1

ARI Research Note 88-02
(Revised June 1987)

**Project A:**
Improving the Selection, Classification and Utilization of Army Enlisted Personnel

# Literature Review: Utility of Temperament, Biodata, and Interest Assessment for Predicting Job Performance

Leaetta M. Hough, Editor
*Personnel Decisions Research Institute*

for

Selection and Classification Technical Area
Lawrence M. Hanser, Chief

Manpower and Personnel Research Laboratory
Newell K. Eaton, Director

ari

U. S. Army
Research Institute for the Behavioral and Social Sciences

January 1988

Approved for public release; distribution unlimited.

# U. S. ARMY RESEARCH INSTITUTE

# FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

**EDGAR M. JOHNSON**
**Technical Director**

**WM. DARRYL HENDERSON**
**COL, IN**
**Commanding**

Research accomplished under contract
to the Department of the Army

Human Resources Research Organization

Technical review by:

**Ilene Gast**
**Deirdre Knapp**

This report, as submitted by the contractor, has been cleared for release to Defense Technical Information Center (DTIC) to comply with regulatory requirements. It has been given no primary distribution other than to DTIC and will be available only through DTIC or other reference services such as the National Technical Information Service (NTIS). The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

A192 109

# REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Personnel Decisions Research Institute
Report Number 116

5. MONITORING ORGANIZATION REPORT NUMBER(S)
ARI Research Note 88-02

6a. NAME OF PERFORMING ORGANIZATION
Human Resources Research
Organization

6b. OFFICE SYMBOL *(If applicable)*
HumRRO

7a. NAME OF MONITORING ORGANIZATION
U.S. Army Research Institute for
the Behavioral and Social Sciences

6c. ADDRESS *(City, State, and ZIP Code)*
1100 South Washington Street
Alexandria, VA 22314

7b. ADDRESS *(City, State, and ZIP Code)*
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
- -

8b. OFFICE SYMBOL *(If applicable)*
- -

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
MDA903-82-C-0531

8c. ADDRESS *(City, State, and ZIP Code)*
- -

10. SOURCE OF FUNDING NUMBERS

| PROGRAM ELEMENT NO. | PROJECT NO. | TASK NO. | WORK UNIT ACCESSION NO. |
|---|---|---|---|
| 6.37.31.A | 202637<br>31A792 | 2.3.2 | 2.3.2.C1 |

11. TITLE *(Include Security Classification)*
Literature Review: Utility of Temperament, Biodata, and Interest Assessment for Predicting Job Performance

12. PERSONAL AUTHOR(S)
L.M. Hough, editor (Personnel Decisions Research Institute)

13a. TYPE OF REPORT
Interim Report

13b. TIME COVERED
FROM 10/82 TO 9/85

14. DATE OF REPORT *(Year, Month, Day)*
January 1988

15. PAGE COUNT
226

16. SUPPLEMENTARY NOTATION Personnel Decisions Research Institute was subcontractor for the principal contractor, Human Resources Research Organization. This research note (OVER)

17. COSATI CODES

| FIELD | GROUP | SUB-GROUP |
|---|---|---|
| | | |
| | | |

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
First-Tour Evaluation Literature Review Predictors
Interest Assessment Personnel Classification Biodata
~~Job Performance~~ Personnel Selection (OVER)

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The research described was performed under Project A, the U.S. Army's large-scale, multiyear manpower and personnel effort for improving the selection, classification, and utilization of Army enlisted personnel. This report is one of three derived from an extensive literature review aimed at identifying many types of constructs that might be used to enhance the accuracy of the present Army screening system (the other two reports deal, respectively, with cognitive abilities and psychomotor abilities). The present report is divided into three sections, each dealing with the utility of one type of information for predicting job performance. The section on temperament discusses traits as the basis for temperament assessment, several methods of scale construction, comparison of psychometric properties, a proposed taxonomy of temperament scales, criterion-related validity, and various moderator variables. The section on biographical data discusses measurement methods and concerns, structure and conceptual issues, and validity research. The section on
(OVER)

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Lawrence M. Hanser

22b. TELEPHONE *(Include Area Code)*
202/ 274-8275

22c. OFFICE SYMBOL

ARI RESEARCH NOTE 88-02

16. Supplementary Notation (continued)

is part of "Project A: Improving the Selection, Classification, and Utilization of Army Enlisted Personnel" (Human Resources Research Organization, American Institutes for Research, Personnel Decisions Research Institute, U.S. Army Research Institute).

18. Subject Terms (continued)

Project A
Temprament

19. Abstract (continued)

interest assessment discusses various methods of measuring interests, models and theories, and validity research. Keywords: Personnel selection.

# PREFACE

This Research Note is one of three that presents the results of a literature review conducted as part of Project A, a large-scale, multiyear research program intended to improve the selection and classification system for initial assignment of persons to U.S. Army Military Occupational Specialties. The research is sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences.

The three Research Notes each cover a separate domain of measures of human abilities, interests, and other attributes. Their titles are:

- *Literature Review: Cognitive Abilities--Theory, History, and Validity* by Jody L. Toquam, VyVy A. Corpe, Marvin D. Dunnette, and Margaret A. Keyes.
- *Literature Review: Validity and Potential Usefulness of Psychomotor Ability Tests for Personnel Selection and Classification* by Jeffrey J. McHenry and Sharon R. Rose.
- *Literature Review: Utility of Temperament, Biodata, and Interest Assessment for Predicting Job Performance* by Leaetta M. Hough, Editor.

The findings presented in these documents were used in the development of a battery of new tests and inventories for use in Project A. The focus of that development effort was to identify abilities and other human attributes that seemed "best bets" for predicting soldiers' job performance, and then to develop new measures for those attributes. These Research Notes, however, have usefulness beyond that particular applied problem. Many issues pertinent to the measurement and use of human abilities are described and discussed in each of these compilations.

The Research Notes describe the results and findings of the literature review, but do not describe the literature search process itself. Therefore, we provide a description of that process in the Preface of each volume.

The literature search was conducted by three research teams from the Personnel Decisions Research Institute. Each team was responsible for one of the three fairly broadly defined areas of human abilities or characteristics that are reported in the Research Notes: cognitive abilities; psychomotor abilities; and non-cognitive characteristics such as vocational interests, biographical data, and measures of temperament. While these domains were convenient for purposes of organizing and conducting literature search activities, they were not used as (nor intended to be) a final taxonomy of possible predictor measures.

The major part of the literature search was conducted in late 1982 and early 1983. Within each of the three areas, the teams carried out essentially the same steps:

1. Compile an exhaustive list of reports, articles, books, or other sources that were possibly relevant to Project A.

2. Review each item and determine its relevancy for the project's general purposes by examining the title and abstract (or other brief review).

3. Obtain the sources identified in the second step as being relevant.

4. For relevant materials, conduct a thorough review and transfer applicable information onto special review forms developed for the project.

In the first step, several activities were designed to insure that the list would be as comprehensive as possible. Several computerized searches of relevant data bases were performed. Across all three ability areas, more than 10,000 potential sources were identified via the computer searches. (Of course, many of these sources were identified as relevant in more than one area, and were thus counted more than once.)

In addition to the computerized searches, reference lists were obtained from recognized experts in each area, emphasizing the most recent research in the field. Several annotated bibliographies were obtained from military research laboratories. Finally, the last several years' editions of research journals that are frequently used in each ability area were scanned, as were more general sources such as textbooks, handbooks, and appropriate chapters in the *Annual Review of Psychology* (which reviews the most recent research in a number of conceptually distinct areas of psychology).

The majority of the items identified in the first step proved not relevant to the applied purpose--that is, the identification and development of promising measures for personnel selection in the U.S. Army. These nonrelevant sources were weeded out in Step 2.

The relevant sources were obtained and reviewed, and team members completed two forms for each source: an Article Review form and a Predictor Review form (several of the latter could be prepared for each source). These forms were designed to capture, in a standard format, the essential information about the reviewed sources, which varied considerably in their organization and reporting styles.

The Article Review form contained seven sections: citation, abstract, list of predictors (keyed to the Predictor Review forms), description of criterion measures, description of sample(s), description of methodology, other results, and reviewer's comments. The Predictor Review form also contained seven sections: description of predictor, reliability, norms/descriptive statistics, correlations with other predictors, correlations with criteria, adverse impact/differential validity/test fairness, and reviewer's recommendations (about the usefulness of the predictor). Each predictor was tentatively classified into an initial working taxonomy of predictor constructs.

The Review forms and the actual sources that had been located were used in two primary ways for Project A purposes. First, three working documents were written, one for each of the three areas. These working documents later evolved into the three Research Notes named above. These documents identified and summarized the literature with regard to issues important to the research being conducted, the most appropriate organization or taxonomy of the constructs in each area, and the validities of the various measures for different types of job performance criteria. Second, the predictors identified in the review were subjected to further, structured scrutiny in order to select tests and inventories for use in later activities of Project A.

As a set, the three Research Notes should provide a valuable resource for scientists, researchers, and personnel practititoners interested in the measurement of individual differences in humans for various applied purposes, but especially for selection and classification.

## CONTENTS

Page

Page

## LIST OF APPENDICES

## LIST OF TABLES: SECTION 1

## LIST OF TABLES: SECTION 1 (CONTINUED)

## LIST OF TABLES: SECTION 2

## LIST OF TABLES: SECTION 3

## LIST OF FIGURES: SECTION 2

## LIST OF FIGURES: SECTION 3

John D. Kamp and Leatta M. Hough

# SECTION 1

# UTILITY OF TEMPERAMENT FOR PREDICTING JOB PERFORMANCE

## Overview

In this section the literature on the use of self-report temperament scales for applied prediction problems is reviewed. Almost from its inception, the field of temperament, or personality, assessment has been characterized by controversy. These controversies, all of which have implications for applied temperament assessment, are reflected in the five major topic areas covered in this section:

(1) The very foundation of traditional self-report temperament assessment, the concept of the trait, has been challenged. The discussion of this topic shows how this challenge has served, not to repudiate the assessment of temperament traits, but to make it a stronger enterprise.

(2) Even among those adopting trait views of temperament, vastly different methods of constructing self-report temperament scales have been advocated and employed. These various methods are discussed and their psychometric properties are evaluated. They are then integrated into a construct-guided approach to temperament assessment that is recommended for applied prediction problems.

(3) In order to implement this construct-guided approach, an organization of the apparently chaotic domain of proposed temperament constructs and measures is desirable. To meet this need, a taxonomy of temperament constructs and scales is proposed and evaluated empirically.

(4) Views on the usefulness of temperament assessment for applied prediction problems range from extreme advocacy to extreme skepticism. The criterion-related validity evidence for self-report temperament scales is reviewed in the framework of the proposed predictor taxonomy. In the course of this review, predictor-criterion relationships of greater and lesser promise are identified, and suggestions are made for maximizing the criterion-related validity of temperament scales.

(5) Finally, the potential effects of possible moderator variables, such as faking, sex, and ethnicity, on criterion-related validity have been of concern in applied temperament assessment. The review of this literature shows that definite answers to most of these questions about moderator variables are not available, and directions for future research and practice are suggested.

## Traits as the Basis for Temperament Assessment

The self-report assessment of individual differences in temperament has traditionally been based on the concept of the trait. The validity of this approach has, however, been challenged. In this subsection, traditional trait approaches and the alternatives that have been proposed are described and considered in the context of their implications for temperament assessment in applied psychology.

## Trait Conceptions

Traditional trait conceptions are best exemplified in the writings of Gordon Allport and Henry Murray. Allport (1937) proposed a view of traits as tendencies to respond in certain consistent ways within classes of functionally equivalent situations. Murray (1938) centered his motivational theory of temperament around the trait concept of the "need," which he referred to as "an organic potentiality or readiness to respond in a certain way under given conditions" (p. 61). These two definitions make explicit one of the two basic assumptions underlying trait conceptions, that traits are expected to express themselves in behavioral consistencies across a range of trait-relevant situations. The second fundamental assumption is that traits display temporal stability; that is, the behavioral consistencies that constitute traits are relatively enduring over time. This assumption is reflected in Guilford's (1959) definition of a trait as "any distinguishable, relatively enduring way in which one individual differs from others" (p. 6). As will be seen later, there are vast differences among the various trait approaches to temperament assessment on such fundamental issues as what the primary traits of temperament are and how best to construct scales to measure them. Nevertheless, all trait approaches ultimately rest on the assumptions of the cross-situational consistency and temporal stability of behavior.

## The "Situationist" Challenge

In 1968, Walter Mischel published a highly influential book that spurred a period of intensive examination and heated debate concerning the validity of temperament trait conceptions. Mischel asserted, on the basis of a review of the literature, that correlations between cross-situational indicators of the same trait rarely exceed .30, except when such indicators are based on self-report. Mischel attributed the apparent evidence of cross-situational behavioral consistency found in self-report (as indexed, for example, by high intercorrelations among the items of a self-report trait scale) to the tendency of people to impose an illusion of trait-like consistency on their perceptions of their own behavior, when behavioral variability across situations is actually the rule. In other words, in their implicit personality theorizing, people tend to act, mistakenly, as trait psychologists.

Although Mischel did not question the temporal stability of behavior (cf. Mischel & Peake, 1982), his pessimistic view of the extent of cross-situational behavioral consistency led him and others to conclude that traits of temperament are merely convenient fictions, with little or no basis in reality. Mischel argued for a social learning perspective in which behavior is seen as explainable primarily by the situational context in which it occurs rather than by traits. Mischel's 1968 book thus initiated the "situationist" challenge to trait approaches to the assessment of temperament.

As pointed out by Bowers (1973), "situationism" represents an amalgamation of many somewhat different positions, all essentially behavioristic in emphasis, rather than a single, monolithic strategy. Situationism has most commonly been interpreted as the view that the amount of variance in behavior explained by differences in situations is far greater than the amount explained by individual differences in temperament (Bowers). Situational determinants of behavior are investigated through experimental methods, while individual differences determinants are investigated using correlational methods, and the debate over which of these two approaches is best suited to the advancement of psychology is an old one (cf. Cronbach, 1957).

However, an emphasis on the importance of situational influences does not invalidate an individual differences approach based on traits. This is because there are in fact individual differences in behavior in nearly all situations, and, as Golding (1975) has pointed out, individual differences in behavior (e.g., talkativeness) could theoretically be perfectly consistent across situations differing vastly in the overall levels of behavior they evoke (e.g., funerals vs. parties). It is the task of temperament assessment to predict such individual differences in behavior within situations, regardless of how much of the overall cross-situational behavioral variance they ultimately account for.

## The "Interactionist" Alternative

An alternate approach, called "interactionism," has been proposed in an effort to achieve a reapproachment of those views of temperament stressing situational effects and those stressing traits. Interactionism recognizes the importance of individual differences in behavior, but also argues that these individual differences are highly situation-specific rather than displaying trait-like consistency across a wide range of situations. Proponents of this approach assert that the prediction of behavior requires taking into account both the person and the situation. This seemingly revolutionary view has been heartily endorsed (e.g., Bowers, 1973; Pervin, 1978), so much so that one reviewer concluded that "if interactionism is not the Zeitgeist of today's personality psychology, it will probably be that of tomorrow's" (Ekehammar, 1974, p. 1045).

As with situationism, the exact meaning of interactionism has been interpreted in many different ways (Olweus, 1977). The interpretation that has been most extensively subjected to empirical analysis pertains to the size of the statistical interaction term in analyses of variance incorporating main effects for both persons and situations, as exemplified by studies done by Endler and Hunt (1966, 1969) on anxiousness. In these studies, subjects rated their typical responses across a variety of response modes to each of a number of different potentially anxiety-provoking situations. When these ratings were analyzed, the interaction between persons and situations accounted for significantly more of the overall variance than did the main effects for either persons or situations. The combined results of 11 such components of variance studies reviewed by Bowers (1973) lead to essentially the same conclusion. The application of the interactionist view to temperament assessment is illustrated by the S-R Inventory of General Trait Anxiousness (Endler & Okada, 1975), which yields separate trait anxiety scores for each of four general classes of situations, such as interacting with other people and encountering physical danger.

As revolutionary as interactionism may have seemed, the necessity of taking both personal and situational characteristics into account in predicting behavior is actually inherent in the traditional trait conceptions that interactionism sought to supplant. As Tellegen (1981) has reminded us, classical conceptions such as those of Allport and Murray define traits as dispositions to exhibit characteristic behaviors in functionally equivalent situations. Although there is thus no paradigmatic difference between interactionism and traditional trait conceptions, the practical difference is that interactionists like Endler have viewed functionally equivalent situations as much narrower (e.g., interacting with people) than have traditional conceptions, which view traits as operative across a wider range of situations (e.g., all ego-threatening situations).

This basic difference reduces to a question of the extent to which behavioral consistencies are relatively broad or specific to more circumscribed classes of situations, which in the limiting case would require a different trait for every single situation that could be encountered. Interactionists have interpreted the large person by situation interaction effects emerging from the components of variance studies as empirical evidence for the situation-specificity of individual differences. However, Golding (1975) has cogently argued that such results are not decisive unless these interactions are also shown to be meaningful and replicable rather than idiosyncratic. The practical usefulness of interactionism can perhaps best be tested through future research comparing the predictive validity of traditional trait measures with that of measures of more situation-specific dispositions.

## The Defense of Trait Measurement

In the face of challenges from the situationist and interactionist positions, some have defended the construct validity of temperament trait measures. One important aspect of assessing construct validity is through the correlations of a test with criterion variables related to the construct that the test purports to measure (Cronbach & Meehl, 1955). As pointed out by Mischel (1968), the temperament assessment literature is replete with low validity coefficients. However, some defenders of traditional trait assessment have pointed to conceptual and methodological inadequacies as a cause of the many instances of apparent lack of validity of trait measures (e.g., Block, 1977; Hogan, Desoto, & Solano, 1977). Block's colorful statement of this position can scarcely be improved upon: "It may well be that the current dismal assessment of the personality literature depends too heavily on the poor 'batting average' our sloppy empiricism has attained. Home runs have been averaged with strike-outs, and clearly there have been many of the latter. But some people know how to play ball and others do not. What if the home runs are hit by competent, resourceful athletes while the strike-outs come from the blind and the infirm" (p. 41).

In fact, conceptually and methodologically sound studies do often support the construct validity of self-report trait measures, with the .30 validity "ceiling" proposed by Mischel (1968) regularly being exceeded (e.g., Block, 1977; Hogan et al., 1977; Jackson & Paunonen, 1980). In addition, empirical studies published since the reviews just cited have shown substantial validity for self-report trait measures as predictors of trait ratings by others (Cheek, 1982; McCrae, 1982) and experimental laboratory behavior (Monson, Hesley, & Chernick, 1982).

One particular methodological factor that has been shown to contribute to the appearance of poor validity for self-report temperament measures is criterion unreliability. It has been shown that the low correlations commonly found between self-report trait measures and single-item criteria derived from a single behavioral act or rater increase markedly when these criteria are combined into more reliable aggregates (Cheek, 1982; Epstein, 1979, 1980; Rushton, Brainerd, & Pressley, 1983). Criterion aggregation is based on the same principle as that used to form reliable self-report trait scales from composites of individual items whose intercorrelations are relatively low. Overall, the spirit of the defense of traditional temperament trait measurement is captured well by Epstein's (1977) defiant proclamation: "Traits are alive and well."

## Temperament Assessment in Applied Psychology

What implications do the various positions and research findings that have been described have for the use of temperament assessment for applied prediction problems? The situationist contention that more of the overall variance in behavior is due to situational differences than to differences in personal characteristics has little or no bearing on the question of how to predict individual differences in behavior within situations. The interactionists have offered an answer to this question: Measure trait conceptions that are more situationally circumscribed than those embodied in traditional trait measures. This prescription has a certain common-sense appeal for many applied problems. For example, it seems plausible that a measure of fear of heights could prove more valid than a general harm-avoidance scale in forecasting the success of skyscraper window-washers. However, the comparative validity of situation-specific versus broader trait measures has yet to be adequately tested. A further problem with this approach is the necessity of measuring an ever greater number of traits as the domain of situations is divided more and more finely. The cost of this measure must be weighed against the relative economy of assessment afforded by broader scales.

The situationist and interactionist challenges have been useful in motivating a careful scrutiny of trait approaches to temperament assessment. Defenders of trait approaches have shown that research that is methodologically sound (e.g., criteria are reliable) and conceptually sound (i.e., predictors and criteria are chosen on the basis of plausible hypotheses) does support the validity of trait measures. This evidence has come largely from studies using basic research criteria. However, attention to these same factors should also maximize the predictive validity of temperament trait measures for applied psychology criteria.

From a broader perspective, demonstrations of validity for traditional self-report trait measures in predicting applied psychology criteria perhaps do as much as anything else to support the viability of the trait approach to temperament assessment. Evidence of such validity, particularly in the prediction of criteria reflecting psychological adjustment, is presented later in this section.

## Methods of Scale Construction

Many different methods of constructing self-report scales to measure temperament traits have been advocated and employed. These various methods can be grouped into three major categories: purely rational, external, and internal. All three methods begin with the assembly of an initial item pool based on rational hypotheses about the domain of constructs to be measured. The strategies then diverge.

In the purely rational approach, the grouping of items into scales is based solely on the rational judgment of the test constructor, while in the other two approaches items are assigned to scales on the basis of empirical evidence. In the external scale development approach, items are selected based on evidence of their nontest correlates (e.g., the effectiveness of items in differentiating criterion groups like delinquents and nondelinquents). Scale construction using the internal approach is guided by evidence of the intercorrelations among the items in the pool (e.g., in the use of orthogonal factor analysis to construct scales with optimum convergent and discriminant properties).

The rationales and specific measures associated with the purely rational, external, and internal methods are described below in more detail. The psychometric properties of the different kinds of scales are then compared in the following subsection.

## The Purely Rational Method

Focus and Rationale. The focus of the purely rational approach to scale construction is on maximizing content validity, the extent to which a measure adequately samples the content domain embraced by a construct. In the content validity approach of purely rational scale construction, rational judgment is necessary both to specify the content domain and to assess the adequacy with which a measure samples that domain.

Jackson (1971) has provided the clearest statement of the premise underlying this approach. Jackson asserted in his first "principle" of valid temperament assessment that scale items must be derived from an explicit definition of the trait to be measured. While recognizing the probabilistic relationship between questionnaire responses and the underlying trait, Jackson argued that "issues, questions, or situations directly bearing on the characteristics of interest will show the highest probability of reflecting them" (p. 232). According to Jackson, the purely rational approach will also result in scales with maximum criterion-related validity. Of course, correlations with other variables are necessary to establish empirically the criterion-related and construct validity of purely rational scales.

Measures. Many of the earliest temperament scales were constructed on a purely rational basis (Meehl, 1945), and throughout the history of temperament assessment, a vast number of unpublished measures have been devised for research purposes using purely rational construction. Among published multiscale temperament inventories that are currently in widespread use, however, only the *Edwards Personal Preference Schedule* (EPPS; Edwards, 1959) was constructed through purely rational methods, with scales targeted toward 15 of the psychological needs proposed by Murray (1938).

## The External Method

Focus and Rationale. The focus of the external approach is on maximizing criterion-related validity. This focus is evident in Meehl's (1945) classic explication of the rationale underlying the external approach.

Meehl claimed that it is naive to assume, as had constructors of the earliest temperament scales, that answers to test questions serve as accurate substitutes for direct behavior samples. It was asserted that, although questionnaire responses are related to nontest behavior, the nature of that relationship is often not obvious. In support of this argument, Meehl provided empirical examples from the *Minnesota Multiphasic Personality Inventory* (MMPI) in which individuals with recognized clinical syndromes were differentiated from normals on the basis of item responses that either seem irrelevant to the syndrome (e.g., "I sometimes tease animals" is scored on the MMPI Depression scale) or are actually opposite to how an observer would evaluate the respondents' behavior (e.g., "psychopaths," who are quite rebellious, agree less often than normals with the MMPI item: "I have been quite independent and free from family rule").

Meehl (1945) contended that empirical findings such as these call into question the rational, a priori approach to temperament scale construction, which assumes that "the psychologist building the scale has sufficient insight into the dynamics of verbal behavior and its relation to the inner core of personality that he is able to predict beforehand what certain sorts of people will say about themselves when asked certain sorts of questions" (p. 297). Rather, Meehl concluded that the most defensible approach to temperament scale construction requires selecting items that empirically differentiate groups recognizable on the basis of their nontest behavior as manifesting different levels of the disposition in question. The substantive meaning of any given external scale, although initially restricted to the specific criterion classification used to develop that scale, can be fleshed out through its pattern of relationships with other variables, the "bootstrapping" approach to developing construct validity (Cronbach & Meehl, 1955).

