Person, Job and Organizational Fit
Enter Score	☐	☐	☐	☐	☐	☐	☐	☐	☐
Exceptional	7	7	7	7	7	7	7	7	7
Above Average	6	6	6	6	6	6	6	6	6
	5	5	5	5	5	5	5	5	5
Average	4	4	4	4	4	4	4	4	4
	3	3	3	3	3	3	3	3	3
Below Average	2	2	2	2	2	2	2	2	2
Obviously Deficit	1	1	1	1	1	1	1	1	1
No data/ Can't Tell	0	0	0	0	0	0	0	0	0

☐ **Hiring Decision**

1 — No Hire; 4 — Undecided/ Reserve Judgment; 7 — Definite Hire

☐ **Estimate of Potential For Future Advancement**

1 — Little or No Potential; 3 — Somewhat Unlikely; 4 — Undecided/ Reserve Judgment; 5 — Likely; 7 — High Likely

EMPLOYMENT HISTORY

Please provide a complete job history. Make sure you leave no gaps.

1.1a. Past Full-Time Employment

Company and Address (Please list locations worked in past.)	Type of Business	Estimated Actual Average Hrs./Wk. Worked	Salary Weekly (Approx.)	Dates (Mo.& Year) From	To	Job Duties (see Key below). Circle roles performed in each job and indicate number of people supervised, if applicable.	Reasons for Leaving -- Be specific. List all reasons for leaving.
						(see Key below) Roles performed: INS, C, INT, A, R, N, QC Title: DSNS # ____ ISNS# ____	
						(see Key below) Roles performed: INS, C, INT, A, R, N, QC Title: DSNS # ____ ISNS# ____	
						(see Key below) Roles performed: INS, C, INT, A, R, N, QC Title: DSNS # ____ ISNS# ____	

** KEY: Formal Titles Held: - Titles:

The following roles were performed: = Roles performed:
Instructing = INS; Coordinating = C; Interviewing = INT;
Advising = A; Representing = R; Negotiating = N;
Quality Control = QC
If Direct Supervision Occurred, Indicated Number Supervised = DSNS# ____
If Indirect Supervision Occurred, Indicate Number Supervisod = ISNS# ____

1.1b. Please list part-time employment (include jobs during high school and college).

TYPE OF WORK/JOB TITLE	FIRM'S NAME AND ADDRESS	APPROXIMATE DATES

1.2. Were there any special assignments relating to any of the jobs provided in question #1 that are worth noting? Yes _____ No _____ If yes, please explain _____

1.3. In the past, what type of work have you liked best? _____

1.4. What type of work do you see yourself best fitted for? (Be specific)

1.5. What part of your last job you had did you like least? _____

1.6. What position did you aspire to in your last company of employment and what prevented you from reaching it? _____

EDUCATION

2.1. Formal Educational History -- Starting with high school, provide all formal educational experiences that you took part in, including college, technical schools, graduate school, etc.

School and Location	Area of Study Curriculum Major/Minor - Describe Briefly	Date of Attendance	Indicate Certificate Degree or Diploma	Date Graduated	Annual Grade Pt. Avg. Adapt to 4.0 scale

2.2. List all of the titles of college management and mathematical courses that you have had in the past. Also, list all other courses that you consider relevant in furthering your development in the food industry. After each course, provide the grade that you received.

2.3. Were you an honor student in High School? _____ in College? _____

2.4. Informal Educational History -- list all informal seminars, coursework (e.g., vocational educational experiences, Cornell Courses), conferences that you would consider important in furthering your development in the food industry. Also, provide dates attended and length in days.

2.5. From your past educational experiences --
a) What subjects do you like best? (Please list)

b) What subjects did you find most difficult?

2.6. From the tests that you have taken in the past, what scores have you received?

SAT _____

ACT _____

Graduate Boards _____
(Give name and score)

Intelligence Tests _____
(List name of test and scores)

Other Tests _____
(List names of tests and scores)

2.7. If you attended school past High School, how did you finance it?

_____

_____

_____

_____

_____

_____

_____

_____

_____

2.8. a) What extracurricular activities did you take part in during High School? (Please list)

_____

_____

_____

_____

_____

b) What leadership position did you hold in what clubs during this time?