Measures. The two most widely-used products of the external approach are the MMPI and the *California Psychological Inventory* (CPI; Gough, 1975). The MMPI was originally designed for use in psychiatric diagnosis. The constructors of the MMPI, Starke Hathaway and J. Charnley McKinley, were impressed by the success E. K. Strong had achieved in measuring vocational interests using scales constructed by empirically contrasting the item responses of members of various occupations with those of "men-in-general." Adapting this method to their purposes, Hathaway and McKinley generated a large pool of items designed to tap diverse aspects of psychiatric symptomatology and administered these items to members of various psychiatric diagnostic groups and to normals.

The basic clinical scales of the MMPI, such as Depression and Schizophrenia, were assembled from those items for which the responses of the appropriate psychiatric group were significantly different from those of normals. Although the scales of the MMPI were originally designed to detect various psychopathological syndromes, they have also been found to have meaning within normal subject populations (Dahlstrom & Welsh, 1960). The listings of citations in the most recent Mental Measurements Yearbook (Buros, 1978) indicate that the MMPI has been by far the most widely used of all temperament instruments, and its popularity shows no signs of abating (Lanyon, 1984).

Harrison Gough, influenced by the early MMPI milieu at the University of Minnesota, applied the external scale development approach to the assessment of adaptive temperament constructs, in contrast to the emphasis of the MMPI on psychopathological characteristics. Gough designed the scales of the CPI to measure what he calls folk concepts, described by Gough (1965) as "variables used for the description and analysis of personality in everyday life and in social interaction" (p. 295). The CPI is comprised of 18 scales targeted toward various aspects of successful adjustment that are common in the folk wisdom, such as self-acceptance, sociability, and responsibility. Thirteen of these scales were constructed using the external method, four using the internal method, and one on the basis of item endorsement frequencies. The predominant focus of the external method on criterion-related validity is reflected in Gough's (1975) description of the purpose of the CPI scales as "to forecast what a person will say or do under defined conditions, and to identify individuals who will be described in characteristic ways by others who know them well or who observe their behavior in particular contexts" (p. 5).

The CPI and MMPI are designed to provide comprehensive coverage of the domains of successful adjustment and psychopathology, respectively. In addition, the external approach has often been applied to the development of scales for specific situations, such as predicting performance in a particular job (Ghiselli, 1973; Guion & Gottier, 1965).

### The Internal Method

Focus and Rationale. In contrast to the focus of the external method on maximizing criterion-related validity, scale constructors using the internal method have as a whole been concerned with identifying and measuring functional units of temperament, or traits, and mapping out a structural representation of their interrelationships. This measurement goal dictates the construction of scales consisting of items sharing a large core of common variance, a result best achieved when items are assembled into scales on the basis of empirical evidence concerning their intercorrelations.

Factor analysis is one method that is particularly well suited to exploring the internal structure of a pool of items and constructing scales to reflect that structure. In fact, Cattell (e.g., 1965) has contended that the fundamental, underlying sources of variation in temperament, which he calls source traits, can only be isolated through factor analysis. Similarly, Tellegen (1981) has claimed that scale construction based on a succession of factor analyses applied to an item pool that is continually shaped and reshaped on the basis of the factor-analytic results produces scales that embody "natural dispositional units" (p. 219). As with the purely rational and external methods, empirical evidence beyond that used in scale construction is necessary to establish the construct validity of internal scales.

Measures. Guilford was the first to apply factor analysis to the study of temperament, beginning his effort in the mid-1930s with an attempt to identify distinct components of introversion-extraversion. Over the next 10 years, Guilford isolated at least 15 self-report temperament factors (Guilford, 1975), most of which were consolidated in 1949 into the 10 scales of the *Guilford-Zimmerman Temperament Survey* (GZTS; Guilford & Zimmerman, 1949), the most widely used product of Guilford's work.

Another widely recognized factor system is that developed by Eysenck, who began a long succession of factor-analytic studies using Guilford's items. In contrast to the relative specificity embodied in Guilford's system of 10 to 15 factors, Eysenck has contended that the domain is best represented in terms of only three broad factors, which he calls neuroticism, extraversion, and psychoticism. Eysenck's most recent inventory, the *Eysenck Personality Questionnaire* (EPQ; Eysenck & Eysenck, 1975), yields scale scores for these three factors.

Cattell adopted an approach to the isolation of temperament factors that was quite different from that of Guilford and Eysenck. Rather than attempting to confirm and measure hypothesized factors, such as extraversion, Cattell sought to explore the entire domain of temperament in order to map out all of its fundamental dimensions. Beginning with the list of over 17,000 trait terms compiled from the dictionary and condensed by Allport and Odbert (1936), Cattell used a sequence of procedures based first on a subjective grouping of synonyms and later on cluster analysis to ultimately reduce the number of variables to be factor analyzed to a set of 35 trait clusters, each represented by a bipolar rating scale. Factor analysis of peer rankings on these 35 clusters resulted in 12 oblique factors. Cattell later replicated these 12 factors in factor analyses of self-report data, as well as isolating four additional factors. These 16 oblique factors are repesented by the scales of Cattell's *Sixteen Personality Factor Questionnaire* (16PF; Cattell, Eber, & Tatsuoka, 1970).

More recently, Jackson has published two inventories constructed through internal methods, the *Personality Research Form* (PRF; Jackson, 1967) and the *Jackson Personality Inventory* (JPI; Jackson, 1976). Unlike Guilford, Eysenck, and Cattell, who used exploratory factor analyses to determine what traits to measure, Jackson began construction of each of these two inventories by explicitly defining the consructs to be measured. For the PRF, these were 20 of the manifest needs proposed by Murray (1938). The 15 traits selected for the JPI were considered by Jackson (1976) to have "potential for furthering an understanding of the personality functioning of the normal or non-psychopathologically disturbed individual" (p. 9). For each inventory, Jackson first assembled large numbers of candidate items into provisional sales for each prespecified trait. Scale items were then selected from those showing high correlations with the total provisional scale to which they had been assigned and low correlations with other provisional scales, the same convergent-discriminant criterion used in factor-analytic scale construction.

In addition to the widely used published inventories that have been described, a tremendous number of published and unpublished temperament scales have been constructed through internal methods. In fact, the internal approach has probably been overall the most widely utilized of the three scale construction methods.

## Comparison of Psychometric Properties

### Internal Consistency Reliability and Within-Inventory Correlations

The different measurement goals addressed by the internal and external scale construction methods are typically reflected in differences between scales of the two types in internal consistency reliability and within-inventory correlations. The measurement of functional units of temperament is usually thought to require scales with high internal consistency, and thus the internal method is designed to maximize this property.

For example, mean internal consistency values for content scales[1] calculated from the inventory manuals are .81 for the GZTS (10 scales), .74 for the PRF (20 scales), and .78 for the JPI (15 scales). Even higher internal consistency values are reported in the manuals for two recent internal inventories, the *Comrey Personality Scales* (CPS; Comrey, 1970) and the *Differential Personality Questionnaire* (DPQ; Tellegen, 1982), each developed through extensive series of factor analyses. Mean internal consistency values are .93 for the CPS (8 content scales) and .85 for the DPQ (11 content scales).

External scales, on the other hand, are designed to predict specific "real world" criteria, such as effective leadership, academic achievement, and delinquency. Such criteria typically reflect combinations of a number of underlying temperament characteristics rather than single, unitary traits. For example, as groups, delinquents probably differ from "model citizens" in impulse control, hostility, and emotional stability. Therefore, individual scales best capable of predicting complex criteria must themselves be heterogeneous in content and factorially complex. As a result, the int?rnal consistency reliability of external scales is usually lower than that of internal scales (Nunnally, 1967). The mean of the internal consistency coefficients reported in the *MMPI Handbook* (Dahlstrom et al., 1975, college sample) for 9 of the 10 content scales of the standard MMPI profile is .57, while the corresponding value from the *CPI Handbook* (Megargee, 1972, high school samples) for the 12 external CPI content scales is .65.

Most developers of internal scales have viewed basic temperament traits as relatively independent, as indicated by the common use of orthogonal factor analysis to develop internal inventories. As a result, intercorrelations among scales within most internal inventories are relatively low. For example, scale intercorrelation matrices published in the inventory manuals permit calculation of average intercorrelations among content scales (disregarding signs of the correlations) of .25 for the GZTS, .19 for the PRF, .22 for the JPI, .18 for the CPS, and .16 for the DPQ. The lower the correlations among the scales within an inventory, the more nonredundant information each individual scale provides.

On the other hand, correlations among the scales of any given external inventory should largely reflect the correlations among the criterion variables targeted by the scales of that inventory. The clinical (i.e., content) scales of the MMPI were designed to predict various psychopathological disorders. Because different psychopathological characteristics, such as anxiety and depression, often occur together, it is not surprising that the intercorrelations reported in the *MMPI Handbook* (for the college

[1] Throughout this section, the term "content scales" is used to refer to scales designed to measure substantive temperament characteristics, as opposed to the so-called "validity scales" included in many inventories to measure test-taking attitudes and behavior.

samples) among the 10 MMPI profile clinical scales average .33 in magnitude. Similarly, the CPI scales all are designed to predict aspects of successful adjustment, many of which are conceptually related (e.g., Dominance and Social Presence). Accordingly, the intercorrelations among the 16 CPI content scales, as reported in the *CPI Manual* (Gough, 1975), average .31 in magnitude.

Thus, internal scales tend to have higher internal consistency reliability and lower within-inventory correlations than external scales. It is difficult, however, to draw generalizations about the internal consistency and between-scale correlations of purely rational scales. Because this method of scale construction relies solely on rational judgment, these characteristics vary according to the conceptions and skills of individual test constructors.

### Criterion-Related Validity

As described earlier, Meehl (1945) issued the classic argument that external scales should be more criterion-valid than purely rational scales. Some 25 years later, Jackson (1971) argued that, although Meehl's 1945 position was defensible at the time, understanding gained through temperament research had provided the basis for purely rational construction of maximally valid scales. Going further, Jackson challenged the most sophisticated external methods to match the criterion-related validity that could be produced by skilled, or perhaps even inexperienced, item writers operating on a purely rational basis. These two major position papers highlight a history of controversy over the relative criterion-related validity of temperament scales constructed by different methods.

Unfortunately, nearly all studies providing direct comparisons bearing on the relative validity of purely rational, external, and internal scales have used basic research criteria, such as peer ratings and self-report variables, rather than typical applied psychology criteria, such as academic achievement, job proficiency, and psychological adjustment. The results of these comparative studies with primarily basic research criteria are described first, followed by a subjective appraisal of comparative validity for applied psychology criteria.

Basic Research Criteria. Ashton and Goldberg (1973) conducted a study designed to test Jackson's (1971) challenge for external scales to match the criterion-related validity of purely rational scales. These authors paid each of 15 psychology graduate students and 15 nonpsychologists to rationally construct (on the basis of a trait description) one 20-item scale to measure either sociability, achievement, or dominance. Thus, 10 new purely rational scales, five by psychologists and five by nonpsychologists, were generated for each trait. Averaged peer ratings on each of the three targeted traits were correlated with the new rational scales, one external CPI scale, and one internal PRF scale designed to measure that trait. Subjects were 168 college women. The more spectacular aspect of Jackson's challenge was clearly unsupported, as the average validity of the nonpsychologists' rational scales ($r$=.18) was clearly lower than that of the external CPI scales ($r$=.27). However, the average validity of the psychology students' rational scales ($r$=.29) compared favorably to that of the external scales. The most valid scales were the most internally consistent psychology student scales (average $r$=.34) and those of the PRF (average *r*=.35), these two representing the internal method.

Ashton and Goldberg's (1973) test of Jackson's challenge was extended by Jackson (1975) himself to different samples of item writers and traits. In this study, each of 22 undergraduate psychology students rationally constructed one 16-item scale based on a description of the trait of self-esteem, social participation, or tolerance. A total of 116 college females, comprising pairs of roommates, completed these 22 scales, three like-named CPI scales (Social Presence[2], Sociability, and Tolerance), and the internally developed JPI, which also contains scales for these traits. Self-ratings and roommate ratings on the three traits of interest were used as criteria. For the self-ratings, the average validities across the three traits were .31 for the external (CPI) scales, .44 for the purely rational (average student) scales, and .46 (most internally consistent student) and .51 (JPI) for the internal scales. The corresponding values for the average peer rating validities were .09 (CPI), .25 (average student), .28 (most internally consistent student), and .29 (JPI). The far poorer relative performance of the CPI scales in this study compared to that of Ashton and Goldberg was probably partly due to the use of 16-item subscales drawn randomly from the full CPI scales to make them comparable in length to the psychology student and JPI scales.

Unlike the studies of Ashton and Goldberg (1973) and Jackson (1975), Hase and Goldberg (1967) restricted their comparative validity study to scales constructed from a common item pool, that of the CPI. One purely rational, one external, and two internal sets of 11 scales each were found to have virtually identical average validities in predicting 13 peer rating and biographical criteria in a sample of 201 university freshmen women. However, the lack of differential validity among scales constructed by different methods may have been largely determined by use of a common item pool. In the two later studies in which item pools were allowed to vary between scale construction methods, the internal scales were somewhat more valid than the purely rational and external scales.

Validity of Subtle Versus Obvious Items. Investigations of the validity of subtle versus obvious items provide another body of research related to the comparison of scale construction methods. Meehl's (1945) position paper for the external strategy endorsed the use of scale items whose relationships to the dimension being measured may be nonintuitive, or even counterintuitive. Examples of these from Meehl were given earlier. Jackson (1971), in his argument for rational item selection, hypothesized that so-called subtle items appearing on externally constructed scales are simply mistakes, fortuitous (and invalid) products of the characteristics of the particular comparison groups used in scale development. At issue, then, is whether subtle items enhance scale validity by capitalizing on remote but valid trait indicators only discoverable through external scale construction, or whether they actually attenuate validity relative to that attainable through strict reliance on obvious items.

[2] Although Jackson identified all three CPI scales as examples of the external approach, Social Presence was actually developed by internal methods. However, exclusion of the results for this trait from the comparisons would not have altered the conclusions.

Most comparisons of the validity of subtle versus obvious items have used the MMPI as the predictor instrument and other self-report scales as criteria. Exceptions to the use of test data as criteria are the studies by Wiener (1948), McCall (1958), and Duff (1965), all of which found obvious MMPI items to be more valid than subtle items. Wiener divided each of five MMPI scales into subscales consisting of subtle and obvious items and contrasted the mean subscale scores of 50 veterans who were successful in school or on-the-job training with those of 50 who were unsuccessful. For each of the five scales, the subtle subscale was less valid (median $r_{pb}$=-.08) than the obvious subscale (median $r_{pb}$=.24), so that the total MMPI scale was actually less valid (median $r_{pb}$=.15) than the obvious subscale alone. McCall found the same pattern when studying the validity of the MMPI Depression scale for differentiating 41 depressive from 41 nondepressive psychotics. The mean differences between these groups can be expressed as point-biserial correlations of .26 for the "face valid" subscale, .03 for the "irrelevant" subscale, and .22 for the total scale. Finally, for each of three MMPI scales, Duff correlated the subtlety of individual scale items with their validity for differentiating a psychiatric sample targeted by that scale from normals. Correlations between item subtlety and item validity were -.48, -.38, and -.22 for the three scales.

Among studies using test data as criteria, Gynther and Burkhart (1983) have described the results of an extensive series of comparisons of the validity of subtle and obvious MMPI items. A variety of criteria, all test data, were correlated in various studies with obvious, neutral, and subtle subscales of four MMPI scales. The general trend of results for two of the scales (Depression and Hysteria) was the same as for the earlier studies just summarized--the subtle subscales diluted the validities of the full scales to levels below those achieved by the obvious subscales alone. The subtle subscales for two other scales (Psychopathic Deviate and Mania) did offer some evidence of enhancing overall validity. In a study that did not use the MMPI, Holden and Jackson (1979) found negative correlations, ranging between -.20 and -.44, between subtlety scores of PRF items and the empirical validities of those items against three self-report criteria.

These findings indicate that subtle items, often considered a unique virtue of the external approach to scale construction, are less valid than obvious items and may actually detract from overall scale validity. However, all of this research has been conducted in contexts where there has been no apparent motivation to fake responses. In one study that included directed faking instructions, Holden and Jackson (1981) found no significant differences between the validity of subtle and obvious PRF subscales for predicting a composite self-report criterion. However, evidence that is reviewed later in this section shows that the effects of faking in actual decision-making contexts are not validly represented by the results of studies in which subjects are explicitly instructed to fake. Research is needed to determine whether the greater validity of obvious than subtle items in situations without motivation to distort generalizes to decision-making contexts.

Applied Psychology Criteria. Taken as a whole, studies providing direct comparisons have found internal scales and obvious items to be more criterion-valid than externa' scales and subtle items. However, this conclusion may well apply only to the kinds of criteria, such as peer ratings and self-report variables, that have nearly always been used in these studies. Criterion constructs measured in these ways are more likely to resemble the homogeneous, unidimensional temperament constructs embodied in internal scales than the more heterogeneous, factorially complex constructs represented by external scales. On the other hand, this same line of reasoning suggests that external scales might be somewhat more valid than the other kinds of scales for predicting most applied psychology criteria, at least at the level of individual scale correlations.

As discussed earlier, applied psychology criteria are usually multifaceted, and these facets may be related to individual differences in several temperament variables. At the level of individual scale correlations, factorially complex external scales that tap the various temperament components of complex criteria could be expected to be more valid than the more homogeneous, unidimensional scales produced by the internal and purely rational methods. Although the evidence on this question is only suggestive, there are some indications that this may indeed be the case.

Studies of the criterion-related validity of temperament scales for predicting five major classes of applied psychology criteria are reviewed in some depth later in this section. Unfortunately, none of those studies was specifically designed to directly compare the validity of measures of the same constructs developed by different methods, and too few included scales representing different construction methods, regardless of the constructs assessed, to permit any meaningful comparison on that basis.

In the absence of any more substantial evidence, a subjective appraisal of that body of criterion-related validity research suggests that individual external scales, particularly those of the CPI, have tended to be somewhat more valid than purely rational and internal scales. This evaluation echoes that offered by Goldberg in 1972. While recognizing the possibility that newer, psychometrically more refined (i.e., internal) inventories might ultimately demonstrate greatest validity, Goldberg concluded that "at least for the next five years, the knowledgeable practitioner should be able to provide more valid nontest predictions from the CPI than from most other comparable instruments on the market today" (p. 96).

In fact, it is probably fair to characterize the "conventional wisdom" as holding that the externally constructed MMPI and CPI are the published inventories most widely used in applied settings because they are "the tests that work." It has also been suggested that temperament scales specially developed for specific prediction situations using external methods have proven more valid in those situations than have other types of scales (e.g., Guion & Gottier, 1965).

Integration. Do these appraisals, subjective as they may be, indicate that externally constructed scales are the temperament measures of choice for applied prediction situations? Not necessarily. When an overall criterion is analyzed into its homogeneous components, it should be possible to predict applied psychology criteria at least as well using weighted composites of homogeneous scales that target these components as with heterogeneous external scales (Nunnally, 1967).

Nunnally has recommended the assembly of a "catalog" of homogeneous temperament measures shown through factor analytic studies to represent the domain. The use of weighted composites of such scales for applied prediction problems offers both practical and theoretical advantages over the use of external scales. On the practical side, weighting homogeneous scales optimally for each specific prediction situation should prove more valid than using previously developed external scales, in which the number of items measuring each component fixes its contribution to the overall scale score and which have been developed using criteria that may differ somewhat from those of current interest. This approach is also more efficient than constructing new external scales for every new situation. The theoretical advantage stems from the more apparent interpretability of constructs represented by homogeneous internal scales compared to those embodied in heterogeneous external scales, which facilitates a more conceptual understanding of relationships between predictors and criteria.

In such a construct-guided approach to prediction, internal structure and external relations are both important. However, the successful application of this approach hinges on two very important requirements. One is a very comprehensive and accurate analysis of the criterion behavior to be predicted as, for example, through a good job analysis. The other involves the explication of the domain of temperament constructs and the identification of measures of those constructs. A taxonomy of temperament constructs and existing measures of them is proposed below.

## Structure of The Temperament Domain: A Proposed Taxonomy of Temperament Scales

### Background

A construct-guided approach to the prediction of applied psychology criteria from temperament scales would be greatly aided by a "catalog" or taxonomy specifying the primary constructs in the temperament domain and identifying measures of these constructs. By the same token, a taxonomy of temperament scales would facilitate the meaningful organization and interpretation of the results of previous criterion-related validity studies using temperament measures. None of the three major recent reviews of the validity of temperament scales in predicting job performance has proceeded in this fashion. Guion and Gottier (1965) grouped results of individual studies by inventory, and Ghiselli and Barthol (1953) and Ghiselli (1973) simply averaged results for all temperament measures within a number of different job types. Such methods of summarization could obscure relations between temperament predictor constructs and various criterion constructs that may actually be supported in the literature.

The organizing function that a taxonomy of temperament scales could serve is particularly desirable in light of the tremendous proliferation of temperament traits and measures that have been advanced. Multiscale temperament inventories are now the most numerous of all types of published tests (Jackson & Paunonen, 1980). The already staggering number of published tests undoubtedly increases many fold when all of the various unpublished scales that have been devised are included. These hundreds of published and unpublished scales stem from a vast diversity of conceptions of the important sources of variation in temperament. A hint of this diversity can be gleaned from the previous descriptions of a few of the most widely used multiscale inventories, which are designed variously to measure psychiatric syndromes (MMPI), folk concepts (CPI), basic needs proposed by Murray (EPPS. PRF), and dimensions isolated through various methods of factor analysis (GZTS, EPQ, 16PF, CPS, DPQ).

To further complicate matters, ıt is difficult to assess the comparability of temperament scales on the basis of scale names alone. As Ghiselli (1973) has succinctly stated, "In some cases different names are used to denote the same, or very nearly the same, quality, and in others the same name is used to denote quite different qualities" (p. 464). For this reason, a categorization of temperament scales according to constructs cannot simply be based on a rational sorting of scales according to their labels, but must instead rely on empirical evidence concerning between-scale correlations. Beyond that, some system of temperament constructs is necessary to provide the categories for such an analysis.

Unfortunately, there is little agreement on the identification or even the approximate number of primary or "first-order" (in factor-analytic terminology) temperament dispositions, those traits that most self-report temperament scales are designed to measure. For example, among researchers who have used factor analysis to explore the domain, Cattell et al. (1970) claimed that their list of 24 primary factors was exhaustive, whereas Guilford (1975) suggested 58.

Current researchers are converging, however, on a smaller number of more general or "higher order" sources of variation that underlie the diverse concepts tapped by various self-report temperament scales. The seminal work in this area was carried out by Tupes and Christal (1961) and Norman (1963). These authors found five basic dimensions emerging from factor analyses of peer ratings and nominations. Norman called these dimensions Surgency, Emotional Stability, Agreeableness, Conscientiousness, and Culture.

Goldberg (1981) has enthusiastically endorsed the primacy of these five higher order dimensions in the self-report domain as well. Other researchers who endorse fewer than five higher order dimensions nevertheless tend to identify factors recognizable in terms of these five (e.g., Block, 1965; Eysenck & Eysenck, 1969; Guilford, 1975; Tellegen, 1982). In the most recent summary of this work, Hogan (1983a) has proposed a slightly modified set of six basic higher order dimensions. Four of these are aligned with the last four Norman factors listed previously, but Hogan maintains that the Dominance and Affiliation components of the Norman Surgency factor are independent enough to warrant separate assessment.

A framework consisting of these five or six higher order temperament dimensions, then, allows integration of the positions of many major researchers in the field and thus provides the most compelling basis for a taxonomic effort. Accordingly, as part of the present review, a taxonomic analysis was carried out in which the scales of a dozen major temperament inventories were classified into a taxonomy of six higher order dimensions patterned after those proposed by Hogan. The names chosen for these six dimensions are: Potency, Affiliation, Adjustment, Agreeableness, Dependability, and Intellectance. The alignments of the higher order dimensions in the systems previously mentioned with these six are shown in Table 1. The classification of scales into these categories was performed on the basis of published correlations. Analysis of these correlations also allowed an empirical evaluation of the viability of the proposed classification system.

Table 1

Alignment of Higher-Order Temperament Factors Proposed by Various Researchers with the Present Six-Category Taxonomy

| Source | | | | | | |
|---|---|---|---|---|---|---|
| Present System | Potency | Affiliation | Adjustment | Agreeableness | Dependability | Intellectance |
| Hogan (1982) | Ascendance | Sociability | Adjustment | Likeability | Self-Control | Intellectance |
| Tupes and Christal (1961) | ———Surgency——— | | Emotional Stability | Agreeableness | Dependability | Culture |
| Norman (1963) | ———Surgency——— | | Emotional Stability | Agreeableness | Conscientiousness | Culture |
| Block (1965) | | | Ego Resiliency | | Ego Control | |
| Eysenck and Eysenck (1969) | ———Extraversion——— | | Neuroticism | ———Psychoticism——— | | |
| Guilford (1975) | ———Social Activity——— | | Emotional Stability | Paranoid Disposition | Introversion | |
| Tellegen (1982) | Positive Affectivity | | —Negative Affectivity— | | Constraint | |

## Method

The 12 current multiscale temperament inventories that appear to be most widely utilized in basic and applied research were selected for analysis. These are: *California Psychological Inventory* (CPI; Gough, 1975), *Comrey Personality Scales* (CPS; Comrey, 1970), *Differential Personality Questionnaire* (DPQ; Tellegen, 1982), *Edwards Personal Preference Schedule* (EPPS; Edwards, 1959), *Eysenck Personality Questionnaire* (EPQ; Eysenck & Eysenck, 1975), *Gordon Personal Profile-Inventory* (GPPI; Gordon, 1978), *Guilford-Zimmerman Temperament Survey* (GZTS; Guilford, Zimmerman, & Guilford, 1976), *Jackson Personality Inventory* (JPI; Jackson, 1976), *Minnesota Multiphasic Personality Inventory* (MMPI; Dahlstrom et al., 1972, 1975), *Omnibus Personality Inventory* (OPI; Heist & Yonge, 1968), *Personality Research Form* (PRF; Jackson, 1967), *Sixteen Personality Factor Questionnaire* (16PF; Cattell et al., 1970). Although all instruments are designed to provide self-report assessment of temperament variables, these inventories are extremely diverse in rationale, purpose, and scale construction methodology. The scales included in the taxonomic analysis were the 146 content scales from the 12 inventories (i.e., all scales except the "validity" scales). These are listed in Appendix A.

A search of the literature was undertaken in an attempt to locate entries for as many of the cells of the 146 x 146 between-scale correlation matrix as possible. Primary sources of correlations were test manuals and handbooks, as well as research reports located during the course of the literature search in the temperament domain. Among the 12 inventories, there are 12 possible intra-inventory and 66 possible inter-inventory correlation matrixes, making a total of 78 matrixes. The literature search resulted in one or more potential candidates for 40 of these 78 matrixes. In those few instances when more than one source for the same correlation matrix was located, the results were always quite similar, so the source reporting the larger sample was utilized. Many of the 40 sources selected on this basis reported correlations separately for more than one sample, usually males and females. In these cases, the separate matrixes were averaged. In all, 5,313, or just over 50%, of the 10,585 possible entries in the 146 x 146 matrix were obtained. The 40 sources contributing these correlations and brief descriptions of the sample characteristics are also provided in Appendix A.

Eleven of the 38, or almost one-third, of the missing inter-inventory correlation matrixes are accounted for by the absence of published correlations between the CPS and the other inventories. Other than the CPS, each inventory was related with between two (JPI) and eight (CPI) other inventories. Sample sizes were quite variable, ranging from 45 (CPI with EPPS) to 50,000 (MMPI with itself), with a median sample size of 217. Most of the matrixes were based on college samples.

An initial grouping was carried out by classifying each of the 146 scales into either one of the six higher order content categories described above or a seventh category, Miscellaneous scales. Each scale was tentatively assigned to one of the seven categories on the basis of item content, available factor-analytic results, and a visual evaluation of its correlations with other scales. As previously mentioned, correlations between the CPS scales and scales from the other inventories were unavailable. However, the published results of factor analyses of the CPS along with the GZTS (Comrey, Jamison, & King, 1968) and with the EPQ and 16PF (Comrey & Duffy, 1968) allowed fairly confident assignment of the CPS scales.

Inspection of the resulting within-category correlation matrixes allowed identification of scales whose relationships poorly fitted prior expectations. For each of these scales, calculation of mean correlations with the scales constituting various other content categories allowed classification into the appropriate content category or the Miscellaneous category. In implementing this procedure, an effort was made to make reasonable assignments of as many scales as possible to actual content categories, rather than assigning scales en masse to the Miscellaneous category in order to maximize the "purity" of the final result.

## Results

One hundred and seventeen of the 146 scales (80%) were classified into higher order content categories as follows: Potency (26), Adjustment (23), Agreeableness (16), Dependability (23), Intellectance (17), Affiliation (12). The remaining 29 scales were grouped in the Miscellaneous category. The means and standard deviations of the within-category and between-category correlations, as well as the number of correlations on which these values are based, are given in Table 2. These means are based on the alignment of scales within each content category in the same direction (e.g., the correlations between "Neuroticism" scales and "Emotional

Stability" scales were reversed in sign). An attempt was made to align all scales assigned to the Miscellaneous category in the "socially desirable" direction.

Table 2

Mean Within-Category and Between-Category Correlations for Six-Category Taxonomy

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Potency | Mean r=.46<br>SD r=.16<br>N r=146 | | | | | | |
| Adjustment | Mean r=.20<br>SD r=.18<br>N r=321 | Mean r=.43<br>SD r=.19<br>N r=165 | | | | | |
| Agreeableness | Mean r=.04<br>SD r=.17<br>N r=173 | Mean r=.24<br>SD r=.16<br>N r=162 | Mean r=.37<br>SD r=.14<br>N r= 44 | | | | |
| Dependability | Mean r=-.08<br>SD r= .16<br>N r= 286 | Mean r=.13<br>SD r=.20<br>N r=276 | Mean r=.06<br>SD r=.17<br>N r=166 | Mean r=.34<br>SD r=.18<br>N r=121 | | | |
| Intellectance | Mean r=.12<br>SD r=.15<br>N r=175 | Mean r=.02<br>SD r=.14<br>N r=193 | Mean r=.04<br>SD r=.16<br>N r= 94 | Mean r=-.12<br>SD r= .18<br>N r= 162 | Mean r=.40<br>SD r=.19<br>N r= 52 | | |
| Affiliation | Mean r=.09<br>SD r=.21<br>N r=157 | Mean r=.00<br>SD r=.16<br>N r=150 | Mean r=.10<br>SD r=.17<br>N r= 98 | Mean r=.08<br>SD r=.14<br>N r=160 | Mean r=-.14<br>SD r= .15<br>N r= 84 | Mean r=.33<br>SD r=.16<br>N r= 45 | |
| Miscellaneous | Mean r=.09<br>SD r=.17<br>N r=392 | Mean r=.12<br>SD r=.18<br>N r=419 | Mean r=.02<br>SD r=.18<br>N r=215 | Mean r=.02<br>SD r=.18<br>N r=361 | Mean r=.04<br>SD r=.17<br>N r=242 | Mean r=-.04<br>SD r= .15<br>N r= 208 | Mean r=.05<br>SD r=.20<br>N r=246 |
| | Potency | Adjustment | Agreeableness | Dependability | Intellectance | Affiliation | Miscellaneous |

The mean correlations shown in Table 2 display an appropriate convergent-discriminant structure. The means within content categories, which range from .33 for Affiliation to .46 for Potency, all are substantially higher than all of the between-category means. The actual scales assigned to each category and the correlation matrixes within content categories are included in Appendix A. Within each content category the scales can be roughly grouped into measures of more specific aspects of the highly general dispositions. However, this subgrouping is only offered as tentative. Many scales tap more than one aspect of a particular higher order dimension and, indeed, some scales appear to mix aspects of different higher order dimensions. Scales of the latter kind typically display low internal consistency. Nevertheless, such rough subdivisions of scales within higher order content categories help to flesh out the substance of the more general temperament dispositions. Results within each category are now briefly discussed.

Potency. The 26 scales classified under Potency display the highest degree of interrelationship of any of the six content categories, with an average correlation of .46. The correlations among and between scales measuring dominance, outgoingness, and the five CPI scales included in this category are particularly high, the CPI scales tapping both of these more specific aspects of Potency. Scales measuring energy level are more weakly related to the rest, but more strongly related to Potency than to any other content category.

Adjustment. The mean correlation among these 23 scales is also quite high, .43. The core concept of this dimension is embodied in the scales measuring emotional stability. These 11 scales, all developed through internal consistency methods, display an average intercorrelation of .63. The CPI and MMPI scales in this category are more heterogeneous, having been developed through external methods, and as a result their patterns of correlations are more variable. In particular, the correlations between the CPI and MMPI scales (other than K) are the lowest in the matrix.

Agreeableness. This category contains fewer scales (16) than either of the first two and fewer within-category correlations (44) than any of the other categories. The mean of these 44 correlations is .37. More specific aspects of Agreeableness are (lack of) cynicism, (lack of) aggression, cooperativeness, and helpfulness (nurturance).

Dependability. This category represents one of the more diverse groupings of content, as reflected by the average correlation of .34 among the 23 scales. The varied aspects of Dependability include nondelinquency, orderliness, nonimpulsiveness, cautiousness, and harm avoidance.

Intellectance. The Intellectance category (17 scales) is dominated by scales from just two inventories, the OPI (7 scales) and the JPI (4 scales), and intra-inventory correlations from these two account for 27 of the 52 correlations in the matrix. The average within-category correlation is .40. Scales measuring (absence of) dogmatic views, cognitive complexity, original thinking, intellectual interests, and reflectiveness are represented here.

Affiliation. Hogan's (1983a) contention that Potency and Affiliation are conceptually and empirically distinct seems reasonably well supported by the average correlation of .33 among the 12 scales grouped into the Affiliation category, as compared to the average value of only .09 between Affiliation and Potency scales. The scales in this category can be roughly cast along a continuum of social dependency, from simply preferring to be with people (affiliation), to not wanting to be independent (lack of autonomy), to behaving in a manner designed to win approval (conformity), to wanting to be cared for (succorance).

Miscellaneous. The 29 scales assigned to this category are conceptually distinctive (e.g., PRF Sentience) and/or display little communality with those grouped into content categories (e.g., MMPI Paranoia). There are, however, two identifiable clusters of scales contained in this otherwise heterogeneous category. Evaluated as groups, neither the various Masculinity-Femininity scales nor the various Achievement Motivation scales are strongly enough related to any content categories to be reasonably included. Although these constructs may often be worth evaluating for specific applied prediction problems, they simply cannot be well tied into the present six-category scheme. Among all 29 Miscellaneous scales, the average correlation is only .05.

Between-Category Correlations. The highest average between-category correlation, which is clearly lower than any of the within-category means, is .24 between Adjustment and Agreeableness. This is not an unexpected result, in light of both the conceptual relationship between these two and previous empirical findings. Indeed, Guilford (1975) has proposed the merger of his second-order Emotional Stability (Adjustment) and Paranoid Disposition (Agreeableness) factors into an even higher-order factor, one that he calls Emotional Health.

The second highest average between-category correlation is .20 between Potency and Adjustment. This correlation is probably due to the positive relationship between the social confidence component of Potency and the high self-esteem component of Adjustment. All of the other between-category means are smaller than .14 in magnitude.

### Conclusions

It would obviously have been possible to increase the mean within-category correlations by successively relegating more weakly related scales to the Miscellaneous category, up to the limit of the highest single between-scale correlation in each category. However, such a procedure would also successively reduce the usefulness of the classification scheme for interpreting results of previous criterion-related validity studies in terms of substantive predictor categories. The results presented in Table 2, based on the assignment of 80% of a very diverse group of scales to content categories, provide sufficient empirical justification for the proposed classification system. Therefore, this taxonomy is used in the following subsection as a framework for organizing and interpreting the results of criterion-related validity studies that have used self-report temperament scales as predictors.

## Criterion-Related Validity

The validity of self-report temperament scales for predicting various applied psychology criteria bears importantly on both basic and applied research issues in temperament assessment. As discussed earlier, demonstrations of criterion-related validity for temperament scales in general support the viability of the trait conceptions upon which these scales are based. Such evidence also argues against the interpretation of temperament scale scores as primarily reflecting response sets rather than substantive temperament characteristics, an issue that is discussed later. More generally, the criterion-related validity of temperament scales reflects on their construct validity (Cronbach & Meehl, 1955). In applied research, evidence concerning the criterion-related validity of temperament scales provides information on the likelihood that they will contribute to the prediction of criteria of interest.

In this subsection we review the results of criterion-related validity studies using temperament scales as predictors of organizationally relevant criteria. The discussion is organized in the framework of five major categories of criterion variables: education, training, job proficiency, job involvement/withdrawal, and adjustment. Within each category, the results for military and civilian studies are considered separately and in combination. Military studies included in the present review consist of those published subsequently to the 1948 review by Ellis and Conrad of the validity of temperament inventories in military settings. The review of civilian studies is largely restricted to the period from 1960 to the present. Several earlier reviewers of civilian studies have covered the period prior to 1960, and their findings are also discussed..

As part of the review, criterion-related validity coefficients for measures grouped into 10 predictor categories are summarized in tabular form within each of the five criterion categories just listed. Eight of the 10 predictor categories derive from the taxonomy of temperament scales described earlier. These eight consist of the six highly general construct categories--Potency, Adjustment, Agreeableness, Dependability, Intellectance, and Affiliation--plus the two more circumscribed content categories, Achievement Motivation and Masculinity-Femininity. The latter two were grouped among the Miscellaneous scales of the taxonomy because of their empirical distinctiveness from the higher order categories. The remaining scales grouped into the Miscellaneous category cannot be aligned with any temperament construct, so it is not meaningful to summarize the validity results for these scales as a group.

The two additional predictor categories fall outside of the temperament taxonomy proposed here. One of these is the widely studied locus of control construct. Individuals with an internal locus of control believe that they control the outcomes that they experience through their actions, while those with an external locus of control believe that what happens to them is largely determined by fate or chance. Locus of control has almost always been operationalized using the Rotter Internal-External Locus of Control (I-E) scale (Rotter, 1966). Since the present taxonomy is of multiscale temperament inventories, the locus of control construct was not included. However, the conceptual and empirical distinctiveness of locus of control from the constructs included in the taxonomy, combined with evidence supporting its relationships with a number of aspects of behavior in organizations (see Spector, 1982, for a review), warrant its inclusion in the summary tables as a separate predictor category.

The tenth predictor category is used to summarize results gleaned from military research reports that did not describe the content of temperament predictors used in sufficient detail to allow their classification into any of the previous nine construct categories. Although these Unclassified Military scales cannot be aligned with any unified temperament construct, they have invariably been scored in the direction expected to correlate positively with favorable outcomes, so that the combination of positive and negative validity coefficients for different scales in this category is unambiguous.

To be included in a validity summary table, a study had to satisfy three conditions: (a) validity results were reported either in terms of some index of correlation--product-moment correlation (*r*), point-biserial correlation ($r_{pb}$), biserial correlation ($r_{bis}$), tetrachoric correlation ($r_{tet}$), or the correlation ratio (*eta*)--or in such a manner that they could be converted to one of those correlational indexes according to methods described by Glass (1977); (b) the predictors could be classified into one of the 10 predictor categories being used, on the basis of either the empirical classification of major inventory scales previously described or scale descriptions provided in the reserch report; (c) numerical results were reported for all temperament predictors studied, so that the results of the summary tables would not be biased toward higher validity coefficients because of the inclusion of studies reporting only the results that reached practical or statistical significance. Some of the criterion-related validity studies located in the literature search that did not meet these conditions for inclusion in the summary tables are included in the discussion.

## Educational Criteria

Criterion-related validity coefficients for temperament scales as predictors of educational criteria are summarized in Table 3. All studies summarized in Table 3 were conducted in civilian settings and used some index of grade-point average (GPA) as the criterion.

As can be seen, the Achievement scales have shown the most evidence of usefulness as predictors of educational performance. Such positive results have been obtained for CPI Achievement via Independence (Farr, O'Leary, Pfeiffer, Goldstein, & Bartlett, 1971; Gough, 1964; Griffin & Flaherty, 1964), EPPS Achievement (Weiss, Wertheimer, & Groesbeck, 1959), and PRF Achievement (Harper, 1975). The eight validity coefficients for these three scales derived from six different studies (those five listed above plus Johnson, 1973) have a median of .30, and only one of the eight is below .23. The one coefficient of .32 for the Rotter I-E Locus of Control scale (Nord, Connelly, & Daignault, 1974) is perhaps related to the success of the Achievement scales. That is, although the I-E scale was not designed as an achievement motivation scale, some of its content does reflect the degree of belief in the efficacy of hard work. Among the studies mentioned above, Weiss et al., Gough, and Nord et al. all found temperament scales to add significantly to aptitude measures in the prediction of GPA. In summary, various achievement-related scales have demonstrated practically useful levels of validity as predictors of educational criteria.

## Training Criteria

Military. In an early review of the validity of temperament scales in military practice, Ellis and Conrad (1948) summarized results separately for adjustment criteria (to be discussed later) and "performance" criteria. Twenty-seven of the 36 validity results grouped by Ellis and Conrad under "performance" criteria involved pass/fail outcomes among military trainees, while the remaining nine pertained to proficiency in military jobs. Hence this section of their review is most applicable to the prediction of success in military training.

Table 3

Summary of Criterion-Related Validity Studies That Used Temperament Predictors and Educational Criteria

| Temperament Predictor Category | Number of Studies | Median Sample Size | Number of Different Predictors | Number of *r*'s | Median *r* |
|---|---|---|---|---|---|
| Potency | 12 | 194 | 16 | 42 | .06 |
| Adjustment | 11 | 229 | 16 | 43 | .14 |
| Agreeableness | 6 | 162 | 6 | 9 | .03 |
| Dependability | 12 | 194 | 11 | 24 | .13 |
| Intellectance | 6 | 162 | 4 | 6 | .17 |
| Affiliation | 2 | 64 | 5 | 5 | -.03 |
| Achievement | 6 | 288 | 3 | 8 | .30 |
| Masculinity | 6 | 213 | 4 | 8 | -.16 |
| Locus of Control | 1 | 50 | 1 | 1 | .32 |

Ellis and Conrad concluded that temperament scales had been generally ineffective in predicting "performance" (i.e., training and job proficiency) criteria; however, most of the negative results for training studies were found among Air Force pilot trainees. For example, defining successful prediction as at least one predictor scale achieving a statistically significant validity coefficient of .20 or more, only 2 the 15 results summarized for pilot trainees were successful, compared to 9 of the 12 results for other groups of trainees. Ellis and Conrad did not describe the results that they summarized in sufficient detail (e.g., validity results for all predictors) to allow their inclusion in the present summary tables. However, since most of the temperament inventories used in early military practice, such as the *Personal Inventory*, the *MMPI*, and the *Bell-MMPI-Army Adjustment Inventory*, were designed for psychiatric screening (Ellis & Conrad, 1948), it is likely that most of the scales would be classified in the Adjustment category of the predictor taxonomy. In conclusion, the early studies summarized by Ellis and Conrad suggested some promise for temperament scales, perhaps primarily those in the Adjustment category, as predictors of success in military training for groups other than pilots.

The present review resulted in the summary of validity coefficients for training criteria shown in Table 4. Criteria used in the studies summarized in Table 4 included objective (e.g., course grades), subjective

Table 4

Summary of Criterion-Related Validity Studies That Used Temperament Predictors and Training Criteria

| | Summary Information for Training Criteria | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Military Studies | | | | | Civilian Studies | | | | | Combined Military & Civilian Studies | | | | |
| Temperament Predictor Category | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r |
| Potency | 7 | 163 | 8 | 27 | .13 | 3 | 130 | 7 | 9 | .14 | 10 | 143 | 8 | 36 | .13 |
| Adjustment | 8 | 185 | 7 | 20 | .20 | 3 | 130 | 7 | 8 | .16 | 11 | 156 | 7 | 28 | .19 |
| Agreeableness | 5 | 185 | 1 | 5 | .08 | 0 | — | -- | — | — | 5 | 185 | 1 | 5 | .08 |
| Dependability | 8 | 146 | 5 | 16 | .12 | 3 | 130 | 3 | 4 | .18 | 11 | 144 | 5 | 20 | .12 |
| Intellectance | 5 | 185 | 1 | 5 | .19 | 0 | — | — | — | — | 5 | 185 | 1 | 5 | .19 |
| Affiliation | 0 | — | — | — | — | 0 | — | — | — | — | 0 | — | — | — | — |
| Achievement | 2 | 98 | 1 | 3 | .46 | 1 | 141 | 1 | 1 | .20 | 3 | 120 | 1 | 4 | .33 |
| Masculinity | 1 | 81 | 1 | 2 | .06 | 1 | 141 | 1 | 1 | .14 | 2 | 98 | 1 | 3 | .09 |
| Locus of Control | 1 | 146 | 1 | 1 | .30 | 1 | 140 | 1 | 1 | .27 | 2 | 143 | 1 | 2 | .29 |
| Unclassified Military Scales | 3 | 254 | 8 | 8 | .18 | | | | | | 3 | 254 | 8 | 8 | .18 |

(e.g., instructor ratings), go-no go (e.g., pass/fail), and hands-on training performance measures. The summary information for military training criteria is based on a relatively small number of studies. Given this limitation, the Adjustment (median *r*=.20) and the Unclassified Military (median *r*=.18) predictor categories have the highest median correlations that are based on more than five coefficients, while smaller amounts of evidence support the validity of Achievement (median *r*=.46 based on three coefficients) and Locus of Control (one coefficient of .30) scales.

Temperament scales in general, and Achievement scales in particular, have predicted military classroom performance rather well. For example, the Achievement via Independance scale of the CPI proved the most valid of all CPI scales administered to each of four samples from three different studies of military classroom performance. Rosenberg, McHenry, Rosenberg, & Nichols (1962) related all scales of the CPI to course grades of Army psychiatric assistant trainees and found predictive validity coefficients of .46 (*N*=98 in an 8-week course) and .47 (*N*=64 in a 4-week course) for Achievement via Independence in separate samples. Predictive validity coefficients greater than .30 were also found for six other CPI scales, including Tolerance (Adjustment category, *r*=.42), in the sample of 98, and for four other CPI scales, including Capacity for Status (Potency category, *r*=.44), in the sample of 64. In both samples, Achievement via Independence added significantly to an aptitude measure in multiple correlations.

In a study by Collins (1967) using the entire CPI, Achievement via Independence and Tolerance both correlated .41 with class rank in a sample of 59 Army drill sergeant trainees. (This study is not included in Table 4 because correlations were not reported for all CPI scales.) Datel, Hall, and Rufe (1965) administered three CPI scales to 290 Army foreign language trainees prior to a 47-week training course. Achievement via Independence again predicted classroom performance, in this case course grades, better than the other CPI scales, although its validity was quite a bit lower (*r*=.14) than in the other two studies.

The three studies whose results are grouped in the Unclassified Military scales category of Table 4 all used rating criteria. In one of these studies, six scales of the Army Self-Description Blank produced validity coefficients ranging between .15 and .20 in predicting peer ratings of combat potential which were gathered 16 weeks after testing for 166 Army field artillery trainees (Adjutant General's Office, 1957). In the same study, a predetermined two-scale composite had a predictive validity of .21. Berkhouse and Cook (1961) found the combat suitability scale of the Army Self-Description Blank to be unrelated (*r*=.02) to combined supervisory and peer ratings of 254 Army Special Forces trainees. However, the Classification Inventory, a non-cognitive measure consisting of 25 interest items, 25 temperament items, and 75 mixed biodata and temperament items, achieved a predictive validity of .23 against this criterion, one of the higher values among the various cognitive and non-cognitive predictors employed. Finally, Tubiana and Ben-Shakhar (1982) developed a temperament scale that correlated .34 with combined supervisory and peer ratings gathered after the 2-to 8-month basic training period for 459 Israeli soldiers. The one study contributing the validity coefficient of .30 for the Rotter Locus of Control scale (Lied & Pritchard, 1976) also used a rating criterion, instructor ratings of effort, in a sample of 146 Air Force technical trainees.

Studies included in this review showed mixed results for attempts to predict go-no go military training criteria such as pass/fail and voluntary withdrawal from a training course (unfavorable discharge during military training is included in the adjustment criterion category). The two studies of go-no go military training criteria that are included in Table 4 were both reported by Gordon (1978) and involved the GPPI as the predictor instrument. In one study, two of four GPPI scales showed statistically significant biserial correlations with a pass/fail outcome among 396 Navy sonarman trainees. These two scales were Restraint ($r_{bis}$=.27), classified in the Dependability category, and Emotional Stability ($r_{bis}$=.22), classified with the Adjustment scales. In the second GPPI study, eight scales were correlated with pass/fail among 185 Navy underwater demolition trainees. Four of the eight scales produced statistically significant biserial validity coefficients, the two highest being .28 for Personal Relations, an Agreeableness measure, and .22 for Emotional Stability. Both of these studies with the GPPI used predictive validation designs.

Among studies of go-no go military training criteria that cannot be included in the summary table, Webster, Booth, Graham, and Alf (1978) included the CPS among intended predictors of completion versus noncompletion of the 14-week Navy Hospital Corps training school for 600 males and 600 females. One of the eight content scales of the CPS (Activity) added significantly to the other predictors in multiple correlations for males and for the total sample, but none did so for females alone. Correlations for individual scales were not reported.