_____

_____

_____

_____

_____

2.9. a) What extracurricular activities did you take part in during college and/or graduate school? (Please list)

b) What leadership positions did you hold during this time?

HEALTH HISTORY

3.1. How would you rate your state of health?

___ poor ___ fair ___ average ___ good ___ excellent

3.2. Do you exercise regularly? _____ Please describe.

_____

_____

_____

_____

3.3. Do you have any sleeping or eating problems? _____ If yes, please describe.

_____

_____

Are you _____ underweight _____ normal weight _____ overweight?

How many hours sleep each night do you receive? _____

3.4. How would you rate your energy level?

___ poor ___ fair ___ average ___ high ___ extremely high

If energy level is high, please explain how you control it. (Be specific).

_____

_____

3.5. Do you drink alcoholic beverages? _____ If yes, how many drinks per week? _____

3.6. Have you had any familial, personal, or health difficulties related to alcohol use? If yes, please explain.

_____

_____

_____

3.7. Are you on any medication? _____ If yes, please describe.

_____

_____

_____

3.8. a) Have you used illegal drugs in the past? Yes _____ No _____
Or, do you currently use them? Yes _____ No _____ If yes, describe.

_____

_____

_____

b) Have you been convicted of any alcohol or drug related legal offenses in the past? Yes _____ No _____ If yes, describe and provide dates.

_____

_____

_____

3.9. Do you smoke? Yes _____ No _____ If yes, what and how much on a weekly basis? _____

_____

3.10. Do you have problems in any of the sensory areas? Yes ____ No ____

If yes, please describe, along with corrective actions taken.

Hearing	Yes _____	No _____
Vision	Yes _____	No _____
Taste	Yes _____	No _____
Touch	Yes _____	No _____
Body awareness	Yes _____	No _____
Muscle sensing ability	Yes _____	No _____

_____

_____

_____

_____

3.11. How would you rate your overall body coordination?

___ poor ___ fair ___ average ___ good ___ excellent

3.12. How would you rate your physical strength?

___ poor ___ fair ___ average ___ good ___ excellent

3.13. When did you last have a physical examination and for what purpose?

_____

_____

3.14. Have you ever been treated for or received benefits for an accident or illness connected with workman's compensation? Yes _____ No _____
If yes, give nature of injury, date, and complete details including number of days lost from work. _____

_____

_____

_____

3.15. Your answers to the following questions are to help make sure we place you in a job safe to yourself and others and in making such health studies as necessary.

Information you supply will be held in confidence by the Human Resources Department, except as it may be necessary that others be informed in regard to your employment with this company only.

Have you have had or now have any of the conditions listed below:
(Please answer for each condition.)

YES	NO	CONDITION	YES	NO	CONDITION
		Allergies			Hernia or Rupture
		Arthritis or Rheumatism			Kidney Trouble
		Asthma			Loss of Eye or Limbs
		Back Trouble			Mental Disturbance
		Cancer, Cyst, Tumor, or Growth			Nerve Trouble or Neuritis
		Diabetes			Pneumonia
		Epilepsy or Fits			Recent Weight Loss or Gain
		Eye Trouble (Serious)			Rheumatic or Scarlet Fever
		Fainting or Dizzy Spells			Skin Trouble or Rash
		Foot Trouble			Stomach Trouble
		Fractures of Bones			Sugar or Albumin in Urine
		Gall Bladder Trouble			Tuberculosis
		Knee Trouble (Trick)			Varicose Veins
		Hearing Difficulties			Hypertension
		Heart Trouble			Other

If you answered YES to any of the above questions, show item number and explain briefly. (Use reverse if necessary):

_____

_____

_____

_____

_____

LOCATIONAL PREFERENCE

4.1. In what areas would you be willing to relocate? (Please list. See map below for locations.)

PERSONAL GOALS AND PHILOSOPHY

5.1. What is important to you in life? (Please list in order of priority, starting with those of most importance.)

5.2. What are your future goals and objectives? (Be specific and and provide long-term and short-term goals with time frames. Please rank major areas in order of importance and list goals in order of importance.)