Griffin and Mosko (1977) reviewed Naval aviation attrition research conducted from 1950 through 1976. Consistent with the negative results summarized earlier by Ellis and Conrad (1948) in predicting successful completion of Air Force pilot training, these authors concluded that, although numerous temperament scales have been studied as predictors of Naval aviation attrition, the results were generally unsuccessful. This conclusion was further supported by a later study by Griffin and Hopson (1978) in which increments in multiple correlations produced by various OPI scales in predicting 2-year academic and non-academic attrition among Navy and Marine aviation trainees failed to hold up under cross-validation. However, even efforts to predict pilot attrition from temperament variables have not been uniformly unsuccessful. For example, Voas (1957) reported that scales of the MMPI and GZTS demonstrated significant validity for all types of attrition among almost 2,000 Navy aviation trainees during the first 16 weeks. According to Voas, the strongest relationships were produced by "anxiety" (i.e., Adjustment) scales, although actual validity coefficients were not given.

Civilian. Ghiselli (1973) included temperament measures in his large-scale review of the validity of psychological tests in civilian occupational settings during the period from 1920-1971. Despite drawing on both published and private sources, Ghiselli was able to report validity evidence for temperament measures against training criteria for only one of eight broad occupational categories. This was an average validity of -.11 for training in protective occupations, based on fewer than 500 total cases.

The present review met with little more success in locating evidence concerning the validity of temperament scales as predictors of civilian job training criteria. The median correlations shown in Table 4 for training criteria in civilian settings all are based on fewer than 10 coefficients. Despite the general paucity of information, it is worth noting that police academy performance has been successfully predicted using both the CPI (Hogan, 1971) and the MMPI (Bernstein, Schoenfeld, & Costello, 1982), with the highest individual scale validity coefficients reaching about .30 in both of these studies. Also, Tseng (1970) found a median correlation of .27 between the Rotter I-E scale and instructor ratings on six work proficiency dimensions in a sample of 140 vocational rehabilitation trainees.

Summary. Since so few studies relating temperament scales to civilian training criteria are available, the median validity coefficients shown in Table 4 for training criteria in military and civilian studies combined largely reflect the results for military studies alone. In many cases, attempts to predict military training criteria from temperament scales have been fairly successful, with the highest individual scale validity coefficients generally ranging between .20 and .30 in these studies. Adjustment scales have shown the most general evidence of validity across different kinds of military training criteria. A smaller body of evidence also supports the validity of Achievement scales as predictors of performance in the classroom aspects of military training.

## Job Proficiency Criteria

Military. The 1948 review by Ellis and Conrad, described in the section on training criteria, included nine results from studies of the prediction of military job proficiency from temperament scales. Only one of these nine results was statistically significant. The results of the present review for job proficiency criteria are summarized in Table 5. Criteria used in these studies were either ratings, rankings, or nominations by supervisors or peers or archival production criteria, which include objective indexes of production and administrative actions such as promotions. The literature comprising military studies of job proficiency criteria is in many ways similar to the literature comprising military studies of training criteria. First, the number of available studies is not large. Second, among the predictor categories with more than five coefficients, the Adjustment and Unclassified Military scales have the highest median validity values, .18 for both. Finally, the best single-scale predictors in individual studies have generally shown validity coefficients in the .20-.30 range.

Military researchers have naturally been interested in the prediction of combat proficiency. The most important study of combat performance to date, called Fighter I, was published by Egbert et al. in 1958. These researchers gathered peer nomination and/or individual interview information on more than 1,000 soldiers under combat conditions in Korea in 1953. From this group, 310 men were selected for psychological evaluation. These 310, consisting of 166 good combat performers, called "fighters," and 144 poor combat performers, called "non-fighters," were felt by Egbert and associates to represent the upper and lower 15-20% of the distribution of combat proficiency. The men completed a 40-hour battery of psychological tests over a 5-day period. Since predictor information was gathered after

Table 5

Summary of Criterion-Related Validity Studies That Used Temperament Predictors and Job Proficiency Criteria

Summary Information for Job Proficiency Criteria

| Temperament Predictor Category | Military Studies |  |  |  |  | Civilian Studies |  |  |  |  | Combined Military & Civilian Studies |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r |
| Potency | 5 | 269 | 15 | 19 | .08 | 18 | 77 | 19 | 46 | .07 | 23 | 94 | 22 | 65 | .07 |
| Adjustment | 5 | 269 | 17 | 19 | .18 | 18 | 86 | 19 | 46 | .08 | 23 | 95 | 22 | 65 | .11 |
| Agreeableness | 4 | 200 | 5 | 6 | .00 | 10 | 104 | 9 | 16 | .05 | 14 | 104 | 12 | 22 | .03 |
| Dependability | 5 | 269 | 11 | 14 | .06 | 22 | 96 | 12 | 35 | .18 | 27 | 97 | 15 | 49 | .11 |
| Intellectance | 4 | 200 | 5 | 6 | .00 | 9 | 107 | 4 | 10 | .02 | 13 | 107 | 7 | 16 | .01 |
| Affiliation | 1 | 310 | 2 | 2 | -.02 | 3 | 101 | 4 | 4 | .03 | 4 | 206 | 4 | 6 | -.02 |
| Achievement | 2 | 290 | 1 | 2 | .15 | 2 | 46 | 1 | 2 | .32 | 4 | 159 | 1 | 4 | .24 |
| Masculinity | 1 | 310 | 3 | 3 | .23 | 6 | 66 | 4 | 7 | .03 | 7 | 69 | 4 | 10 | .10 |
| Locus of Control | 1 | 207 | 1 | 1 | .19 | 4 | 101 | 1 | 6 | .31 | 5 | 102 | 1 | 7 | .25 |
| Unclassified Military Scales | 1 | 212 | 9 | 25 | .18 |  |  |  |  |  | 1 | 212 | 9 | 25 | .18 |

the criterion classification had been performed, the study was concurrent in design.

The major self-report temperament measures included in the predictor battery were the MMPI, CPI, and 16PF. The two groups of scales showing the largest differences between the fighter and non-fighter groups were the Adjustment and Masculinity/Femininity scales. Converting the *t*-values reported by Egbert et al. to point-biserial correlations, the 13 Adjustment scales included on the standard profiles of the three inventories produced a median validity coefficient of .20, with 10 of the 13 coefficients between .17 and .25. Fighters also scored in the more masculine direction than did non-fighters on the profile Femininity scales of the CPI ($r_{pb}$=.28), 16PF ($r_{pb}$=-.23), and MMPI ($r_{pb}$=-.20). These results are reflected in the median validity coefficients shown in Table 5 for the Adjustment and Masculinity categories.

Comparisons were also reported for a number of MMPI scales that are not included on the standard inventory profile. Among these, the highest validity coefficients were for scales called Ego Strength ($r_{pb}$= .27), a measure of Adjustment, and Femininity ($r_{pb}$= -.31). Egbert and associates also used analyses of the responses of fighters and non-fighters to the items of the CPI and MMPI to construct an empirical scale from each inventory to differentiate the two groups. The cross-validated *t*-value for the 56-item empirical CPI scale translates to a point-biserial correlation of .40, with a corresponding value of .34 for the 64-item empirical MMPI scale. These latter values in particular point to the potential usefulness of temperament measures in predicting combat proficiency.

The Fighter I study provided the impetus for the development of the Army Self-Description Blank (ASDB). In various research efforts, scales were developed, usually through empirical item analyses, for infantry and for other occupational specialties. As described in the section on training criteria, six scales of the ASDB produced modest predictive validity coefficients (.15-.20) against peer ratings of combat potential among Army field artillery trainees (Adjutant General's Office, 1957). However, these were ratings of combat potential rather than actual combat performance, the criterion used in the Fighter I study.

Johnson and Kotula (1958) used the ASDB to develop empirical and rational scales separately for each of three Army jobs: cooks, clerks, and mechanics. The efforts of these researchers to develop scales with differential validity for the three different job types were largely unsuccessful, in that the eight validity coefficients for scales for targeted occupations had a median of .20, while the 17 validity coefficients for scales for untargeted occupations had a median of .16. These 25 validity coefficients constitute the results shown in Table 5 for Unclassified Military scales.

The overall pattern of validity for the temperament keys in combination with their low correlations with Army aptitude tests led Johnson and Kotula to conclude that temperament measurement could add to the classification efficiency of aptitude tests. Regarding the lack of differential validity for the various keys, these authors speculated that "a 'general job adjustment' factor may be operative to some extent in a military setting where emphasis is placed on military as well as on occupational duties" (p. 10).

In other words, performance in all military jobs probably reflects to some extent a common factor of adaptation to the military lifestyle. This general military adaptation factor may be more predictable from temperament scales than are the more technical aspects of proficiency in individual military jobs. If the temperament characteristics required for successful adaptation to military life are best captured by scales in the Adjustment category of the predictor taxonomy, this could explain the general pattern of moderate validity for these scales as predictors of both military training and military job proficiency criteria.

The summary values in Table 5 for military job proficiency criteria include the results of five additional studies. As part of a study of the prediction of AWOL, Drucker and Schwartz (1973) also gathered supervisory ratings of both military skills and leadership potential for 269 Army enlistees. All four content scales of the CPI that were included as predictors produced statistically significant predictive validity coefficients against the military skills ratings: .31 for Responsibility (Adjustment category), .25 for Dominance (Potency category), .17 for Achievement via Independence (Achievement category), and .16 for Socialization (Dependability category). The leadership potential ratings were used to divide the sample into four groups. One-way analyses of variance showed that these groups differed significantly in their mean scores on all four CPI content scales.

Gordon (1978) reported statistically significant validity coefficients for two of the eight scales of the GPPI, Ascendance (*r* =.29) and Vigor (*r*=.29), in predicting supervisory ratings of 72 Army Special Forces officers at a 9-month follow-up. These two scales are both included in the Potency category of the predictor taxonomy. In general, it seems plausible that Potency scales would be more highly related to the job proficiency of military officers than of enlisted personnel. However, Shenk, Watson, and Hazel (1973) found effectiveness ratings of nearly 6,000 Air Force officers to be virtually unrelated to scales constructed by Norman to measure each of the five higher order temperament factors that he has proposed, including one aligned with Potency (see Table 1).

Gordon (1978) reported that the scores of Navy company commanders in charge of basic training on two of the eight GPPI scales were significantly correlated with an index of the performance of their trainees. These two scales were Restraint (*r*=.32) and Cautiousness (*r*=.28), both in the Dependability category. In the one other study included in Table 5, Broedling (1975) found a correlation of .19 between the Rotter Locus of Control scale and supervisory ratings of 207 Naval personnel, 80 officers and 127 enlisted.

Finally, in an untabled study spanning the job proficiency and adjustment criterion categories used in this review, Hoiberg and Pugh (1978) formed a composite 2-year Naval effectiveness criterion from records of promotions and demotions, type of discharge, and delinquency. Subjects were nearly 8,000 enlisted men and women in seven occupations who were administered the CPS during their last week of training school. Correlations with the 2-year effectiveness criterion were reported for three of the eight CPS content scales, as well as other predictors, for 3,959 of these subjects. The two CPS scales in the Dependability category, Social Conformity ($r$=.21) and Orderliness ($r$=.16), along with two biodata items (education and school expulsions/suspensions), were the most valid predictors in the combined cognitive and non-cognitive predictor battery. In addition, Social Conformity scores were reported to be positively correlated with job and Navy satisfaction.

Civilian. Three major reviews of the validity of temperament scales in predicting job proficiency in civilian settings have been published since 1950. Ghiselli and Barthol (1953) summarized evidence concerning the validity of temperament scales against job proficiency criteria for eight different occupational categories from 1919 to the time of their review. Sixty percent of their material came from the published literature, while 40% was drawn from private sources. The authors did not attempt to organize the results according to temperament constructs; instead, they excluded results for "traits that appeared to have little or no importance for the job in question" (p. 18), and pooled the results not excluded on that basis. The findings reported by Ghiselli and Barthol are given in Table 6. As can be seen, the most favorable results were obtained for sales occupations, and the least favorable for supervisors and service workers.

Guion and Gottier (1965) updated the Ghiselli and Barthol (1953) review by summarizing all results of civilian studies relating temperament measures to occupational criteria that were published in the *Journal of Applied Psychology* and *Personnel Psychology* from 1952-1963. These authors did not summarize the results in any quantitative fashion, except by stating that temperament measures "have had predictive validity more often than can be accounted for simply by chance" (p. 141). Guion and Gottier concluded that, although instances of successful prediction of occupational criteria from temperament scales had been reported, no generalized principles of prediction using temperament measures could be discerned from the overall results. They went on to suggest that greater conceptual guidance in the development and selection of predictors, criteria, and research designs might enable the discovery of such principles.

Ghiselli's (1973) most recent review covered the period from 1920-1971 and was similar to the earlier review by Ghiselli and Barthol (1953) in that: (a) published findings were supplemented with a great deal of information obtained by Ghiselli from private sources; (b) results were organized by occupational categories, but not by temperament constructs; (c) only the results for measures of temperament traits that were evaluated by Ghiselli as pertinent to a particular occupational category were included in the summary for that category. The findings of Ghiselli's 1973 review for job proficiency criteria are summarized in Table 7. As can be seen, the mean validities for temperament scales arrived at by Ghiselli are quite respectable, ranging from .16 to .50 with a median value of .26.

The results of the present review for job proficiency criteria in civilian settings can be seen in Table 5. The highest median value based on an appreciable number of validity coefficients is .18 for the Dependability category (35 coefficients). There is also some suggestion of the validity of Locus of Control (median $r$=.31 based on six coefficients) and Achievement (median $r$=.32 based on only two coefficients) scales as predictors of civilian job proficiency.

Table 6

Results of Ghiselli and Barthol's (1953) Review of the Validity of Temperament Scales for Job Proficiency Criteria

| Occupation | Total Number of *r*'s | Total Number of Cases | Weighted Mean $r$[a] |
|---|---|---|---|
| General Supervisors | 8 | 518 | .14 |
| Foremen | 44 | 6433 | .18 |
| Clerks | 22 | 1069 | .25 |
| Sales Clerks | 8 | 1120 | .36 |
| Salesmen | 12 | 927 | .36 |
| Protective Workers | 5 | 536 | .24 |
| Service Workers | 6 | 385 | .16 |
| Trades and Crafts | 8 | 511 | .29 |

[a] These were computed using *z*-transformations weighted by sample size. (Adapted from Ghiselli, 1973)

Table 7

Results of Ghiselli's (1973) Review of the Validity of Temperament Scales for Job Proficiency Criteria

| Occupational Category | Total Number of Cases | Weighted Mean $r$[a] |
|---|---|---|
| Managerial | >10,000 | .21 |
| Clerical | 1000 - 4999 | .24 |
| Sales | 1000 - 4999 | .31 |
| Protective | 500 - 999 | .24 |
| Service | 100 - 499 | .16 |
| Trades and Crafts | 100 - 499 | .29 |
| Industrial | <100 | .50 |

[a] These were computed using *z*-transformations weighted by sample size. (Adapted from Ghiselli, 1973)

Overall, however, the median validity coefficients in Table 5 are quite low, far lower than those reported both by Ghiselli and Barthol (1953) and later by Ghiselli (1973). The reason for this discrepancy probably lies in Ghiselli's restriction of his summaries to results for measures that appeared relevant to predicting performance in each occupational category, as compared to the present pooling of results for all job types within each temperament construct category regardless of the apparent relevance of the construct for the individual jobs. Certainly temperament scales, as with other predictors, could be expected to display some differential validity across different occupations, a phenomenon that would not be revealed by the present method of summarization. Indeed, Guion and Gottier's (1965) rather pessimistic conclusion about the generality of predictive validity for temperament measures was based on an evaluation of overall results for all types of predictors in all types of jobs.

By drawing heavily on evidence from private as well as published sources, Ghiselli was able to gather sufficient information to enable a more fine-grained classification of results. For that reason, his 1973 review probably provides the best summary to date of the level of validity that has been previously attained by job-relevant temperament measures as predictors of job proficiency in civilian settings, although no information linking specific temperament constructs to specific occupations is provided.

The previous discussion suggests that temperament predictors are likely to achieve higher correlations with job performance when they are judiciously selected on the basis of their relevance for the job in question. Researchers who have undertaken validation efforts for temperament scales in civilian occupational settings do not appear to have generally adopted this strategy. Statements indicating that temperament predictors were selected on the basis of job analysis or some other type of evaluation of the temperament characteristics likely to lead to success in the job being studied are rarely found in the literature.

This evaluation echoes that of Guion and Gottier (1965), who stated that in the studies they reviewed "clearly thought-out relationships between predictors and criterion (are) uncommon" (p. 158). Typically, one or another of the multiscale temperament inventories has been chosen to provide the temperament predictors because it is "established" or "minimizes response sets," rather than because some or all of the scales have been hypothesized to be related to specified aspects of performance in the specific job being studied. Such "broadside" approaches (Guion & Gottier) are likely to yield a few useful validity coefficients for individual scales and many neglible ones, which, taken together, would be expected to produce an overall validity picture such as that in Table 5.

Although selecting temperament predictors on the basis of targeted job characteristics is probably the most important factor in maximizing their validity for occupational purposes, other factors may also contribute. For example, in most situations temperament characteristics are likely to be more strongly related to nontechnical than to technical aspects of job performance. Direct evidence for this contention derives from a study in which Graham and Calendo (1969) collected supervisory ratings on both types of criteria for a sample of 69 clerical workers. While only 2 of 40, or 5%, of the correlations with technical proficiency criteria were significant (*p*<.05), 20 of 64, or 31%, of the correlations with work-related personal characteristics were significant (the median of these 20 was .29), with ratings of employee "attitude" and "stability of temperament" proving most predictable. Also, Hogan, Hogan, and Busch (1984) developed a measure of interpersonal competence, the Service Orientation Index, that was found to be significantly correlated with various nontechnical job proficiency criteria among 37 nursing students in clinical training (*r*=.31), 30 nursing home employees (*r*=.42), 100 insurance clerks (*r*=.25), and 56 truck drivers (*r*=.34).

Following the preceding line of reasoning, it seems likely that temperament predictors in general would have greater validity in occupations in which temperament-related, nontechnical aspects of performance contribute heavily to overall job proficiency. Sales work is an example of an occupation in which traits of temperament seem relatively important to success. Tables 6 and 7 show that in the two Ghiselli reviews all mean validity coefficients for sales occupations were greater than .30. Also, reasonably successful prediction of job proficiency among police has been accomplished using both the CPI (Hogan, 1971; Mills & Bohannon, 1980) and the MMPI (Costello, Schoenfeld, & Kobos, 1982). As mentioned earlier, there is also some evidence for the validity of these two inventories in predicting training performance among police.

Just as temperament characteristics may be more valid for predicting performance in certain civilian occupations than in others, certain temperament characteristics may be more generally related than others to performance across many or all occupations. Achievement-related characteristics seem particularly likely to show a generalized pattern of validity. Indeed, in the two studies just mentioned that related the CPI to police performance, the CPI Achievement via Independence scale had validity coefficients of .32 (Hogan, 1981) and .31 (Mills & Bohannon, 1980), which were among the highest correlations in both of these studies.

Locus of control is a characteristic that may be a component of achievement or work motivation, in that individuals with an internal locus of control believe that they control the outcomes that they experience through their own behavior, while those with an external locus of control believe that such outcomes are largely determined by fate or chance. Thus, "internals" would be expected to work harder than "externals" because they expect hard work to lead to desired outcomes.

In fact, several studies (summarized by Spector, 1982) have shown relations between the Rotter Internal-External Locus of Control (I-E) scale and components of expectancy theories of job motivation such as that described by Vroom (1964). More directly, statistically significant correlations have been found between the I-E scale and proficiency in a diversity of jobs, including scientists and engineers (Dailey, 1979), vocational rehabilitation counselors (Majumder, MacDonald, & Greever, 1977), government managers, technical specialists, and staff personnel (Heisler, 1974), and mental hospital volunteers (Hersch & Scheibe, 1967). As shown in Table 5, the median validity of the I-E scale in six samples from these four studies is .31.

Integration and Summary. As can be seen in Table 5, the predictor categories that have proven most valid in military studies of job proficiency (e.g., Adjustment, Masculinity) are generally different from those that have been most valid in civilian research (e.g., Dependability, Locus of Control). For that reason, the median values for combined military and civilian studies are quite low overall. The moderate validity of Adjustment scales in military settings may reflect a relationship with a military adaptation factor that is common to all military jobs rather than with the technical aspects of proficiency in individual military jobs. Further, the results of the Fighter I study show that temperament scales, particularly Adjustment and Masculinity/Femininity measures, hold some promise for the prediction of combat performance.

Just as temperament predictors may be more strongly related to the military adaptation aspects than to the technical aspects of military job proficiency, some evidence from civilian settings also suggests that they may predict the nontechnical components of successful job performance better than the technical components. Temperament characteristics related to achievement or work motivation seem likely to contribute to proficiency across a wide range of jobs, and there is some evidence supporting the general validity of the Rotter I-E scale.

However, most temperament characteristics probably vary in their contributions to proficiency across different jobs. The apparent failure of most previous research to select temperament predictors on the basis of job requirements has likely contributed to the picture of low overall validity in Table 5. In the two reviews (Ghiselli, 1973; Ghiselli & Barthol, 1953) in which relationships with apparently irrelevant predictor scales were excluded, indexes of typical validity have been considerably higher. The practical implication is that temperament scales are likely to yield the greatest predictive payoff, both in civilian and military settings, when they are developed or chosen on the basis of hypothesized relationships with criterion behaviors.

## Job Involvement/Withdrawal Criteria

The criteria evaluated in this category are military reenlistment and civilian turnover, job tenure, and absenteeism. Summary statistics for studies reporting results that are amenable to tabulation are given in Table 8. Unfortunately, extremely few such studies are available.

Military. The only study relating self-report temperament scales to reenlistment that was located was the one by Shenk et al. (1973), mentioned earlier, in which scales constructed by Norman to measure higher order temperament factors were administered to nearly 6,000 Air Force officer trainees. None of the 15 correlations between these scales and active duty versus inactive status of the officers 6 years after testing was greater than .08 in magnitude.

Civilian. The two most recent reviews that included evaluations of temperament scales as predictors of job involvement/withdrawal criteria were published by Schuh (1967) and Porter and Steers (1973). Schuh reviewed 10 different studies, all published between 1951 and 1963, that used either self-report or projective temperament instruments to predict employee tenure. The variability of the results led Schuh to conclude that "some tests used in some situations yield good predictive results" (p. 145). Porter and Steers interpreted the results of the literature they

Table 8

Summary of Criterion-Related Validity Studies That Used Temperament Predictors and Job Involvement/Withdrawal Criteria

Summary Information for Job Involvement/Withdrawal Criteria

| | Military Studies | | | | | Civilian Studies | | | | | Combined Military & Civilian Studies | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Temperament Predictor Category | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r |
| Potency | 1 | 5951 | 3 | 3 | -.02 | 2 | 51 | 3 | 10 | .11 | 3 | 69 | 6 | 13 | .04 |
| Adjustment | 1 | 5951 | 3 | 3 | -.03 | 2 | 62 | 4 | 13 | .20 | 3 | 51 | 7 | 16 | .17 |
| Agreeableness | 1 | 5951 | 3 | 3 | -.03 | 1 | 107 | 1 | 2 | .03 | 2 | 107 | 4 | 5 | -.02 |
| Dependability | 1 | 5951 | 3 | 3 | -.02 | 3 | 71 | 4 | 12 | .16 | 4 | 74 | 7 | 15 | .14 |
| Intellectance | 1 | 5951 | 3 | 3 | -.07 | 2 | 51 | 2 | 6 | -.12 | 3 | 69 | 5 | 9 | -.09 |
| Affiliation | 0 | -- | -- | -- | -- | 1 | 50 | 1 | 4 | .09 | 1 | 50 | 1 | 4 | .09 |
| Achievement | 0 | -- | -- | -- | -- | 0 | -- | -- | -- | -- | 0 | -- | -- | -- | -- |
| Masculinity | 0 | -- | -- | -- | -- | 1 | 50 | 1 | 4 | .03 | 1 | 50 | 1 | 4 | .03 |
| Locus of Control | 0 | -- | -- | -- | -- | 0 | -- | -- | -- | -- | 0 | -- | -- | -- | -- |

[a] The criterion is reenlistment.

[b] The criteria are job tenure and low absenteeism.

reviewed as suggesting that organizational withdrawal is more prevalent among workers at the extreme ends, both upper and lower, of various temperament trait continua. Although Porter and Steers cited the results of six studies of turnover and absenteeism as supporting this "polar" hypothesis, none of these studies actually analyzed the relationship between temperament trait extremity and withdrawal.

Probably the most comprehensive recent study of the relationship between temperament variables and civilian job involvement/withdrawal was reported by Bernardin (1977). In this study, 12 scales of the 16PF were correlated with both job tenure and absenteeism in each of two samples ($N$=51 and $N$=48) of male telephone sales workers, resulting in four validity coefficients for each predictor scale. The same four scales were most predictive of both tenure and (low) absenteeism in both samples. Two of these were from the Adjustment category ($r$'s ranging between .17 and .29), one was from the Dependability category ($r$'s ranging between .21 and .40), and one was from the Potency category ($r$'s ranging between .16 and .23). Weighted composites of these four scales were developed in a double cross-validation design, producing validity coefficients of .38 and .31 against tenure in the cross-validation phase. Finally, analyses of the relationships between 16PF scale score extremity and tenure and absenteeism failed in every instance to support the Porter and Steers (1973) "polar" hypothesis.

The results of a couple of earlier studies are consistent with Bernardin's (1977) finding of relationships between Adjustment and Dependability scales and civilian job involvement/withdrawal. Loudermilk (1966) found the Personnel Reaction Blank (Gough, 1971), a measure of Dependability, to predict both job tenure ($r$=.15) and low absenteeism ($r_{pb}$=.29) among paper and lumber mill employees more validly than any of the other tests included in his predictor battery. (The Personnel Reaction Blank also correlated .33 with a job performance criterion in this study.) Also, Sinha (1963) reported a correlation of .39 between the Taylor Manifest Anxiety Scale (Taylor, 1953), a measure of Adjustment, and absenteeism among industrial workers in India. The statistical summary of the relationships between temperament scales and job involvement/withdrawal criteria in civilian settings can be seen in Table 8. Because most of the correlations represented in this section of the table derive from the Bernardin (1977) study, the median correlations essentially reflect his positive findings for scales in the Adjustment and Dependability categories.

Summary. Conclusions about the relationships between temperament scales and job involvement/withdrawal criteria are tenuous at best; however, there is some indication that measures of Adjustment and Dependability are the most promising.

## Adjustment Criteria

This important criterion category includes several indicators of social and psychological adjustment which are grouped into the subcategories of unfavorable discharge and delinquency in military settings, and delinquency and substance abuse in civilian settings. Summary tables are presented separately for unfavorable discharge, delinquency, and substance abuse, as well as for all adjustment criteria combined.

Military. As mentioned before, Ellis and Conrad (1948) devoted one section of their early review of the validity of temperament scales in the military to the prediction of adjustment criteria. In this section, they summarized a total of 70 results from studies using such predictor instruments as the Personal Inventory, the Cornell Selectee Index, and the MMPI. Criteria in nearly all cases consisted of some form of psychiatric judgment, either the results of psychiatric screening procedures carried out prior to enlistment, psychiatric ratings or formal diagnoses of active duty personnel, or discharge or disenrollment for psychiatric reasons. While documenting several weaknesses in the designs and statistical analyses producing these 70 results, Ellis and Conrad's overall conclusion was that "in the overwhelming majority of studies, the instrument in question proved to have some value for screening or diagnostic purposes. The near-unanimity of favorable results is impressive" (p.386). However, Ellis and Conrad did not report sufficient information on these studies to allow their inclusion in the present summary tables.

At least three later studies, summarized in Table 9, support the efficacy of temperament measures as predictors of unfavorable discharge. In one of these, Benton and Bechtoldt (1955) obtained the discharge records of a group of Navy enlistees who had been administered a 20-item version of the Personal Inventory during their first week of recruit training. Inspection of the content of this scale indicates it is similar to the content of scales in the Adjustment category. Subjects were classified into five reason-for-discharge categories: normal ($N$=500), psychiatric during training ($N$=64), psychiatric after completion of training ($N$=83), physical disability during training ($N$=50), physical disability after completion of training ($N$=27).

Comparisons of the Personal Inventory scores of the normal discharge group with those of the four unfavorable discharge groups showed that psychiatric discharge was more predictable when it occurred during recruit training ($r_{pb}$=.36) than when it occurred after the successful completion of training ($r_{pb}$=.17). Discharge for physical disability was also moderately predictable during training ($r_{pb}$=.21), but not afterwards ($r_{pb}$=.01). Analysis of variance results reported by Benton and Bechtoldt allow calculation of a single correlational index (*eta* coefficient) to characterize the overall differences among the Personal Inventory means of the five discharge groups. This value, .36, is the one used in Table 9 to represent the results of this study.

In a study reported by Carleton, Burke, Klieger, and Drucker (1957), the discharge records of a group of Army enlistees ($N$=1,536) who were discharged before serving 18 months were classified as either favorable ($N$=1,324) or unfavorable ($N$=212). Another group of enlistees who were discharged after serving 18 months or more ($N$=1,594) were also classified as favorably ($N$=1,354) or unfavorably ($N$=240) discharged. During training, all subjects had been administered the Army Personality Inventory, an instrument consisting of 300 items from the MMPI. Because the MMPI is heavily saturated with Adjustment variance (e.g., Tellegen, 1964), most of the characteristics tapped by the Army Personality Inventory likely fall into this category of the predictor taxonomy. However, because the particular items and scoring keys used by Carleton et al. were not described in the research report, their results are shown in the Unclassified Military scales category of Table 9.

Table 9

Summary of Criterion-Related Validity Studies That Used Temperament Variables to Predict Unfavorable Military Discharge (An Adjustment Criterion)

| Temperament Predictor Category | Number of Studies | Median Sample Size | Number of Different Predictors | Number of *r*'s | Median *r* |
|---|---|---|---|---|---|
| Potency | 0 | -- | -- | -- | -- |
| Adjustment | 2 | 1,996 | 2 | 2 | -.43 |
| Agreeableness | 0 | -- | -- | -- | -- |
| Dependability | 0 | -- | -- | -- | -- |
| Intellectance | 0 | -- | -- | -- | -- |
| Affiliation | 0 | -- | -- | -- | -- |
| Achievement | 0 | -- | -- | -- | -- |
| Masculinity | 0 | -- | -- | -- | -- |
| Locus of Control | 0 | -- | -- | -- | -- |
| Unclassified Military Scales | 1 | 1,565 | 16 | 20 | .22 |

These results all were expressed as cross-validated biserial correlations, whose values were .21 and .20 for two separate scoring keys, and .29 and .26 for two separate two-scale composites, in the group discharged before 18 months. In the group discharged after 18 months, four separate scoring keys produced correlations ranging between .12 and .24, with 12 different composites of from two to six predictor scales showing validity coefficients ranging between .14 and .31 (median=.23). The same methods of scoring were more valid in the earlier discharge group than in the later discharge group in all four such comparisons reported. This result, as with Benton and Bechtoldt's (1955) finding that unfavorable discharge was more predictable during recruit training than after successful completion of training, is probably due to the presence of a larger proportion of more seriously maladjusted individuals in the earlier discharge than in the later discharge groups. It is exactly these individuals who should be most readily identifiable using temperament scales.

Hoiberg, Hysham, and Berry (1973) administered the 115-item Recruit Temperament Survey (RTS), among other predictors, to a large sample of Navy enlistees beginning initial training. Scale scores of 1,643 enlistees discharged for psychiatric reasons during training were compared with those of a control group of 1,625 who successfully completed training. The RTS was by far the most valid of all predictors studied, its *t*-value converting to a point-biserial correlation of .45. In addition, an empirically derived 17-item subscale of the RTS, consisting primarily of Adjustment items, produced a point-biserial validity coefficient of .49. Although this value was derived from the same sample used to develop the RTS-17, the sample size was substantial enough that minimal validity shrinkage might be expected upon application to a new sample, and so this value is used in Table 9 to represent the results of this study. Thus, in all three studies reviewed here, temperament scales most closely aligned with the Adjustment category of the predictor taxonomy provided evidence of useful validity in predicting unfavorable discharge, especially in the earlier stages of military service.

All studies relating temperament scales to military delinquency criteria that were located used scales from either the CPI or MMPI as predictors. The results of these studies are summarized in Table 10. Drucker and Schwartz (1973) administered five CPI scales, as well as other predictors, to more than 2,000 Army enlistees beginning basic training. Information from subsequent reports allowed determination that 60 of these soldiers had been reported AWOL once or more during basic training, while 2,012 had not. A series of *t*-tests showed that the AWOL subjects scored lower than the non-AWOLS on the four content scales of the CPI that were used, translating to biserial predictive validity coefficients of -.38 for Responsibility (Adjustment category), -.33 for Achievement via Independence (Achievement category), -.30 for Socialization (Dependability category), and -.16 for Dominance (Potency category). Drucker and Schwartz were able to obtain information on AWOL during initial unit assignments for a subset of 301 members of the original sample. Of these, 32 had been reported AWOL once or more, while 269 had not. Biserial correlations with this criterion were higher than in the basic training sample: -.51 for Responsibility, -.43 for Dominance, -.35 for Socialization, and -.32 for Achievement via Independence. As previously described, these four scales were also significantly predictive of supervisory ratings of both military skills and leadership potential among the 269 non-AWOLS in initial unit assignments.

In a study using a criterion similar to AWOL, Knapp (1964) administered the CPI Socialization scale (Dependability category) to 82 Navy enlisted men working aboard the same ship and also examined their service records for evidence of previous disciplinary offenses. The 36 subjects in this sample who were classified as offenders had been primarily guilty of unauthorized absences. The lower scores of these offenders than of the nonoffenders on Socialization can be expressed as a biserial correlation of -.39.

At least three concurrent validation studies have compared temperament scale scores of military prisoners with those of control samples of nonprisoners. Clark (1952) compared the MMPI scores of each of three groups of Army prisoners (total *N*=136), differing in psychiatric diagnosis, with those reported earlier by Schmidt (1945) for a group of 98 nonprisoners. Twenty-six of 27 mean differences for content scales were significant across the three prisoner versus nonprisoner comparisons, translating to average point-biserial correlations as high as .64 for Psychopathic Deviate and .62 for Mania, both Dependability scales.

Gough and Peterson (1952) administered a 42-item version of CPI

Table 10

Summary of Criterion-Related Validity Studies That Used Temperament Variables to Predict Delinquency (An Adjustment Criterion)

| | Military Studies | | | | | Civilian Studies | | | | | Combined Military & Civilian Studies | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Temperament Predictor Category | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r |
| Potency | 1 | 1187 | 1 | 2 | -.30 | 2 | 2891 | 8 | 13 | -.26 | 3 | 2072 | 8 | 15 | -.26 |
| Adjustment | 2 | 301 | 4 | 5 | -.41 | 2 | 2891 | 7 | 13 | -.42 | 4 | 1410 | 10 | 18 | -.42 |
| Agreeableness | 0 | — | — | — | — | 1 | 747 | 1 | 1 | -.31 | 1 | 747 | 1 | 1 | -.31 |
| Dependability | 5 | 683 | 3 | 7 | -.39 | 4 | 1827 | 8 | 11 | -.44 | 9 | 1128 | 10 | 18 | -.43 |
| Intellectance | 0 | -- | -- | -- | -- | 1 | 747 | 1 | 1 | -.24 | 1 | 747 | 1 | 1 | -.24 |
| Affiliation | 0 | — | -- | — | — | 0 | -- | — | -- | — | — | — | -- | — | — |
| Achievement | 1 | 1187 | 1 | 2 | -.33 | 1 | 2959 | 1 | 2 | -.42 | 2 | 2482 | 1 | 4 | -.35 |
| Masculinity | 1 | 234 | 1 | 1 | -.42 | 1 | 2959 | 1 | 2 | .01 | 2 | 2891 | 2 | 3 | -.02 |
| Locus of Control | 0 | -- | -- | -- | -- | 0 | — | -- | -- | — | 0 | — | -- | -- | — |

Socialization (Dependability category) to 99 stockade prisoners and 1,092 newly inducted soldiers at the same Army post. The lower scale scores of the prisoners translate to a point-biserial correlation of -.32. When Datel (1962) replicated this comparison at the same Army post roughly 10 years later using 303 incoming stockade prisoners and 762 newly inducted soldiers, the separation produced by the full 54-item CPI Socialization scale was even greater ($r_{pb}$=-.58).

The summary results for the prediction of delinquency in military settings shown in Table 10 derive from the five studies just described that compared offenders defined by unauthorized absence or imprisonment with nonoffenders. Other studies have addressed a more difficult prediction problem that goes beyond the prediction of delinquency in unselected military samples, the differentiation of members of samples of military delinquents in terms of their number of offenses.

With respect to AWOL, Clark (1948) found that 55 recidivists did not differ significantly from 45 first-offense AWOLs on any of the scales of the MMPI. Gough and Peterson (1952), however, did find a modest difference ($r_{pb}$=-.17) on CPI Socialization (Dependability category) between 144 repeated offenders and 209 first offenders among Air Force stockade prisoners. The greatest success in this regard was achieved by Knapp (1963), who found CPI Socialization to correlate -.28 with number of delinquent offenses (with length of service partialed) in a sample of 92 Navy brig prisoners.

As would be expected from the results reviewed in this section, the summary values that are shown in Tables 9 and 10 for the validity of temperament scales in predicting military adjustment criteria are impressive. Although the number of correlations on which these medians are based is not large, the sample sizes involved in these results generally are. The greatest amount of evidence pertains to scales in the Adjustment and Dependability categories, both of which have median validity coefficients around -.40.

Civilian. The summary of results for civilian studies of delinquency can also be seen in Table 10. Consistent with the results of military studies, temperament measures have performed very well in predicting civilian delinquency, with most of the strongest findings being produced by Dependability (median $r$= -.44) and Adjustment (median $r$=-.42) scales.

A number of different criteria of delinquency have been studied in research with civilians. Among studies in occupational settings, Gordon (1978) reported that six of the eight GPPI scales differentiated those steel plant workers who would later have records of violation of company regulations from those who would not. Based on a total sample of 747, the highest predictive validity coefficients were achieved by the two GPPI scales in the Dependability category ($r_{pb}$=-.46 for Cautiousness and $r_{pb}$= -.44 for Responsibility) and the one in the Adjustment category ($r_{pb}$=-.36 for Emotional Stability). Jones (1980) reported correlations ranging between .32 and .42 for three scales measuring antisocial attitudes (Dependability category) with self-reported amount of employee theft in a mixed occupational sample of 39 subjects. In a study of 1,672 high school students, Rathus, Fox, and Ortins (1980) found that the MacAndrew alcoholism scale of the MMPI (MacAndrew, 1965) had a significant regression weight (no correlations were reported) in multiple regression analyses of self-reported frequency of each of eight delinquent behaviors, including theft, fighting, and vandalism, as well as in analyses of self-reported alcohol and drug use.

Gough has reported two large-scale studies of temperament scale differences between institutionalized delinquents and nondelinquents. In one study (1966), he compared the CPI scale scores of delinquent and nondelinquent males in each of two different samples. The first sample consisted of 881 delinquents and 2,146 nondelinquents, the second of 409 delinquents and 2,482 nondelinquents. In the first sample, the largest differences occurred on the Socialization (Dependability category, $r_{pb}$= -.66) and Responsibility (Adjustment category, $r_{pb}$=-.61) scales. The results in this sample were used to develop two "social maturity" composites, one of three scales and one of six scales, both of which weighted Socialization and Responsibility most heavily. In the second sample, Socialization ($r_{pb}$= -.52) and Responsibility ($r_{pb}$=-.50) were again the most highly related of the individual scales to criterion status, with the three-scale composite ($r_{pb}$=-.60) and the six-scale composite ($r_{pb}$= -.63) producing even greater separation of the delinquent and nondelinquent groups.

The second large study reported by Gough involved comparisons of delinquents and nondelinquents on the Personnel Reaction Blank (Gough, 1971), a scale modeled after CPI Socialization, among both males and females. Substantial differences were found between 505 delinquent and 1,626 nondelinquent males ($r_{pb}$=-.58) and between 114 delinquent and 1,408 nondelinquent females ($r_{pb}$=-.57).

The summary of results of civilian studies relating temperament scales to alcohol and drug abuse is given in Table 11. Scales in the Dependability category, with a median validity of -.42, have been both strongly and uniquely predictive of substance abuse.

The results summarized in Table 11 derive from two distinct bodies of research, one of addiction and one of substance use in the nonaddictive range. Studies of addiction have focused on individuals formally diagnosed as alcoholic or drug (usually narcotics) addicted. Without exception, this research has been carried out in institutional settings, either hospitals, substance abuse treatment programs, or criminal detention facilities. Probably because of its often routine use in these settings, the MMPI has been the temperament instrument used in nearly all of this research. On the other hand, studies of substance use in the nonaddictive range have related a wide variety of temperament measures to variations in the self-reported use of alcohol and illicit drugs, primarily marijuana, in nonaddicted samples. Undoubtedly because of their accessibility, high school and college student samples have predominated in this type of research. Representative results from studies of both addiction and substance use in the nonaddictive range are described below.

Three basic research designs have been used among studies of the MMPI characteristics of alcoholics and drug addicts. The most common approach has simply been the description of mean MMPI profiles and frequency of 2-point scale elevation codes (determined by the two highest scale scores on individual profiles) among substance addicted samples. Such results suggest *t*-score (mean=50, SD=10) elevations on all MMPI clinical (i.e.,

Table 11

Summary of Criterion-Related Validity Studies That Used Temperament Variables To Predict Civilian Substance Abuse (An Adjustment Criterion)

| Temperament Predictor Category | Number of Studies | Median Sample Size | Number of Different Predictors | Number of *r*'s | Median *r* |
|---|---|---|---|---|---|
| Potency | 6 | 124 | 11 | 16 | .09 |
| Adjustment | 11 | 148 | 19 | 32 | -.14 |
| Agreeableness | 2 | 120 | 3 | 4 | -.03 |
| Dependability | 10 | 171 | 11 | 22 | -.42 |
| Intellectance | 2 | 120 | 1 | 2 | .20 |
| Affiliation | 2 | 120 | 2 | 4 | -.07 |
| Achievement | 1 | 148 | 1 | 1 | .26 |
| Masculinity | 7 | 148 | 4 | 8 | -.18 |
| Locus of Control | 0 | -- | -- | -- | -- |

content) scales for substance addicts relative to the MMPI normative sample, with the Psychopathic Deviate (Dependability category) scale producing the greatest differentiation.

Representative of the results for alcoholics is the description of the mean profile of a sample of more than 1,000 male alcoholics by Hodo and Fowler (1976). In this mean profile, all of the clinical scales were elevated above 60, Psychopathic Deviate and Depression (Adjustment category) were above 70, and Psychopathic Deviate showed the greatest elevation. The description of a comparable-sized sample (*N*=871) of narcotic addicted males (Berzins, Ross, & Monroe, 1971) showed exactly the same results, except that the third highest scale in the mean profile, the Schizophrenia scale (Adjustment category), was also elevated above 70. This similarity in the MMPI characteristics of alcoholics and narcotic addicts has also been documented through studies directly comparing samples of the two types, both on the standard profile scales (Hill, Haertzen, & Davis, 1962) and on the MacAndrew alcoholism scale (Kranitz, 1972).

Although indexes of temperament scale differences between substance addicts and normals that are based on inventory norms are informative, comparisons with control groups are more appropriate (Apfeldorf, 1981). In addition, the comparison of MMPI scale means between substance addicts and appropriate control samples provides a statistical index, the *t*-ratio, that can be both tested for statistical significance and converted to a correlational index ($r_{pb}$). The more popular version of the control group design has involved contrasting alcoholics and drug addicts with other deviant groups, primarily psychiatric patients and criminals, with the goal of discovering temperament characteristics that distinguish addiction among the various forms of deviance. For example, the study by Hill et al. (1962) cited above found a dramatic similarity in the mean profiles of samples of 184 alcoholics, 192 drug addicts, and 195 criminals, thus establishing a link between the prediction of substance abuse and the prediction of delinquency. However, such comparisons do not parallel the problem of predicting substance abuse among members of a typical occupational applicant group.

The research design yielding results most relevant to the present review, that comparing substance addicts with normal controls, has been the least prevalent among MMPI studies of addiction. Most of the studies of this kind that have been carried out have been concurrent in design and suggest large MMPI differences between substance addicts and normals, particularly on the Psychopathic Deviate (Dependability category) scale (Gilbert & Lombardi, 1967; Hampton, 1953; Manson, 1949) and on the empirically derived MMPI alcoholism (unclassified) scales (Hampton, 1953; Hoyt & Sedlacek, 1958; Rich & Davis, 1969; Vega, 1971). Well over half of the differences on these scales in the studies just cited convert to point-biserial correlations of .50 or higher.

Although such results suggest considerable promise for the identification of current addiction using temperament scales, studies with predictive validation designs are necessary to provide evidence on the prediction of future addiction. In one such study, reported by Loper, Kammeier, and Hoffman (1973) and Hoffman, Loper, and Kammeier (1974), the scale scores of a sample of 32 hospitalized alcoholics who had taken the MMPI in college an average of 13 years previously were compared with those of a randomly chosen sample of 148 college classmates. Three scales differentiated the two groups at statistically significant levels, Psychopathic Deviate ($r_{pb}$=.24) and Mania ($r_{pb}$=.20), these being the two MMPI scales in the Dependability category, and the MacAndrew alcoholism scale ($r_{pb}$=.18). More accurate prediction could probably be expected over a shorter time interval.

In agreement with the particular prominence of high Psychopathic Deviate scores among MMPI characteristics of current and future substance addicts, temperament scales from the Dependability category have also proven the best predictors of substance use in the nonaddictive range. For example, Smart and Fejer (1969) found the greatest MMPI differences between 100 LSD users and 46 nonusers on Psychopathic Deviate ($r_{pb}$=.60) and Mania ($r_{pb}$=.44). Comrey and Backer (1970) found the two Dependability category scales of the CPS, Social Conformity and Orderliness, to have the highest correlations with self-reported amount of marijuana use among 209 college students. The correlations were -.54 for Social Conformity and -.30 for Orderliness. In a replication by Knecht, Cundick, Edwards, and Gunderson (1972) using 69 college students, the correlations were -.60 for Social Conformity and -.48 for Orderliness, as well as -.46 for 16PF G: Conscientious, another Dependability scale. Knecht et al. also reported a correlation of -.56 for Social Conformity in a second sample of 66 college students.

The two Dependability category scales from the CPI, Socialization and Flexibility, were the most successful of all CPI scales in differentiating among levels of self-reported marijuana use in college student samples studied by Hogan, Mankin, Conway, and Fox (1970), Weckowicz and Jannson (1973), and Green and Haymes (1973). The Hogan et al. (1970) study was the only one of the three to report correlational indexes, *eta* coefficients which were .46 for Flexibility and .39 for Socialization. Dunnette et al. (1980) found correlations of .60 for a "rebellion" scale consisting of temperament and biodata items and .29 for a temperament scale of "impulsiveness" (both Dependability scales) using amount of alcohol and drug use as a criterion in a sample of over 3,000 high school students.

As was true of addiction research, most studies relating temperament measures to substance use in the nonaddictive range have been concurrent in design. In one predictive study, Gulas and King (1976) found GPPI Responsibility, another Dependability measure, to differentiate entering college freshmen, all of whom were nonusers of drugs, in terms of whether they would be identified as heavy drug users or as continued nonusers ($r_{pb}$=-.38) three years later.

Summary. The results of studies of the prediction of the various adjustment criteria discussed here from temperament scales are combined into the single summary shown in Table 12. Comparison of the values in Table 12 with those in the earlier summary tables in this section reveals that temperament scales have been much more strongly related to adjustment criteria than to criteria in any of the other four major criterion categories.

In particular, scales from the Dependability category, with a median validity of -.43 against combined adjustment criteria in combined military and civilian research, and from the Adjustment category, with a corresponding value of -.33, have demonstrated substantial evidence of strong relationships. These summary values are based on a body of research showing consistent relationships between Dependability scales and military and civilian delinquency and civilian substance abuse, and between Adjustment scales and military delinquency and unfavorable discharge and civilian delinquency.

## Summary of Criterion-Related Validity Evidence

The construct-guided approach to prediction that is advocated here, which involves selecting temperament predictors on the basis of hypothesized relationships with criterion constructs, has apparently not been adopted in most previous research attempting to predict applied psychology criteria from temperament scales. Rather, in most studies all available temperament predictors have been correlated with all available criteria.

For that reason, the median validity coefficients shown in the present summary tables may be somewhat misleading, particularly for the training performance, job proficiency, and job involvement/withdrawal criterion categories. The median validity coefficients within each of these categories reflect a mixture of results for different jobs and different criterion measures, some of which probably are conceptually related to any particular predictor construct category and some of which probably are not.