Personal:

Vocational:

Educational:

Financial: (Please indicate minimum accepted salary for management trainee position. Please indicate minimum accepted salary for store management position after training; and what you would want to earn five years from now, disregarding inflation.)

5.3. What are your plans for self-development in the next five years? (Be specific and list in order of priority.)

5.4. List the qualities in people that you admire most.

5.5. List the qualities in people that you most dislike.

5.6. a) List the qualities that you would like to see a manager possess. (Underline those qualities that you have already developed.)

b) Describe what you would consider an inadequate boss would be like. (Underline those qualities that you possess.)

5.7. What is your philosophy of management? How did you develop it?

5.8. Why are you applying for a store management position and why at ?

PERSONAL HISTORY

6.1. a) Outside of your educational experiences, what clubs and organizationals have you taken an active part in? (List dates, as well as clubs and organzational names.)

b) Please list the leadership positions that you have held in the above clubs and organziations (include dates and positions that were held).

6.2. List awards that you have received in the past. Also, please list important accomplishments that you have performed.

6.3. What types of hobbies as well as recreational and social activities do you take part in? (Please list.)

_____

_____

_____

_____

_____

_____

6.4. What kinds of materials do you read on a regular basis?

Types of Books: _____

_____

Number of Books per year: _____

Periodicals, magazines, journals, etc., read on a regular basis. (Please list publication names.)

_____

_____

_____

Newspapers. (Please list names and number of days per week read and average number of hours per weeks spent on reading them.)

_____

_____

_____

Other reading materials. (Please list.)

_____

_____

_____

6.5. Number of hours per week spent watching TV (7 days). _____

Number of hours spent watching news. _____

Number of hours spent in listening to radio news. _____

6.6. Whom in the past have you sought guidance from and admire? (Please list their names and their relationship to you. Also, indicate whether you still keep in touch with them on a regular basis.)

6.7. What states or countries have you lived in? (Please list.)

6.8. What states and countries have you traveled to? (Please list.)

6.9. Please list what you estimate your areas of strength to be as a person.

6.10. Please list what you estimate your areas of weakness to be as person.

6.11. Do you ever feel under stress? ___ Yes ___ No How do you deal with it?

6.12. What have you done in the last five years to improve yourself as a person? (Please list.)

Answer each of the following questions below. Circle T for true and F for false.

T F 1. I really get discouraged at times.

T F 2. I consistently feel that life is worthwhile.

T F 3. It seems that in the past, people used to have more fun that they do now.

T F 4. I get tense as I think of all the things lying ahead of me.

T F 5. I frequently have spells of the blues.

T F 6. I wake up fresh and rested most mornings.

T F 7. Overall, the rich man is better off than the poor man.

T F 8. I enjoy people a lot.

T F 9. Supervisors typically expect too much from their supervisees.

T F 10. When playing competitive games with others, winning is more important than enjoying the company of other people.

T F 11. Being successful is a matter of willpower.

T F 12. I enjoy the planning of activities and the making of decisions with regard to how things should get done.

T F 13. People often ask me for advice or for assistance in making decisions.

T F 14. The fear of getting caught is the reason behind people acting honestly.

T F 15. I am usually part of and aware of what is going on in the group I belong to.

T F 16. I prefer to work with others.

T F 17. Most people inwardly dislike putting themselves out to help others.

T F 18. If they could gain by it, most people would be willing to lie.

T F 19. At times, I have ups and downs without apparent reason.

T F 20. I handle changes quite well.

T F 21. It's okay to try to grab all a person can get in this world.

T F 22. I often feel grouchy.

T F 23. Ethical and moral issues need to be taken quite seriously.

T F 24. People will regularly use unfair means to gain profit or to take advantage rather than to lose it.

T F 25. I admire those who can turn a fast buck, even when it is a little shady.

T F 26. This would be a happier world, if people followed moral codes versus trying to stretch them.

T F 27. I do not always tell the truth.

T F 28. When diplomacy and persuasion are needed to get people moving, others usually ask me to do it.