Table 12

Summary of Criterion-Related Validity Studies That Used Temperament Variables to Predict Adjustment Criteria

| | Summary Information for Combined Adjustment Criteria | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Military Studies | | | | | Civilian Studies | | | | | Combined Military & Civilian Studies | | | | |
| Temperament Predictor Category | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r | Number of Studies | Median Sample Size | Number of Different Predictors | Number of r's | Median r |
| Potency | 1 | 1187 | 1 | 2 | -.30 | 8 | 171 | 14 | 29 | -.13 | 9 | 180 | 14 | 31 | -.17 |
| Adjustment | 4 | 724 | 6 | 7 | -.41 | 13 | 170 | 20 | 45 | -.25 | 17 | 234 | 22 | 52 | -.33 |
| Agreeableness | 0 | -- | -- | -- | -- | 3 | 171 | 4 | 5 | -.03 | 3 | 171 | 4 | 5 | -.03 |
| Dependability | 5 | 683 | 3 | 7 | -.39 | 14 | 388 | 17 | 33 | -.43 | 19 | 388 | 17 | 40 | -.43 |
| Intellectance | 0 | -- | -- | -- | -- | 3 | 171 | 2 | 3 | .18 | 3 | 171 | 2 | 3 | .18 |
| Affiliation | 0 | -- | -- | -- | -- | 2 | 120 | 2 | 4 | -.07 | 2 | 120 | 2 | 4 | -.07 |
| Achievement | 1 | 1187 | 1 | 2 | -.33 | 2 | 2891 | 1 | 3 | -.37 | 3 | 2072 | 1 | 5 | -.33 |
| Masculinity | 1 | 234 | 1 | 1 | -.42 | 8 | 168 | 4 | 10 | -.11 | 9 | 170 | 4 | 11 | -.13 |
| Locus of Control | 0 | -- | -- | -- | -- | 0 | -- | -- | -- | -- | 0 | -- | -- | -- | -- |
| Unclassified Military Scales | 1 | 1565 | 16 | 20 | .22 | | | | | | | | | | |

For example, Agreeableness scales seem more conceptually related to the interpersonal aspects of performance in social service jobs than to the technical aspects of performance in manufacturing jobs.

Thus, these median validity coefficients, which are combined into the single, overall summary shown in Table 13, provide information only about the most general relationships between temperament predictor and applied psychology criterion constructs. Although these summary values are informative, they are not intended to serve as a substitute for analyzing criterion behaviors and matching them with appropriate temperament predictor constructs in any particular applied prediction situation.

As Table 13 shows, educational criteria have in general been predicted fairly well by measures of achievement-related constructs, including locus of control, with .30 a representative value.

The number of studies relating temperament measures to training criteria is small, particularly in civilian settings. The available evidence from military studies suggests a general pattern of moderate validity for Adjustment scales (median $r$=.20) across different criteria of military training performance. Those few studies relating Achievement scales to performance in the military training classroom have found substantial correlations.

There are also relatively few available studies of the validity of temperament scales as predictors of job proficiency in military settings. However, the very important Fighter I study has demonstrated the potential usefulness of temperament measures, particularly Adjustment and Masculinity scales, as predictors of combat proficiency. Overall, the Adjustment scales (median $r$=.18) have shcwn the greatest amount of evidence of validity for military job proficiency criteria. The consistent, moderate level of validity demonstrated by Adjustment scales for both military training and job proficiency criteria is perhaps due to relationships with military adaptation factors rather than with technical proficiency factors.

Values summarizing the validity of temperament scales in predicting job proficiency criteria in civilian settings are far lower in the present review than in earlier reviews by Ghiselli and Barthol (1953) and Ghiselli (1973), with the most general evidence of validity being offered by the Dependability (median $r$=.18, based on 35 coefficients) and Locus of Control (median $r$=.31, based on six coefficients) categories. The discrepancy is probably due to the practice in the earlier reviews of including only results for temperament predictors that aopeared relevant to the job in question. Indeed, much of the research that has attempted to predict job proficiency from temperament scales has apparently been carried out in the absence of hypotheses about specific predictor-criterion relationships. More conceptually based research efforts may well improve the overall validity picture.

Very few studies of the relationships between temperament scales and job involvement/withdrawal criteria have been carried out, particularly in military settings. The scant evidence that is available from civilian studies suggests that predictors in the Adjustment (median $r$=.20) and Dependability (median $r$=.16) categories may be most generally related to job involvement/withdrawal.

Table 13

Median Criterion-Related Validity Coefficients for Temperament Predictors for All Criterion Categories

| Temperament Predictor Category | Educational Criteria | Training Criteria | Job Proficiency Criteria | Job Involvement/ Withdrawal Criteria | Adjustment Criteria |
|---|---|---|---|---|---|
| Potency | .06(42)[a] | .13(36) | .07(65) | .04(13) | -.17(31) |
| Adjustment | .14(43) | .19(28) | .11(65) | .17(16) | $\boxed{-.33}$(52) |
| Agreeableness | .03(9) | .08(5) | .03(22) | -.02(5) | -.03(5) |
| Dependability | .13(24) | .12(20) | .11(49) | .14(15) | $\boxed{-.43}$(40) |
| Intellectance | .17(6) | .19(5) | .01(16) | -.09(9) | .18(3) |
| Affiliation | -.03(5) | -- | -.02(6) | .09(4) | -.07(4) |
| Achievement | $\boxed{.30}$(8) | $\boxed{.33}$(4) | $\boxed{.24}$(4) | -- | $\boxed{-.33}$(5) |
| Masculinity | -.16(8) | .09(3) | .10(10) | .03(4) | -.13(11) |
| Locus of Control | $\boxed{.32}$(1) | $\boxed{.29}$(2) | $\boxed{.25}$(7) | -- | -- |
| Unclassified Military Scales | -- | .18(8) | .18(25) | -- | $\boxed{-.22}$(20) |

[a] The number in parentheses next to each median value is the number of correlations on which that median is based.

NOTE: Median correlations above .20 are indicated by a box.

Temperament scales have proved much stronger predictors of adjustment criteria, such as unfavorable discharge from the military, delinquency, and substance abuse, than of criteria in the other categories. Median correlations on the order of -.40 characterize the relationships of Dependability predictor scales with substance abuse and delinquency, and of Adjustment predictor scales with delinquency and unfavorable discharge. Validity coefficients of this magnitude for these criteria are likely to be found only in the non-cognitive predictor domain.

Because the constructs represented by the temperament predictor categories used in this review are essentially uncorrelated, weighted composites of valid scales from different temperament categories should prove substantially more valid than individual scales for all criteria. Also, the correlations between temperament scales and measures of cognitive and psychomotor abilities are typically low. Thus, the incremental validity provided by temperament scales is likely to be significant and of practical use in predicting criteria that are highly relevant to the Army.

## Moderator Variables

Temperament scales have been shown to provide valid predictors of a number of important criteria. However, variables that may moderate the criterion-related validity of temperament scales must also be considered in applied temperament assessment problems. Research on five potential moderators that have been studied is reviewed in this subsection. The first three that are considered comprise different kinds of test-taking attitudes and behavior: nonpurposeful responding, response sets, and faking. The remaining two classes of potential moderators that are discussed are temperament characteristics and group membership (i.e., sex and ethnicity).

### Nonpurposeful Responding

Nonpurposeful responding can be defined as answering the items of a temperament measure in a random, careless, or capricious fashion. Scale scores produced by individuals responding in these ways are necessarily invalid. Although no empirical evidence exists on the prevalence of nonpurposeful responding, it undoubtedly varies among situations and samples of respondents.

The most common approach to the detection of nonpurposeful responding has been the construction of "communality" or "infrequency" scales consisting of items likely to be answered in a particular way only by individuals responding nonpurposefully. One method of constructing such scales is through empirical analyses of the item endorsement frequencies for an entire inventory item pool; those items with the most extreme splits in endorsement frequencies are chosen to constitute the scale. Gough (1975) used this approach to develop the CPI Communality scale. A sample item from this scale is "I must admit that people sometimes disappoint me," which most people answer "True." O'Dell (1971) has applied this empirical approach to the 16PF. Jackson (1967, 1976), on the other hand, rationally constructed items that appeared a priori to be likely to produce extreme endorsement splits for inclusion in the Infrequency scales of the PRF and JPI. The items of these scales are more obvious--an example from the JPI is "My musical compositions have been played in concert halls around the world," which most people can obviously be expected to answer "False."

The efficacy of "communality" or "infrequency" scales has typically been tested through their ability to separate randomly generated protocols from those produced by actual respondents under normal testing conditions. Thus, Gough (1975) reports a study in which 30 CPI answer sheets were completed on the basis of a table of random numbers. The highest individual score in this sample on the 28-item Communality scale, a raw score of 20, was reported to be in the lowest one-half of one percent of all CPI records on file. O'Dell (1971) compared the scores derived from 59 randomly generated protocols with those of 173 psychology students on a 31-item "infrequency" scale constructed for the 16PF. Applying a cutting score of five, O'Dell found that 51 of 59 (86.4%) of the random records were correctly identified, while 10 out of 173 (5.8%) of the actual records would be seen as invalid. The efficacy of "communality" or "infrequency" scales may actually be greater than that implied by the results of these studies, since some of the already small amount of overlap between scores of the random and actual protocols may have been due to records generated by actual respondents answering nonpurposefully.

Thus, the construction of scales from items with extreme splits in endorsement frequencies has proven quite successful in the detection of randomly generated records. In actual practice, extreme scores on such scales must surely indicate either nonpurposeful responding, illiteracy, or severe deviance of cognitive or temperament characteristics. In any case, such extreme scores indicate response records of dubious validity. The inclusion of a "validity" scale of this kind in any inventory or battery of temperament measures is desirable.

## Response Sets

Response sets can be defined as tendencies to respond in characteristic, consistent ways to self-report temperament items regardless of the specific content of those items. Nearly all of the considerable attention that has been paid to this topic has focused on the two response sets called social desirability and acquiescence. Social desirability is seen as a tendency to answer temperament items in the socially approved direction regardless of whether such self-descriptions are accurate. Acquiescence, on the other hand, is viewed as a tendency to indiscriminately respond "True" to True/ False temperament items regardless of their content.

Both of these hypothesized test taking tendencies are conceptualized as individual differences variables; that is, individuals are seen as varying in their tendencies to adopt the social desirability and acquiescence response sets. Response sets pose a threat to the criterion-related validity of temperament scales to the extent that scale scores reflect individual differences in response-set tendencies rather than individual differences in substantive temperament characteristics.

Response Set Interpretations of Temperament Scale Scores. The MMPI has been the primary target of response set interpretations of temperament scale scores. Most procedures for assessing the influence of social desirability on MMPI scale scores have relied on judges' ratings of the social desirability of the content of MMPI items. Such ratings have provided the basis both for assigning social desirability values to content scales and for constructing separate social desirability scales, consisting entirely of true-keyed items rated as highly desirable and/or false-keyed items rated as highly undesirable.

The constructors of MMPI social desirability scales have tried to make them diverse in content in order to rule out explanations of scores on them in terms of substantive temperament constructs. Acquiescence on the MMPI has been studied by dividing content scales into completely true-keyed and false-keyed subscales and by constructing separate acquiescence scales, consisting entirely of true-keyed items that are diverse in content. The proportion of true-keyed items on each MMPI content scale has also been used as an index of that scale's potential for eliciting response acquiescence. These methods for assessing social desirability and acquiescence have been primarily applied to the question of whether the two large factors commonly found for the MMPI are more interpretable in terms of content or in terms of response sets.

Jackson and Messick factor analyzed sets of MMPI scales consisting of true- and false-keyed subscales and a number of social desirability scales in samples of prison inmates (Jackson & Messick, 1961) and psychiatric patients and college students (Jackson & Messick, 1962). Each sample produced two large factors that replicated well across the three samples. One of these factors showed high loadings for the social desirability scales. Further, in all three samples, the loadings of individual scales on this factor correlated over .90 with their social desirability values based on judges' ratings. Jackson and Messick interpreted this factor as reflecting social desirability. The other large factor showed positive loadings for the true-keyed subscales and negative loadings for the false-keyed subscales, and thus was interpreted as acquiescence.

Edwards, Diers, and Walker (1962) used slightly different methods to assess the contributions of response sets to the two largest MMPI factors. These authors scored the MMPI for 58 intact scales (as opposed to true- and false-keyed subscales) in a sample of college students. A correlation of correlations with a social desirability scale led to the interpretation of this factor as desirability. The loadings of the scales on the other large factor correlated .82 with their proportion of items keyed True, prompting the interpretation of this factor as acquiescence. As is discussed shortly, other interpretations of these findings with the MMPI have been offered.

Jackson (1960) used data provided in the manual for the CPI to draw inferences about the contributions of response sets to that inventory. Jackson ranked the CPI scales in seven ways: once according to their potential for eliciting social desirability (amount of change between normal and "fake good" instructions), once according to their potential for elicitng acquiescence (proportion of true-keyed items), and five times according to their correlations with each of five other scales (four from the MMPI) thought to primarily reflect either social desirability or acquiescence. The sizeable correlations among the rankings designed to reflect social desirability and among those designed to reflect acquiescence led Jackson to conclude that CPI scale score variance is primarily attributable to the two response sets. This is the same conclusion that was reached by the authors of the three MMPI studies just described, one that has also not gone uncontested.

Content Interpretations of Temperament Scale Scores. The empirical basis for response set interpretations of temperament scale scores is not as unequivocal as the results of these studies seem to suggest. For example, Tellegen (1965) has shown that many of the most dramatic findings, such as correlations above .90 between the factor loadings of scales and their correlations with social desirability indexes, are spuriously inflated by the use of inappropriate statistical procedures. More important, the various indexes of response sets that have been used in these studies, although intended to be heterogeneous in content, may also be interpretable as measures of substantive temperament constructs. If these response set measures are indeed confounded with regularities of item content, then content explanations of the results of studies using them cannot be ruled out.

For example, Block (1965) contended that true-keyed items across the various MMPI scales reflect general self-control. Accordingly, Block interpreted the major MMPI factor defined by scales or subscales dominated by true-keyed items as a broad Ego Control factor rather than as acquiescence. Block also performed an ingenious empirical study in which he randomly eliminated dominantly keyed items from MMPI scales until modified scales with equal numbers of true- and false-keyed items were obtained. The factor structure of these modified scales, which contained no possible acquiescence variance, was practically identical to the factor structure of the original scales in each of five different samples. These results argue strongly against the interpretation of one of the two major MMPI factors as primarily an acquiescence factor. Beyond that, Rorer (1965) has provided a general review of studies of acquiescence in temperament assessment and concluded that it is of negligible importance.

Block (1965) interpreted social desirability scales as diffuse measures of emotional stability and used the label Ego Resiliency to describe the major MMPI factor defined by these and other scales correlated highly with them. However, item content and social desirability cannot be untangled as neatly as through the balancing of true- and false-keyed items to separate item content from acquiescence. This is because certain temperament characteristics, particularly those related to emotional adjustment, are inherently socially desirable (e.g., cheerfulness) or socially undesirable (e.g., anxiousness), and the measurement of such characteristics with scales that are completely balanced in terms of social desirability is probably impossible.

Given the seemingly inextricable confounding of temperament item content reflecting adjustment with social desirability, evidence concerning the nontest correlates of scales constructed from such items provides the strongest basis for deciding between the two interpretations. The meaningful relationships that have been demonstrated between scales defining the two major factors of the MMPI and various nontest criteria favor a content over a response set interpretation of these factors (e.g., Block, 1965; Tellegen, 1964). More generally, results reviewed earlier in this section provide many instances of substantial criterion-related validity for scales in the Adjustment category of the predictor taxonomy. Such findings are inconsistent with a strict social desirability interpretation of these scales.

Controlling Response Sets. The view that many temperament scales primarily reflect response set variance rather than trait variance has prompted the development of scale construction procedures designed to control response sets. One such procedure involves the use of forced-choice items, which require the respondent to choose the most descriptive of two or more different statements rather than answering True or False to a single statement. Response alternatives can be matched on rated social desirability to control for this response set, while acquiescence is controlled by eliminating the True/False response options. Edwards (1959) combined this approach with purely rational scale construction in the development of the EEPS.

Jackson (1967) used different procedures for controlling response sets in the process of constructing the scales of the PRF. Scale items were selected from a large pool on the basis of a complex formula designed to simultaneously maximize internal consistency and minimize desirability, as determined through correlations with a social desirability scale. Items were selected on this basis with the additional provision that the final scales consist of equal numbers of true- and false-keyed items, to control for acquiescence. Essentially the same methods were used again later by Jackson (1976) in constructing the JPI.

Correlations with response set indexes often differ dramatically between scales such as those of the EPPS, PRF, and JPI, which were constructed using procedures designed to control response sets, and scales such as those of the MMPI and CPI, which were not. For example, correlations of the PRF Desirability scale with the scales of the CPI, as well as with the remaining PRF scales, are provided in the PRF manual (Jackson, 1967). The median correlation of the 16 CPI content scales with PRF Desirability is .53, and the highest correlation is .73. The 20 PRF content scales have a median correlation with Desirability of only .25 and a maximum of .44. However, because of the inherent social desirability of good emotional adjustment, low correlations with social desirability indexes can probably be achieved only at the cost of sacrificing measurement of this important area of the temperament construct domain.

Indeed, only one of the 50 content scales of the EPPS, PRF, and JPI (JPI Anxiety) is classified in the Adjustment category of the temperament scale taxonomy developed here. Because the MMPI alone has four scales classified in this category, and the CPI alone has six, the apparent saturation of these two inventories with "social desirability" variance can be meaningfully reinterpreted as saturation with Adjustment variance. Further, the substantial evidence for criterion-related validity of Adjustment scales indicates that procedures designed to control social desirability in scale construction necessarily result in the loss of valid variance (Hogan, 1978; Lykken, 1978), a case of "throwing out the baby with the bath water." The temperament scales that are most useful to applied psychologists are those that most validly predict criteria of interest, regardless of their correlations with response set indexes.

Summary. Temperament scale score variance that has been attributed by some to response sets has been shown to be valid for predicting important criteria. Therefore, the use of scale construction procedures that incorporate controls for response sets is unlikely to be conducive to the development of scales with maximum criterion-related validity. In all, there is no strong evidence that the criterion-related validity of temperament scales is moderated by response sets.

Faking

Faking on a temperament measure can be defined as the distortion of responses in a deliberate attempt to influence the outcome of the decision that the test is being used for. Restricting the definition of faking to decision-making contexts distinguishes it from the social desirability response set, which involves a less calculated form of distortion that is not restricted to any particular assessment context. Also, it may sometimes be the case that the particular responses generally perceived as optimal for a particular selection situation are not the most socially desirable responses. For example, Dunnette, McCartney, Carlson, and Kirchner (1962) report a study in which both sales applicants and incumbent salesmen received lower scores on scales called Cooperativeness, Conscientiousness, and Calmness when asked to fake the responses of an ideal salesman than when given instructions to respond honestly; cooperativeness, conscientiousness, and calmness are generally considered socially desirable attributes.

In most cases, however, faking can be expected to occur in the socially desirable direction. As a consequence, social desirability scales are among those temperament scales that can be most influenced by faking. Issues pertinent to faking as a potential moderator of the criterion-related validity of temperament scales are the prevalence of faking in selection contexts, the effects of faking on validity, and the detection of faking.

Prevalence of Faking on Temperament Measures in Selection Situations. Many studies have examined the effects on temperament scale scores of specific instructions to fake responses. Such studies have typically tested the same group of subjects twice, first under instructions to respond honestly, then later under instructions to fake, and found sizeable scale score differences between pretest and posttest (Schwab, 1971).

Schwab has contended that the results of many of these studies may not be conclusive due to inadequacies of design, such as the failure to include a control group given honest instructions at both pretest and posttest. Indeed, Schwab found, in a study using four scales of the GPPI and two different samples, that inclusion of a control group reduced the apparent faking effect in six of eight scale by sample comparisons. This finding suggests that the magnitude of directed faking effects may have been overstated in much previous research. Nevertheless, statistically significant faking effects occurred in about half of the analyses conducted by Schwab Thus, the bulk of evidence suggests that respondents can successfully fake scores on self-report temperament scales, at least to some extent, when specifically instructed to try to do so.

Although studies of the effects of faking instructions have demonstrated that temperament scale scores can often be faked, they do not address the question of how much faking actually occurs in selection situations. A clever study by Dunnette et al. (1962) provides evidence that actual job applicants distort their responses far less than those specifically instructed to fake. These authors administered the Dunnette Adjective Checklist to 64 sales applicants as part of the the total selection battery, then again under instructions to fake the responses of an ideal sales candidate after the battery had been completed. Scores in the "faking" condition on all seven content scales of the Checklist differed by statistically significant amounts in the direction of the common "sales type" image from scores in the "applicant" condition. Dunnette and associates concluded on the basis of scores on an empirically derived faking key for the Checklist that only 14% of these subjects had seriously faked their responses to that instrument as actual applicants.

Less direct evidence that the effects of faking are far smaller among actual job applicants than among those specifically instructed to fake derives from a study by Bartlett and Doorley (1967) and from an unpublished study cited by Guilford et al. (1976). Both of these studies compared the scores of college subjects in normal, simulated selection and in deliberate faking conditions. They found scores under simulated selection to be far more similar to scores for the normal than for the faking instructions.

The prevalence of faking in selection contexts has also been estimated through comparisons of the scale scores of job applicants (assumed motivated to fake) with those of job incumbents (assumed motivated to respond honestly). Three such comparisons were reported by Dunnette et al. (1962) in addition to the results from that study that have already been described. In one, scores on six of seven Dunnette Adjective Checklist content scales were more favorable (in terms of the common "sales type" image) for 64 sales applicants than for 62 incumbents, although these differences were not tested for statistical significance. These authors also compared scores on the Dunnette Checklist content and faking keys and on the MMPI Lie (L) and Subtle Defensiveness (K) scales separately for industrial sales applicants ($N$=63, all subsequently hired) versus incumbents ($N$=50), and for retail sales applicants ($N$=96, also later hired) versus incumbents ($N$=70). In both comparisons, applicants scored in the more favorable or faking direction than did incumbents on 8 of the 10 scales. However, only two of these differences reached statistical significance in the industrial sales comparison, while five did so in the retail sales comparison. These same samples were compared by Kirchner (1962) in terms of their EPPS scores. The retail sales comparison again showed more statistically significant differences between applicants and incumbents (4 of 15 scales) than did the industrial sales comparison (none of 15 scales).

Also studying sales jobs, Bass (1957) found that 471 applicants scored in the more socially desirable direction than 265 incumbents on all four scales of the Gordon Personal Profile. All of these differences were statistically significant in this large sample. McClelland and Rhodes (1969), on the other hand, reported no statistically significant differences at all between hospital aide applicants and new incumbents (total $N$=72) on 18 MMPI scales.

Green (1951) compared 70 applicants with 45 incumbent police patrolmen on 10 scales of two early Guilford inventories. Although 5 of 10 mean differences were statistically significant, it is not clear that two of these, lower scores for applicants on measures of sociability and reflectiveness, are consistent with what would be expected if all mean differences were due entirely to faking. This illustrates a major problem with inferring faking from temperament scale score differences between applicants and incumbents: the fact that applicants and incumbents may differ in actual temperament characteristics as well as in their motivation to fake (Schwab & Packard, 1973). For example, applicants are usually younger than incumbents (almost eight years younger on the average in the study by Bass, 1957), and, as the publication of norms for separate age groups for many temperament inventories indicates, temperament scale means often vary with age. Thus, scale score comparisons between applicants and incumbents do not yield completely clear-cut evidence on the prevalence of faking in selection situations.

A better design for assessing the extent of faking on temperament scales in selection contexts requires comparing the scores of applicants who believe that the test is being used in the selection decision with those of applicants who have just been hired. Heron (1956) and Schwab and Packard (1973) have implemented this design, with conflicting results. On an emotional maladjustment scale, Heron found significantly lower scores among 200 applicants for the job of omnibus conductor who were administered the scale as part of the medical screening procedure than among 200 applicants who completed the measure voluntarily just after they had been hired. However, Heron did not state whether or not all applicants in the "selection" condition were all subsequently hired, so it is not clear whether these two groups were or were not exactly comparable. The two groups did not differ significantly on a sociability scale. Schwab and Packard (1973) restricted their study to female applicants for a small-parts assembler job who were all actually hired. They found no statistically significant mean score differences between 26 applicants who were administered the GPPI, ostensibly as the final step of the selection process, and 28 applicants who took the inventory voluntarily just after being informed of the hiring decision. In fact, applicants in the "selection" condition scored in the more socially desirable direction than applicants in the "control" condition on only three of eight GPPI scales.

In summary, there is considerable evidence indicating that scores on many self-report temperament scales can be faked by respondents who are specifically instructed to try to do so. However, it is also clear that the large faking effects produced by instructions to fake cannot be generalized to actual selection situations. Studies comparing the scale scores of applicants with those of incumbents have sometimes found differences on some scales, with applicants obtaining the more favorable scores in most cases, but these differences may be partly or wholly a function of factors other than differential motivation to fake. Among two studies that compared the scale scores of applicants in "selection" and "control" conditions, one found some evidence for faking and the other found none. Thus, it must be concluded that the prevalence of faking on temperament measures in selection situations remains an open question. The available evidence does indicate that faking is far less prevalent than is often assumed.

Effects of Faking on Criterion-Related Validity. The prevalence of faking on temperament scales in selection situations is important only to the extent that faking affects the criterion-related validity of such scales. The usual assumption is that faking on selection tests reduces their criterion-related validity. However, changes in mean scores that may be produced by faking do not mathematically imply changes in correlation coefficients. For example, if all applicants were to alter their scores on a scale by comparable amounts in the same direction, the correlational properties of that scale would remain largely unchanged. In this sense, it is the correlation between scores in selection and nonselection contexts rather than their relative means that is of interest.

Further, Hogan (1983b) has suggested that faking may actually increase the criterion-related validity of temperament scales. According to Hogan, the ability to fake successfully reflects social competence, which itself is related to many criteria. Elsewhere, Hogan, Carpenter, Briggs, and Hansson (in press) have stated "applicants who have the good sense to present themselves in a favorable light on personality measures also seem able to do so on the job." According to this view, then, faking variance is often valid variance.

There is little, if any, empirical evidence concerning the effects on criterion-related validity of faking on temperament measures in selection situations. Dunnette et al. (1962) reported correlations with managerial ratings that were substantially lower when a sample of 45 incumbent salesmen completed the Dunnette Adjective Checklist under faking instructions than when these same salesmen responded to the Checklist under honest instructions. However, this result does not answer the question of how faking in actual selection situations affects criterion-related validity. Dunnette et al. also implied (without actually reporting correlations) that validity coefficients were higher for these sales incumbents responding honestly than for two samples of salesmen who had completed the Checklist as applicants. If it is assumed that the applicants indeed faked, this result would seem to support the contention that faking reduces validity. However, in addition to the absence of significance tests, it appears that Dunnette and associates used a concurrent validation design with incumbents but a predictive design with applicants, which confounds the interpretation of any differences in validity.

In the study by Schwab and Packard (1973) described previously, none of the eight GPPI scales showed statistically significant differences in predictive validity against voluntary turnover between applicants in the "selection" and "control" conditions, although the highest individual scale correlations were reported for the "selection" condition. This would support the position that faking has no effect on criterion-related validity except that the authors found no evidence of faking in this study.

Finally, Ruch and Ruch (1967) speculated that the ability to fake high scores on the K scale of the MMPI, often seen as a social desirability scale, is related to selling ability. Indeed, these authors found that K was significantly correlated ($r$=.39) with supervisory ratings of the effectiveness of 182 sales representatives, and that five other MMPI scales were less valid when they were corrected by K to remove social desirability than when they were not corrected. Since the subjects were presumably motivated to make a good impression when taking the MMPI (results were reported to management), Ruch and Ruch concluded that faking increased validity. However, another explanation is that high scorers on K are better adjusted rather than better fakers (see the previous discussion of the social desirability response set) and thus that the results would have been the same without the alleged motivation to fake.

Among the faking studies that have been described, the design used by

Schwab and Packard (1973) is probably the best that has been proposed so far for assessing both the prevalence of faking and, when faking has occurred, the effects of faking on criterion-related validity. In that approach, both mean scores and predictive validity coefficients were compared between separate groups of applicants and new hires. The same comparisons could also be made for a single group of subjects tested both as applicants and again soon after being hired. Correlations between the two sets of predictor scores would provide evidence on that particular aspect of faking, and the power of statistical tests for differences in means and correlations would be greater than for the independent sample design. However, since counterbalancing could not be used in the single sample design (i.e., no subjects could be tested first as employees and second as applicants), any carry-over effects from repeated testing would not be controlled.

Perhaps there is no single perfect design for testing the effects of faking on temperament scales in selection situations, but much stronger evidence than that which has been reported to date is necessary before any firm conclusions can be drawn.

Detection of Faking. Given the great concern that has been expressed about faking on temperament scales, it is not surprising that methods have been developed to attempt to detect faked records. One such method is the construction of faking detection scales through empirical analyses of differences in item endorsement frequencies between subjects answering under normal instructions and subjects instructed to fake. The Good Impression scale of the CPI is a good example of a scale of this kind. Scales developed in this way function essentially as social desirability scales.

However, the use of social desirability scales to detect faking in actual selection situations is not likely to prove efficacious when false positives (individuals identified as having faked although they actually have not) are of concern; although nearly all faked response records will show high scores on such scales, the records of many individuals who are simply well adjusted will also show high scores.

This problem of overlap in the distributions of scores of faking and well-adjusted respondents on social desirability scales is illustrated by a study by Grow, McVaugh, and Eno (1980) of faking on the MMPI. As part of this study, scores on various MMPI social desirability indexes that have been recommended as useful in the detection of faking good were compared between 50 college students responding to the MMPI anonymously and 50 college students instructed to fake good. The index that minimized false positives while correctly identifying most of the faked records (96%) still identified 38% of the anonymous records as faked. On the other hand, the best index that kept false positives at an acceptably low rate (4%) identified only 46% of the faked records. Looking at these results in another way, the various indexes accounted for only 18-38% of the variance due to faking good. These results are all the more discouraging in light of the evidence described earlier that the amount of faking in actual selection situations is far smaller than when specific faking instructions are used.

One approach to unconfounding faking and actual adjustment is through the rational construction of items reflecting behavior and attitudes that are so extremely virtuous that not even the best adjusted persons could truthfully answer in the socially desirable direction. Scales constructed from items of this kind are so-called "lie" scales, such as MMPI L, the EPQ Lie scale, and DPQ Unlikely Virtues scale. A sample item from the DPQ Unlikely Virtues scale is "I always tell the entire truth." As is true of items constituting "infrequency" scales, good "lie" scale items will seldom be answered in the keyed direction, at least in samples that are not motivated to fake. For example, the DPQ manual (Tellegen, 1982) reports raw score means on the 14-item Unlikely Virtues scale of 1.7 for females and 2.2 for males in the DPQ college normative sample.

The ability of "lie" scales to differentiate faked from honestly completed response records has not been well researched. Regardless of the outcome of such research, in actual selection situations "lie" scales are probably capable of identifying only the most grossly faked response records because of the obviousness of their items. However, since high scores on a "lie" scale are almost surely indicative of either gross faking or nonpurposeful responding, the inclusion of one such scale in a temperament inventory or battery should prove useful. Beyond that, there is probably no completely satisfactory way of distinguishing subtly faked records from those of respondents who have not distorted.

Summary. Clarification of issues regarding faking on temperament scales in actual selection situations awaits further research. Very few studies have been designed in a manner to allow definite conclusions about the prevalence of faking in selection contexts. The bulk of evidence does indicate that "real life" faking effects are far smaller than is often thought.

Another usual assumption is that faking lowers the criterion-related validity of temperament scales. However, it is mathematically possible for faking to have no moderating effect on criterion-related validity, and it has even been suggested that faking may increase validity. There is little or no direct evidence on this question. Potential research designs for evaluating both the prevalence of faking on temperament scales and its effect on criterion-related validity are discussed here.

Finally, although no completely sure-fire method for detecting faking is available, "lie" scales do provide one approach to detecting the most obvious faking. In light of the suggestion that faking may increase criterion-related validity, perhaps it would be better in any given situation not to eliminate records showing high "lie" scale scores until the effects of those records on criterion-related validity in that situation have been investigated.

## Temperament Characteristics

A highly influential paper published by Bem and Allen in 1974 has spawned a line of research into the possibility that certain temperament characteristics may moderate the validity of scores on temperament measures. These proposed moderators can be divided into two trait-specific characteristics, trait consistency and trait observability, and two general characteristics, social communication skill and introspectiveness. All studies of the moderating effects of these variables have been carried out in non-applied settings. Nevertheless, their results provide a basis for deciding whether and where to focus applied investigations.

Trait-Specific Characteristics. Trait consistency and trait observability are two potential moderators that must be assessed individually for every trait that is measured. The concept of trait consistency is that, for any given nomothetic trait, individuals differ in how consistently they manifest their particular levels of that trait from situation to situation. For example, one individual may be somewhat more talkative than average at work, at parties, and with strangers, while another may display talkativeness that is only average at work, somewhat above average at parties, and much more than average with strangers. Both individuals are somewhat above average in their overall levels of talkativeness, but the first is consistently so while the second is not. A moderating effect for trait consistency is based on the idea that the behavior of individuals who are more consistent in their manifestations of a trait should be more predictable, from measures of that trait, than the behavior of less consistent individuals.

It has also been proposed that, for any given trait, individuals differ in how observable their particular levels of that trait are to others. For example, one person who is very self-confident may communicate that strongly to others, while another who is equally self-confident may not act that way quite so outwardly. Thus, ratings by others on a particular trait should be more predictable for individuals whose levels on that trait are highly observable than for those whose levels are less observable. However, theoretically this moderating effect should apply only to the prediction of ratings by others.

Trait consistency was the moderator investigated in the pioneering study by Bem and Allen (1974). These authors proposed two methods of measuring consistency for any given trait. The first method involved simple single-item self-ratings of consistency on that trait, using questions such as "How much do you vary from one situation to another in how friendly and outgoing you are?" The second method was a statistical index of the consistency of individuals' responses to the items of a self-report measure of that trait. This "variance index" reflected primarily the variance of the individual's responses to scale items, which were made using a rating scale format. Bem and Allen divided their sample of 64 college students into high-consistent and low-consistent halves, once on the basis of the single-item consistency ratings, and once on the basis of the "variance index." In keeping with their hypothesis about a moderator effect, correlations between various measures of a trait, including self-report, ratings by others, and behavioral observations, were significantly higher in the high-consistency than in the low-consistency subsamples.

Despite the rather striking findings reported by Bem and Allen (1974), Tellegen, Kamp, and Watson (1982) have questioned the adequacy of their methods for unequivocally demonstrating moderator effects. These authors argued that moderator effects must be assessed after the unmoderated, linear contributions of all variables to prediction of the dependent variable have been accounted for, and they described a multiple regression approach to performing such analyses. Applying this approach to hypothetical examples and computer-simulated data, Tellegen et al. showed that Bem and Allen's method of dividing a sample on the basis of scores on a proposed moderator both fails to exhaust the unmoderated, linear sources of prediction, and can lead to apparent moderator effects that merely reflect statistical artifacts. They also demonstrated the advantages of fully exploiting unmoderated, linear relationships through a reanalysis of Bem and Allen's published data. In this reanalysis, correlations in the entire sample based strictly on linear relations were nearly as high as those reported by Bem and Allen for their high-consistency subsample alone.

Kenrick and Stringfield (1980) studied the moderating effects of both trait consistency and trait observability, using single-item ratings as a measure. College subjects and their parents and peers all rated the subjects according to their level, consistency, and observability on each of the 16 traits measured by Cattell's 16PF. Kenrick and Stringfield divided the subjects into high and low consistency and high and low observability subgroups for each trait. In various analyses, correlations between trait ratings averaged from modestly to substantially higher for high-consistent than for low-consistent subjects. Some evidence was also found for a moderating effect for trait observability.

The methods used by Kenrick and Stringfield (1980) in their apparent demonstration of moderator effects have also been questioned. In a critique of this study, Rushton, Jackson, and Paunonen (1981) provided data to support their contention that subjects rating themselves as highly consistent on a particular trait are more likely to be those at the high and low extremes of the trait distribution than are those who rate themselves as less consistent.

Thus, for example, Kenrick and Stringfield's high-consistency group for the trait of friendliness was likely composed of the more extremely friendly and extremely unfriendly members of their sample, while those more moderate on the friendliness dimension were likely primarily represented in the low-consistency group. As a consequence, Kenrick and Stringfield's correlations between self-ratings, parent ratings, and peer ratings on each trait may have been artifactually inflated for high-consistency subjects by an "expansion of range" and artifactually attenuated for low consistency subjects by a "restriction in range."

Bem and Allen (1974) circumvented this problem by matching subjects in their high- and low-consistency groups on the basis of their self-reported trait levels. Use of the multiple regression approach described by Tellegen et al. (1982) would also have ruled out this source of ambiguity. Because Kenrick and Stringfield did not use either of these procedures, the greater predictability they found for subjects rating themselves as more consistent may have been due to a confounding of a trait consistency and trait extremity rather than to a true moderator effect for consistency. Likewise, their measures of trait observability may have been confounded with trait extremity.

General Characteristics. Two general temperament characteristics, social communication skill and introspectiveness, have been proposed as moderators of the validity of self-report measures of all temperament traits. Social communication skill is presumed to be associated with how well individuals communicate their assessments of their own trait levels--that is, their self-images (Hogan et al., in press)--to others. Thus, correlations between self-reports and ratings by others are expected to increase with increasing levels of social communication skill for all traits. This seems to be a generalization of the proposed moderating effect for trait-specific observability for a global variable reflecting individual differences in observability across all traits. As such, the moderating effect for social communication skill should also apply only to the prediction of ratings by others (cf. Hogan et al.). More introspective individuals, on the other hand, are seen as actually knowing themselves better and, therefore, as providing more accurate self-reports. Thus, any moderating effect of introspectiveness should hold for correlations between self-report and all types of criteria, not just ratings by others.

Peer ratings were the criterion used by Cheek (1982) in an extensive investigation of potential moderators of the validity of self-report trait scales. Cheek evaluated a total of 11 potential moderators, including various measures of social communication skill and introspectiveness, as well as the trait-specific self-ratings of consistency and observability used by Kenrick and Stringfield (1980). The sample consisted of 85 college males. For each of the 11 moderators, subjects were divided at the median into high and low scorers, and correlations between self-report and peer ratings were averaged across four separate traits for each subgroup. Although most of the potential moderators did operate in the predicted direction, none produced statistically significant differences in average correlations between high and low subgroups.

Implications for Applied Research. Although the results of non-applied research into the potential moderating effects of certain temperament characteristics are suggestive, moderator effects that are strong enough to be practically useful have yet to be unequivocally demonstrated for any of these variables. Apparent demonstrations of moderator effects for the trait-specific characteristics of consistency and observability are ambiguous because of methodological problems. A practical liability of these trait-specific moderators is the necessity of assessing them separately for each trait that is measured. The use of single-item self-ratings to do this should not prove too cumbersome, but the reliability of these scores may be questionable. Also, the proposed moderator effects for trait observability and for social communication skill are hypothesized to operate only in applied situations in which the criteria involve ratings by others, (e.g., trait ratings in performance appraisal).

The conceptual basis underlying the proposed moderator effect for introspectiveness probably makes it the most promising candidate for applied research. Because this variable is designed to reflect the actual accuracy of self-report temperament scale scores, its moderating effect should theoretically apply to the relations between self-report and all types of criteria. There is indeed some evidence supporting introspectiveness as a moderator of the correlations between self-report scales and behavior in laboratory situations (cf. Cheek, 1982).

If such a moderating effect could be empirically verified in an applied context, then measures of introspectiveness could be used in the same manner that traditional "validity" scales are now used in both research and applied settings to eliminate self-report records of dubious validity. Alternatively, different prediction equations might be developed for more and less introspective respondents. However, to avoid ambiguities in future assessments of moderator effects for introspectiveness and other variables, the multiple regression approach described by Tellegen et al. (1982) should be used.

Group Membership and Fairness

The possible moderating effect of group membership variables like sex and ethnicity on criterion-related validity--the question of differential validity--is an issue that has been raised in the context of the fairness of selection devices. Another kind of group difference that has implications for fairness, though not a moderator variable issue, is represented by mean differences between groups in predictor scale scores. Such mean differences can result in adverse impact.

As Linn (1978) has pointed out, questions of fairness in selection are most directly addressed through between-group comparisons of prediction systems, that is, regression slopes and intercepts and errors of prediction. Unfortunately, such information has generally not been reported in the available group comparison studies involving temperament predictors. However, the occurrence of mean differences on predictors or differential validity can at least serve as an indicator that questions of fairness may require attention. In this subsection, discussions of the mean differences between groups on temperament scales are followed by a review of the limited evidence on the differential validity of temperament scales.

Mean Score Differences Between Males and Females. Males and females unquestionably differ by practically significant amounts in their raw scores on many, if not most, self-report temperament scales (cf. Maccoby & Jacklin, 1974). For that reason, virtually every published temperament inventory provides separate norms for males and females. In situations where the selection of equal ratios of males and females is desired, separate temperament scale norms will almost certainly be required.

Mean Score Differences Between Blacks and Whites. Many studies have reported temperament scale score differences between blacks and whites. One difficulty in evaluating these results is the common reliance on statistical significance tests as the method for assessing black-white differences. Although these tests speak to the issue of statistical reliability, they do not speak directly to the issue of magnitude or practical significance; for example, large samples may show statistically significant differences that are of little practical importance.

Once the statistical significance of a group difference has been established, indexes such as the amount of overlap of the distributions, the percentage of variance accounted for by group membership, or the number of standard deviation units separating the group means aid in evaluation of the practical significance of the difference. Under most circumstances, these three indices will be highly correlated. Therefore, whenever possible, statistically significant group differences reported in the literature are described here in standard deviation units; unfortunately, many published sources do not provide the information necessary to make such calculations. Results of black-white comparisons on several individual temperament inventories are described first, followed by an integration of these findings in terms of temperament constructs.

Gynther has twice reviewed black-white differences on temperament measures, restricting his 1972 review to the MMPI, and expanding his coverage to all temperament measures in 1979. In the first review, Gynther (1972) summarized information from some 20 sources of MMPI data. Although there was some variability in results from study to study, Gynther concluded that blacks have been found to score consistently higher than whites on three MMPI scales: the Fake Bad (F) "validity" scale, the Schizophrenia scale, and the Mania scale. These differences were said to hold for both normal and institutionalized groups and to persist when demographic variables such as socioeconomic status (SES) were controlled. Gynther, however, did not discuss the magnitudes of these differences.

Gynther (1979) interpreted the results of studies using normal samples published between 1971 and 1978 as supporting his earlier conclusions about black-white MMPI differences. Because most studies comparing the MMPIs of normal groups of blacks and whites have used student samples, the study by King, Carroll, and Fuller (1977) using adult full-time employees is noteworthy. These consisted of 56 black and 56 white employees of a large chemical company, matched on a number of demographic variables. Among the standard MMPI profile scales, blacks scored significantly higher than whites on Mania (.6 SD), while whites scored significantly higher on Paranoia (.5 SD). Whites also scored higher on non-profile scales called Dominance (.7 SD) and Ego Strength (.3 SD). Butcher, Ball, and Ray (1964) reported similar differences among both male and female college students, both when socioeconomic status was controlled for and when it was not.

At least three studies have compared the CPI scale scores of black and white college students. Farr, O'Leary, Pfeiffer, et al. (1971) tested 79 black and 193 white university freshmen. Whites scored significantly higher on 8 of the 18 CPI scales: Capacity for Status (.4 SD), Social Presence (.4 SD), Sense of Well-Being (.3 SD), Tolerance (.4 SD), Communality (.3 SD), Intellectual Efficiency (.3 SD), Flexibility (.3 SD), and Achievement via Independence (statistics not given). As can be seen, all of these differences were smaller than one-half of a standard deviation in magnitude.

Seven of these eight differences were also found by Cross, Barclay, and Burger (1978) in a sample of 772 black and white freshmen and sophomores at a community college. Although means and standard deviations were not given, whites were reported to score significantly higher on seven CPI scales, all of those listed above except Capacity for Status. Blacks scored significantly higher than whites on Good Impression. Despite the many black-white differences, no significant differences were found for different socioeconomic groups, indicating that the two racial groups were probably quite comparable on this variable.

Brown (1974) also found significantly higher CPI scores for whites on Tolerance, Achievement via Independence, and Flexibility among black and white students at a predominantly white university. However, the blacks at this university were found to score more like their white counterparts than like a sample of blacks attending a predominantly black college. Brown's findings with the CPI are consistent with Gynther's (1972) conclusion that black-white differences on the MMPI are affected by demographic variables, but that certain differences remain even after these variables are controlled.

Jones (1978) studied item differences rather than scale score differences between blacks and whites, again with college students, using subsets of the MMPI and CPI item pools. A group of 361 items was administered to 97 black male and female and 129 white male and female junior college students, with blacks and whites matched on SES. The 179 most discriminating items were cluster analyzed, and 10 clusters were interpreted. According to these interpretations, blacks scored higher on social dominance and poise, religious fundamentalism, orderliness, self-criticism, psychological toughness, cynicism and power orientation, and conformity. Whites scored higher on risk taking, psychological vulnerability, and unconventional morality.

Cameron (1971) and Lowe and Hildman (1972) compared whites and blacks on the Eysenck Personality Inventory (EPI). Cameron randomly sampled some 260 white and 150 black urban adults using census information. SES was not controlled. Comparisons of EPI means showed whites scoring significantly higher on Neuroticism (.3 SD), but no significant differences on the Extraversion and Lie scales. There were also no significant differences on a shortened form of the MMPI Ego Strength scale. Cameron concluded that "In sum, both populations appear far more similar psychologically than different" (p. 74). Lowe and Hildman tested over more than 1,600 college students with the EPI. Whites scored significantly higher than blacks on Extraversion, regardless of sex. Cameron's finding of significantly higher Neuroticism scores for whites than blacks held for males only in the Lowe and Hildman sample.

Single studies comparing scores of blacks and whites on the CPS and GZTS have been reported. In a military study, Booth and Berry (1978) administered the CPS to 1,091 black, 192 Hispanic, 186 Asian, and 1,785 white Navy enlistees in two training schools. Although practically all of the differences between blacks and whites on the 10 CPS scales were statistically significant in these large samples, only those differences of at least .3 standard deviation unit in magnitude are listed here (comparisons of whites with Hispanics and Asians are discussed later). Whites scored higher on Trust (.6 SD) and Social Conformity (.5 SD), while blacks scored higher on Orderliness (.4 SD).

Finally, Guilford et al. (1976) reported a study in which black-white differences on the 10 scales of the GZTS were evaluated separately for male and female high school students. The results of the male and female comparisons were quite similar, with blacks of both sexes scoring significantly higher than whites on Ascendance, Sociability, and Emotional Stability, and whites of both sexes scoring significantly higher than blacks on Personal Relations.

From the many differences between blacks and whites that have been reported for individual temperament inventories, can any generalizations be drawn about temperament traits or constructs on which blacks and whites appear to differ, regardless of the particular measure used? Classification of the results of the studies reviewed here according to the taxonomy of temperament scales proposed earlier in this section provides one basis for evaluating this question. Consideration must be restricted to the higher order Potency, Adjustment, Dependability, and Agreeableness categories of the taxonomy because the Intellectance and Affiliation categories are not well represented by scales from the five inventories (MMPI, CPI, EPI, CPS, GZTS) for which results have been described.

Findings for scales in the Potency category are inconsistent. Higher scores reported for whites on the MMPI Dominance scale, on CPI Social Presence (two of three studies) and Capacity for Status (one of three studies), and EPI Extraversion (one of two studies) are countered by higher scores reported for blacks on Jones' (1978) MMPI/CPI social dominance and poise cluster and on GZTS Ascendance and Sociability. Further, the majority of Potency category scales used in the studies described here produced no black-white differences.

Inconsistent and contradictory findings also characterize the results for Adjustment scales. Among indications that blacks score in the direction of poorer adjustment are higher scores on MMPI Schizophrenia and lower scores on MMPI Ego Strength (one of two studies). The CPI studies using college samples have found lower scores for blacks on Tolerance (all three studies), Sense of Well-Being (two of three studies), and Intellectual Efficiency (two of three studies). Also, blacks scored higher on Jones' (1978) self-criticism cluster, the only one of Jones' clusters clearly related to the Adjustment category. On the other hand, blacks have been reported to score lower on EPI Neuroticism (both males and females in one study, males only in another) and higher on GZTS Emotional Stability. Again, many Adjustment scales that have been studied have produced no black-white differences at all.

Contradictory findings seem to be the rule among scales in the Dependability category. There is quite a bit of evidence that blacks score higher than whites on MMPI Mania (Gynther, 1972), and lower than whites on CPI Flexibility (all three CPI studies reviewed here). High scores on both of these scales can be interpreted as, at least in part, indicators of impulsiveness. Another contradiction is provided by the higher scores of whites on both CPS Social Conformity and Jones' (1978) unconventional morality cluster. The findings of higher scores for blacks on CPS Orderliness and Jones' orderliness cluster do agree, but overall there is no consistent pattern of black-white differences on Dependability scales from different inventories.