T F 29. I wait until I am sure that what I am saying is correct, before I speak.

T F 30. I communicate better in person than on a written form such as this.

T F 31. I always insist on doing things as correctly as possible.

T F 32. At times, I feel like swearing.

T F 33. Every once in a while, I put off what I should do today.

T F 34. I enjoy knowing some important people because it makes me feel important.

T F 35. I am a bit shy in meeting new acquaintances.

T F 36. I do not like everyone I know.

T F 37. I don't make smart and sarcastic remarks to others even if I believe that they deserve it.

T F 38. I would usually "have it out" with a person who spreads untrue rumors about me.

T F 39. I am generally quite self-confident.

T F 40. I display poise and social presence.

T F 41. I have a knack for fixing things and for keeping them running properly.

T F 42. I have trouble solving math calculations easily and rapidly.

T F 43. I dread going into a room alone when other people are already meeting there.

T F 44. I typically make friends rather easily.

T F 45. I enjoy influencing and controlling people.

T F 46. In the past, I have had quite peculiar and strange experiences.

T F 47. My hands often shake when I am using them.

T F 48. I have much more to worry about than other people do.

T F 49. Achievement is more gratifying than admiration.

T F 50. If a person takes care of himself, he does not need to worry about other people.

T F 51. Years ago, planning seriously for the future made more sense than it does today.

T F 52. Useless thoughts which keep running through my mind interfere with my ability to concentrate.

T F 53. I am regularly punished without cause.

T F 54. I often get angry at people too quickly.

T F 55. At times, I have difficulty finding enough energy to face and deal with my problems.

T F 56. In choosing a firm for which to work, I would prefer one that had a history of gradual success rather than one that presented quick opportunity for advancement.

T F 57. I am considered a very enthusiastic person.

T F 58. I usually accept criticism well.

T F 59. I am very independent, and I am known for following my own lead.

T F 60. Others would say that I am very persistent.

T F 61. I see personal meaning in my life.

T F 62. I have periods when it is hard to stop a mood of self-pity.

T F 63. I have a hard time acting naturally in front of new acquaintances.

T F 64. I have been cheated out of a number of the best years of my life.

T F 65. I sometimes make decisions too quickly.

T F 66. I have a good sense of humor.

T F 67. Others see me as too serious.

T F 68. I feel terribly dejected when people criticize me in a group.

T F 69. My grammar, vocabulary, and overall language usage are better than average.

T F 70. People say that I am very direct in my dealings with them.

T F 71. I am willing to put in long hours on a regular basis.

T F 72. At times, I gossip a bit.

T F 73. I offer others advice when I see that they need it.

T F 74. I prefer routines to flexible work.

T F 75. I am usually neat and well groomed.

T F 76. If given a chance, I will lead others well.

T F 77. I provide a positive first impression to others in that I am outgoing, personable, and friendly.

T F 78. It is important to understand intellectual matters.

T F 79. I have little trouble asserting myself as needed.

T F 80. I am generally quite neat, prompt, reliable and dependable.

T F 81. I would prefer a job that allows me to move around physically a lot rather than to work in the same place every day.

T F 82. I am of above average intelligence.

T F 83. I see myself as quite practical minded, a person with good common sense.

T F 84. I make significant efforts in order to learn and to continue to develop in my vocational field as well as more generally.

T F 85. Others would say that I am very conscientious and enthusiastic.

T F 86. I am a curious and an inquiring person.

T F 87. I often act on the spur of the moment without thinking.

T F 88. People often misunderstand my actions.

T F 89. My reading and writing skills are better than average.

T F 90. My math skills are better than average.

T F 91. I am good with details.

T F 92. Sometimes I get angry.

T F 93. I like to be in control at all times.

T F 94. I am often giving people advice.

T F 95. I am the kind of person who would happily put in extra long days rather than one who is dependable but who typically leaves at the end of the work day.

T F 96. In talking to others, I get my ideas across to them concisely and fluently.

T F 97. I am quite superstitious.

T F 98. I appreciate constructive criticism because it assists me to grow and to develop.

T F 99. I believe that the best results are obtained by leading a group with similar backgrounds rather than those who are more diverse.

T F 100. I am typically the leader in groups I participate in.

Please check the correct answer for each of the following situations. (Questions 101-106)

101. Another Store Manager tells you in a blunt and distasteful fashion to do something in a different way than you intended. He has no direct authority over you. How should you respond?

_____ Follow his directions immediately.

_____ Do it your own way and say nothing.

_____ Tell him to do the job himself.

_____ Confront him with the fact that it is not within his area of responsibility, and that you will do your own work your own way.

102. A newly promoted Area Supervisor would probably best meet his and the company's objectives by:

_____ Promoting quickly those he thinks deserve it.

_____ Talking with his employees and asking their advice as to necessary changes.

_____ Teaching his employees what true efficiency is.

_____ Making new changes gradually while continuing previous policies.

103. If you were hired as a management trainee at and you were doing work in a store, what would you consider the best approach in order to establish good interpersonal relations with your fellow staff members?

_____ Always speak well of them to the Store Manager.

_____ Try not to pay attention to their errors and to not correct them when they were wrong.

_____ Show interest in your work and try to be cooperative whenever possible.

_____ Set yourself up to perform those jobs that you can do better than they can.

104. You have an employee who does a good job, but he complains continually about the stocking that he has to do. His complaining seems to have a bad effect on the other stockers. The most effective approach to use would be to:

_____ Change him to another department and to another boss.

_____ Talk to him about his attitude and try to work things out.

_____ Tell the other stockers to try to overlook him.

_____ Isolate him and his work from the others.

_____ Let him plan out his own work.

105. A Store Manager is transferred to another town. He was a leader in the community, and he belonged to a variety of organizations. He now would be expected to:

_____ Become more interested in recreational activities.

_____ Dislike his new home.

_____ Over time, take part in the life of the community and further develop his interests.

_____ Start a new life and a new set of interests.

106. Another employee is talking to you about his last trip. The conversation is boring. It would be best to:

_____ Continue to listen politely.

_____ Listen but act disinterested.

_____ Tell him in a straightforward manner that the subject does not interest you.

_____ Continue to look at your watch in an impatient manner.

The following questions are optional. Use any resources that are available to you in answering them.

107. According to Time Magazine, who has the best ice cream in the world?

_____

_____

_____

_____

108. What vitamins are contained in squash?

_____

_____

_____

_____

109. How does one make Haggis?

_____

_____

_____

_____

110. What are salad days, and how did the term orginate?

111. What vitamins cancel out each other?

112. How does one predict pork belly futures?

113. What is Cilantro?

114. What is a baby pig called?

115. Who is Albert Broccoli, and what does he have to do with the green vegetable?

116. When is the best time to plant garlic?

117. Who first invented the ice cream sandwich?

118. Where did the tradition for having lamb at Easter dinner originate?

After filling out this application, if you were to meet socially with its authors, what areas of conversation would you propose outside of your work, and this application blank itself, and why? (It is understood that there is no right answer for this question -- but, please answer it anyway).

Please evaluate this application -- what are its strengths and weaknesses?

VITA

Thomas Winslow Mason was born in Oak Park, Illinois on June 15, 1954. He attended Oak Park-River Forest High School until November, 1971. He received the GED in October, 1972 and, in February, 1973, entered Elmhurst College in Elmhurst, Illinois. He majored in biology and psychology, and received the Bachelor of Science degree in May, 1977.

He moved to Knoxville, Tennessee in December of 1976, where he was employed in industrial sales. In October, 1982 he entered the program in Industrial and Organizational Psychology at The University of Tennessee, Knoxville. He received the Doctor of Philosophy degree with a major in Industrial and Organizational Psychology and a minor in Statistics in June, 1988.

The author is a member of Phi Kappa Phi, Alpha Epsilon Delta, American Psychological Association, American Statistical Association; he was included in the 1977 edition of "Who's Who in American Colleges and Universities." He is a licensed Psychological Examiner in Tennessee and is president of a psychological and statistical consulting firm in Knoxville, Tennessee.