The only measures aligned with the Agreeableness category for which black-white differences have been reported are all measures of cynicism or distrust. Here the results have consistently shown blacks scoring in the direction of greater cynicism. The largest black-white difference on the CPS in Booth and Berry's (1978) Navy sample was on Trust, with blacks scoring lower. Blacks also scored lower than whites on GZTS Personal Relations and higher than whites on Jones' (1978) MMPI/CPI cluster called cynicism and power orientation. These findings are in complete agreement with Gynther's (1972) evaluation of the true source of black-white differences on the MMPI. Gynther concluded that analyses of differences at the item level indicate that black-white differences on the MMPI are attributable to differences in values and perceptions rather than to poorer adjustment among blacks, which is the interpretation suggested by analyses at the scale level. According to Gynther, these differences in values and perceptions reflect greater cynicism and distrust among blacks. Gynther (1979) has also drawn parallels between this cynicism factor and the frequent finding that blacks score in the more external direction than whites on the Rotter Internal-External Locus of Control (I-E) scale, thus expressing less of a belief in their personal control over outcomes.

In summary, despite evidence indicating some consistent black-white differences on scales from individual temperament inventories, the only temperament construct that has differentiated blacks and whites with any consistency across different measures is cynicism or distrust. Even here there seem to be some exceptions. Recall that King et al. (1977) found white employees to score higher than black employees on MMPI Paranoia. Butcher, et al. (1964) found this same difference among both male and female college students, both when SES was controlled for and when it was not. MMPI Paranoia, like all MMPI profile scales, is factorially complex, but these findings do seem to run counter to the notion of greater distrust among blacks. Also, although 6 of the 10 studies reviewed by Gynther (1979) found blacks to score more externally than whites on the I-E scale, the remaining four studies found no difference.

These inconsistencies highlight the overriding conclusion to be drawn from research into temperament scale differences between blacks and whites, which is that such differences are typically quite variable from sample to sample. As Gynther (1979) has stated, "generalizations must be tempered by numerous qualifications" (p. 128). Thus, although there is some basis for anticipating which temperament scales are likely to produce mean score differences between blacks and whites, these expectations may well not be confirmed in any particular sample. Further, the results reviewed here suggest that even when statistically significant mean score differences between blacks and whites are found, such differences are likely to be no larger than one-half of a standard deviation in magnitude.

Mean Score Differences Between Whites and Other Ethnic Groups. Evidence concerning the mean temperament scale scores of whites compared to those of Hispanics, Asians, and Native Americans is fragmentary. The CPS scale score means reported by Booth and Berry (1978) for 1,785 white and 192 Hispanic Navy enlistees show only one scale, Trust, with a mean difference of at least .3 standard deviation. Whites scored just that amount higher than Hispanics on this scale. Gynther (1979) reviewed studies using other temperament measures and concluded that Hispanics tend to score similarly to whites.

The limited body of research comparing whites and Asians was also reviewed by Gynther, and provides some suggestion that Asians may score lower on Adjustment scales. All of these studies used college student samples. Booth and Berry (1978), on the other hand, found nearly identical means on CPS Emotional Stability for the 1,785 white and 186 Asian Navy enlistees that they tested. In this study Asians did score higher than whites on CPS Orderliness (.8 SD) and lower than whites on Masculinity (.3 SD). There were also differences on the "social desirability" (.9 SD) and "infrequency" (.6 SD) scales of the CPS, with Asians scoring higher on both scales.

Studies reviewed by Gynther (1979) which compared whites and Native Americans showed some differences in junior high and alcoholic samples. However, Native Americans scored quite comparably to whites in studies of high school and technical institute students.

Overall, there are too few studies comparing temperament scale scores of whites with those of ethnic groups other than blacks to allow any conclusions about whether, and on what scales, mean differences are likely to be found.

Differential Validity. The generally accepted definition of differential validity is that offered by Boehm (1972). According to this definition, differential validity exists when the validity coefficient for at least one of two subgroups is statistically different from zero and the validity coefficients for the two subgroups are statistically different from each other. As previously mentioned, very little evidence bearing on the differential validity of temperament scales has been reported, particularly for occupational criteria.

Schmitt, Mellon, and Bylenga (1978) evaluated the differential validity between males and females of predictors in seven general categories, one of which was temperament scales. This was accomplished through a meta-analysis of all studies reporting correlations with educational and occupational criteria separately for males and females that were published in *Educational and Psychological Measurement*, the *Journal of Applied Psychology*, and *Personnel Psychology* from 1955 up to the time of the analysis. Over 97% of the pairs of correlations across all predictor categories involved educational criteria, so conclusions from this study apply primarily to these criteria. Out of 80 pairs of correlations for males and females involving temperament predictors, eight showed differential validity, a finding not significantly different from what would be expected on the basis of chance alone. The average validity coefficients reported by Schmitt et al. were .052 for males and .049 for females, these low values almost certainly resulting from the averaging together of positive and negative correlations.

Despite an extensive search, only three studies of the differential validity of temperament scales between different ethnic groups were located, one using an educational criterion and two using job performance criteria. In the educational study, Farr, O'Leary, Pfeiffer, et al. (1971) correlated the CPI scores of 79 black and 193 white university freshmen with first-year GPA. As described earlier, whites scored significantly higher than blacks on 8 of the 18 CPI scales. Whites were also significantly higher on the criterion measure in this sample. Eleven of the 18 CPI scales showed statistically significant validity for either blacks, whites, or both. Mean correlations for these 11 scales were .23 for the total sample, .23 for blacks, and .21 for whites. None of these 11 CPI scales showed significant differences between the two groups in either correlations or regression slopes.

Farr, O'Leary, and Bartlett (1971) reported five separate studies of the differential validity of various predictor tests between black and white employees. In one of these studies, a temperament inventory, the Thurstone Temperament Schedule, was included among the predictors. Scores on this inventory were available for 107 keypunch operators, only 19 of whom were black. Eight different criterion measures were available, six based on supervisory ratings and two based on performance tests. There were no significant black-white differences on any of the seven Thurstone Temperament Schedule scales or on seven of the eight criterion measures. Only 6 of the 56 predictor-criterion relationships for the temperament scales were statistically significant among either blacks, whites, or both. These six correlations averaged .21 in magnitude in the total sample, .38

in magnitude among blacks, and .18 in magnitude among whites. The six tests for differential validity yielded one instance, a correlation that was significantly higher for blacks than for whites. However, because of the small number of blacks, the power of these tests was obviously low.

Finally, Toole, Gavin, Murdy, and Sells (1972) administered the 16PF, among other predictors, to 520 male, semi-skilled airline employees. This sample, consisting of 409 white and 111 minority (predominately black) employees, was further subdivided according to age, so that validity coefficients were reported separately for younger white (*N*=288), younger minority (*N*=75), older white (*N*=121), and older minority (*N*=36) subsamples. The criterion consisted of supervisory ratings summed across 10 dimensions. Toole et al. did not report significance tests for differences between groups in temperament predictor means or validity coefficients. However, significance tests for differential validity can be performed on the correlations that they reported. The most relevant comparisons involve those between the younger white and minority samples and between the older white and minority samples. Among the younger employees, validity coefficients that were statistically significant ($p<.05$) in either the minority sample, the white sample, or both were obtained for 6 of the 16 16PF content scales. The validity coefficients for these six scales averaged .13 in magnitude in both the younger minority and younger white sample, and differential validity is present in one of the six cases, that resulting from a higher correlation for the minority than for the white sample. Among the older employees, only one of the 16PF scales showed statistically significant validity ($p<.05$) in either the minority or white sample, and in this one case there is no differential validity.

Thus, in the three available studies in which validity coefficients for temperament scales were compared between white and minority samples, only two instances of significant differential validity occurred out of a total of 24 comparisons. Both instances involved higher correlations for the minority than for the white sample. Although these 24 comparisons are not statistically independent, the overall result is not inconsistent with what would be expected on the basis of chance alone. This was also true of the finding by Schmitt et al. (1978) of differential validity between males and females in 8 out of 80 comparisons, these involving primarily educational criteria. Although more studies are required before definite conclusions can be drawn, the evidence to date indicates that differential validity is not a characteristic of temperament predictors.

Summary. Mean score differences between males and females on temperament scales are pervasive, so much so that separate norms are required for just about any temperament scale that is to be used for selection. Many instances of mean score differences between blacks and whites have also been reported. However, black-white differences on individual temperament scales and particularly on different measures of the same temperament construct have been quite variable from sample to sample. The only temperament construct showing black-white differences with any degree of consistency across different measures is cynicism or distrust, with blacks scoring higher. There is too little evidence on mean score differences between whites and members of other ethnic groups to allow any generalizations. Evidence concerning the differential validity of temperament scales is also sparse. Preliminary indications are that instances of differential validity both between males and females and between whites and non-whites are likely to be chance occurrences. Overall, the present evidence on mean score differences and differential validity does not discourage the use of temperament scales as selection devices on these grounds.

## Summary of Moderator Variable Research

Five potential moderators of the criterion-related validity of temperament scales have been studied with varying degrees of extensiveness. Nonpurposeful responding necessarily results in invalid response records and can be detected with great accuracy through the use of "communality" or "infrequency" scales. On the other hand, the response sets of social desirability and acquiescence do not appear to pose a threat to the criterion-related validity of temperament scales.

Definite conclusions cannot be drawn about the moderating effects of faking, certain temperament characteristics, or group membership variables until more evidence is available. Faking has generated a great deal of research, but better designs are needed to assess both the prevalence of faking in actual selection situations and the effects of faking on criterion-related validity. On the other hand, the moderating effects of temperament characteristics and group membership variables are relatively new research areas. Among the various temperament characteristics that have been proposed as potential moderators, introspectiveness has the most compelling theoretical rationale, which is that individuals who know themselves better should provide more accurate self-reports.

Differential validity of temperament scales between males and females or between different ethnic groups is not supported by the small amount of available evidence. However, males and females do differ significantly in their raw scores on most temperament scales, necessitating separate norms.

Further, although blacks and whites have not shown consistent differences in their mean scores on different measures of most temperament constructs, the occurrence of differences on specific scales in specific samples has been frequent enough to indicate that ethnic group differences should be evaluated in any selection situation.

## Overall Summary

Self-report temperament assessment has typically been based on the measurement of traits, which can be defined as behavioral tendencies that are relatively consistent across situations and stable over time. The empirical validity of traditional trait conceptions, which view traits as operative across relatively broad classes of trait-relevant situations, has been challenged. The initial challenge came from the "situationist" position, which holds that behavior is explainable primarily by the situational context in which it occurs rather than by traits or individual differences. A later view, "interactionism," recognizes the importance of individual differences in behavior, but also argues that these individual differences are highly situation-specific rather than displaying trait-like consistency across a wide range of situations.

Although neither "situationism" nor "interactionism" has succeeded in replacing trait approaches to temperament assessment, their challenge has served to focus attention on conceptual and methodological inadequacies that have characterized much research in the field. Defenders of trait approaches have shown that research that is methodologically sound (e.g., criteria are reliable) and conceptually well motivated (i.e., predictors and criteria are chosen on the basis of hypothesized relationships) does support the usefulness of trait measures.

The many different methods that have been used to construct self-report temperament scales can be grouped into three major categories: purely rational, external, and internal. Purely rational scale construction relies solely on the rational judgment of the scale constructor, with the goal of maximizing content validity. Psychometric characteristics of purely rational scales vary according to the conceptions and skills of individual test constructors.

External scale construction focuses on maximizing criterion-related validity, and so in this method items are assigned to scales on the basis of their correlations with nontest criterion variables, for example, delinquency versus nondelinquency. In this approach, the use of so-called "subtle" items, items whose relationships to the dimension being measured are not intuitively obvious, is permissible. The factorial complexity of the criteria commonly used to develop external scales is often reflected in predictor scales that are heterogeneous in content and factorially complex.

The focus of internal scale construction has been on identifying and measuring functional units of temperament, or traits, and mapping out a structural representation of their interrelationships. Scale construction based on the intercorrelations of items within an item pool, as, for example, through factor analysis, is best suited to achieving this goal. Internal scales tend to be homogeneous and unidimensional.

The historical controversy concerning the relative criterion-related validity of temperament scales constructed by the three different methods has not been matched by an extensive body of empirical evidence. Nearly all studies providing direct comparisons bearing on the relative validity of purely rational, external, and internal scales have used research criteria, such as peer ratings and self-report variables. In general, internal scales have been found to be somewhat more valid than purely rational or external scales in these studies, probably because peer rating and self-report criteria map best onto the more homogeneous, unidimensional internal scales. On the other hand, it seems plausible that, at the level of individual scale correlations, the more heterogeneous, factorially complex external scales could be more valid for predicting applied psychology criteria, which often reflect combinations of a number of temperament characteristics. Indeed, the large-scale survey of the criterion-related validity literature conducted as part of this review showed a decided bias toward use of the externally constructed MMPI and CPI in applied prediction studies.

However, when an overall behavioral criterion is analyzed into its homogeneous components, as, for example, through a good job analysis, it should be possible to predict applied psychology criteria at least as well using weighted composites of homogeneous scales that target these components as using heterogeneous external scales. Such a construct-guided approach to prediction facilitates a more conceptual understanding of the relationships between predictors and criteria than does a more purely empirical approach.

A construct-guided approach to applied prediction problems would be greatly aided by a classification scheme, or taxonomy, specifying the domain of temperament constructs and identifying measures of these. The number of published and unpublished self-report temperament scales that have been developed is truly enormous, and, to further complicate matters, the comparability of different scales cannot be accurately assessed on the basis of scale names alone. For that reason, published correlations among scales were used as the data for a taxonomic analysis of temperament scales that was conducted for this review.

A framework consisting of six highly general or "higher order" dimensions provided the most compelling basis for the taxonomic analysis. These dimensions have emerged from the factor-analytic work of Tupes and Christal (1961) and Norman (1963) using peer ratings and nominations, and have been extended to the self-report domain by Goldberg (1981) and Hogan (1983a). The six dimensions are: Potency (dominance, outgoingness, energy), Adjustment (emotional stability), Agreeableness (friendliness, cooperativeness), Dependability (conscientiousness, self-control), Intellectance (intellectual interests and values), and Affiliation (desire to be with and dependence on others). The scales from a dozen major published multiscale temperament inventories were classified into these six, plus a seventh Miscellaneous category, on the basis of an analysis of 5,313 published correlations. The average between-scale correlations within the six content categories, which ranged from .33 to .46, were all substantially higher than all of the between-category averages, providing sufficient empirical justification for the proposed classification system.

The construct categories used in this taxonomic analysis were expanded to include Achievement, Masculinity, and Locus of Control, in order to provide a general summary of the criterion-related validity evidence for temperament scales in terms of predictor constructs. Military and civilian studies were grouped into five major categories of criterion variables: educational, training performance, job proficiency, job involvement/withdrawal (military reenlistment, job tenure, absenteeism), and adjustment (unfavorable military discharge, delinquency, substance abuse).

Unfortunately, the construct-guided approach to prediction that is advocated here, which involves selecting temperament predictors on the basis of hypothesized relationships with criterion constructs, has apparently not been adopted in most previous research that has attempted to predict applied psychology criteria from temperament scales. For that reason, the median validity coefficients that were calculated for each predictor construct category within each major criterion category provide a summary of only the most general relationships between temperament predictor and applied psychology criterion constructs. These values probably underestimate the levels of validity attainable when temperament predictors are chosen for their relevance to the particular job or performance situation under study.

These median validity coefficients show that educational criteria have, in general, been predicted fairly well by measures of achievement-related constructs, including locus of control, with .30 a representative value. Scales from the Adjustment category of the predictor taxonomy have demonstrated a consistent, moderate level of validity in military studies of training performance (median *r*=.20) and job proficiency (median *r*=.18). This is perhaps due to relationships with military adaptation factors rather than with technical proficiency factors. The very important Fighter I study has also demonstrated the potential usefulness of temperament measures, particularly Adjustment and Masculinity scales, as predictors of combat proficiency.

Earlier reviews of the validity of temperament scales in predicting job proficiency in civilian settings by Ghiselli and Barthol (1953) and Ghiselli (1973) included in their summaries only results for temperament predictors that appeared relevant to performance in each occupational category that was evaluated. When summarized in this way, the mean validity coefficients for temperament predictors are quite respectable, with a median value of .26 across seven broad occupational categories in the more recent review (Ghiselli, 1973).

The summary values for job proficiency criteria in civilian settings are lower in the present review than in the earlier reviews by Ghiselli, probably because results for all job types were allowed to contribute to the median correlations for each predictor construct category, regardless of the apparent match between the construct category and job requirements. The most general evidence of validity is offered by the Dependability (median *r*=.18, based on 35 coefficients) and Locus of Control (median *r*=.31, based on only six coefficients) categories. Again, however, a construct-guided approach is recommended to maximize validity in any particular attempt to predict job proficiency.

There have been relatively few studies of the relationships between temperament scales and job involvement/withdrawal criteria, particularly in military settings. The small amount of evidence that is available from civilian studies suggests that predictors in the Adjustment (median *r*=.20) and Dependability (median *r*=.16) categories may be most generally related to job involvement/withdrawal.

In an early review of military research, Ellis and Conrad (1948) concluded that temperament scales had proved much stronger predictors of adjustment criteria than of training or job proficiency criteria. The present findings support this conclusion and extend it to civilian settings as well. Median correlations on the order of -.40 characterize the relationships of Adjustment predictor scales with unfavorable military discharge and delinquency, and of Dependability predictor scales with delinquency and substance abuse. These relationships, which are likely to be found only for non-cognitive predictors, highlight the potential usefulness of temperament assessment for organizations concerned with predicting criteria reflecting psychological adjustment.

Variables that have been proposed as potential moderators of the criterion-related validity of self-report temperament scales include three kinds of test-taking attitudes and behaviors--nonpurposeful responding, response sets, and deliberate faking--as well as certain temperament characteristics and group membership variables. Nonpurposeful (e.g., random) responding necessarily results in invalid response records and can be detected with great accuracy through the use of "communality" or "infrequency" scales. Some have contended that scores on many commonly used temperament scales primarily reflect variance due to the response sets of social desirability and acquiescence rather than variance in substantive temperament characteristics. However, the bulk of evidence, particularly that supporting the criterion-related validity of temperament scales that are highly correlated with response set indexes, argues against a strict response set interpretation of scores on these scales.

Faking is distinguished from the social desirability response set by its applicability only to situations in which temperament assessment is used for decision-making purposes, such as personnel selection. Although many studies using specific instructions to fake have shown that scores on most temperament scales can be faked, the extent to which they are faked in actual selection situations remains an open question. The effects of whatever faking may actually occur on criterion-related validity also has yet to be determined. Research designs better than those that have been most commonly used to date are needed to provide more definite answers to these questions.

The potential moderating effects of temperament characteristics and group membership variables like sex and ethnicity are relatively new research areas. Temperament characteristics that have been proposed as moderators of the criterion-related validity of temperament measures can be divided into two trait-specific characteristics, trait consistency and trait observability, and two general characteristics, social communication skill and introspectiveness. Because of methodological ambiguities and weak empirical findings, it cannot be said that moderator effects strong enough to be practically useful have been unequivocally demonstrated for any of these variables. The proposed moderating effect for introspectiveness has the most compelling theoretical rationale, that individuals who know themselves better should provide more accurate self-reports, making it probably the most promising candidate of the four for applied investigations.

A moderator effect for group membership on the criterion-related validity of temperament scales--the question of differential validity--is not supported for either sex or ethnicity by the small amount of available evidence. However, males and females do differ significantly in their raw scores on most temperament scales, necessitating separate norms. Blacks and whites have not shown consistent differences in their mean scores on different measures of most temperament constructs. However, differences on specific scales in specific samples have occurred frequently enough to indicate that ethnic group differences, which could lead to adverse impact, should be evaluated in any selection situation.

## *Section 1 References*

Adjutant General's Office, Personnel Research Branch. (1957, August). *Validation of potential combat predictors: Analysis of personality measures for artillery* (AGO-PRB-RM-57-18). Washington, DC: Author.

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Henry Holt & Co.

Allport, G. W., & Odbert, H. S. (1936). Trait names: A psycholexical study. *Psychological Monographs, 47* (1, Whole No. 211).

Apfeldorf, M. (1981). Are personality test differences between alcoholics and others due to source of sample: A review of MMPI findings. *International Journal of the Addictions, 16*, 449-504.

Ashton, S. G., & Goldberg, L. R. (1973). In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices, and laymen. *Journal of Research in Personality, 7*, 1-20.

Bartlett, C. J., & Doorley, R. (1967). Social desirability response differences under research, simulated selection, and faking instructional sets. *Personnel Psychology, 20*, 281-288.

Bass, B. M. (1957). Faking by sales applicants of a forced choice personality inventory. *Journal of Applied Psychology, 41*, 403-404.

Bem, D. J., & Allen, A. (1974). On predicting some of the people some of the time: The search for cross-situational consistencies in behavior. *Psychological Review, 81*, 506-520.

Benton, A. L., & Bechtoldt, H. P. (1955, June). *The Enlisted Personal Inventory (Part I) as a predictor of personal adjustment after recruit training* (Technical Bulletin 55-6). Washington, DC: Bureau of Naval Personnel.

Berkhouse, R. G., & Cook, K. G. (1961, June). *Development of preliminary screening measures for Special Forces trainees* (AGO-HFRB-RM-61-7). Washington, DC: Adjutant General's Office.

Bernardin, H. J. (1977). The relationship of personality variables to organizational withdrawal. *Personnel Psychology, 30*, 17-27.

Bernstein, I. H., Schoenfeld, L. S., & Costello, R. M. (1982). Truncated component regression, multicollinearity and the MMPI's use in a police officer selection setting. *Multivariate Behavioral Research, 17*, 99-116.

Berzins, J. I., Ross, W. F., & Monroe, J. J. (1971). A multivariate study of the personality characteristics of hospitalized narcotic addicts on the MMPI. *Journal of Clinical Psychology, 27*, 174-181.

Block, J. (1965). *The challenge of response sets*. New York: Appleton-Century-Crofts.

Block, J. (1977). Advancing the psychology of personality: Paradigmatic shift or improving the quality of research? In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, NJ: Erlbaum.

Boehm, V. R. (1972). Negro-white differences in validity of employment and training selection procedures: Summary of research evidence. *Journal of Applied Psychology, 56*, 33-39.

Booth, R. F., & Berry, N. H. (1978). Minority group differences in the background, personality, and performance of Navy paramedical personnel. *Journal of Community Psychology, 6*, 60-68.

Bowers, K. S. (1973). Situationism in psychology: An analysis and critique. *Psychological Review, 80*, 307-336.

Broedling, L. A. (1975). Relationship of internal-external control to work motivation and performance in an expectancy model. *Journal of Applied Psychology, 60*, 65-70.

Brown, N. W. (1974). An investigation of personality characteristics of Negroes attending a predominantly white university and Negroes attending a black college. *Dissertation Abstracts International, 34*, 3980-A.

Buros, O. K. (Ed.) (1978). *The eighth mental measurements yearbook.* Highland Park, NJ: Gryphon Press.

Butcher, J. N., Ball, B., & Ray, E. (1964). Effects of socioeconomic level on MMPI differences in Negro-white college students. *Journal of Counseling Psychology, 11*, 83-87.

Cameron, P. (1971). Personality differences between typical urban Negroes and whites. *Journal of Negro Education, 40*, 66-75.

Carleton, F. O., Burke, L. K., Klieger, W. A., & Drucker, A. J. (1957, May). *Validation of the Army Personality Inventory against a military adjustment criterion* (PRB Technical Research Note 71). Washington, DC: Adjutant General's Office, Personnel Research Branch.

Cattell, R. B. (1965). *The scientific analysis of personality*. Chicago: Aldine Publishing Co.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Champaign, IL: Institute for Personality and Ability Testing.

Cheek, J. M. (1982). Aggregation, moderator variables, and the validity of personality tests: A peer-rating study. *Journal of Personality and Social Psychology, 43*, 1254-1269.

Clark, J. H. (1948). Application of the MMPI in differentiating AWOL recidivists from non-recidivists. *Journal of Psychology, 26*, 229-234.

Clark, J. H. (1952). The relationship between MMPI scores and psychiatric classification of Army general prisoners. *Journal of Clinical Psychology, 8*, 86-89.

Collins, D. J. (1967). Psychological selection of drill sergeants: An exploratory attempt in a new program. *Military Medicine, 132*, 713-715.

Comrey, A. L. (1970). *EITS manual for the Comrey Personality Scales*. San Diego: Educational and Industrial Testing Service.

Comrey, A. L., & Backer, T. E. (1970). Construct validation of the Comrey Personality Scales. *Multivariate Behavioral Research, 5*, 469-477.

Comrey, A. L., & Duffy, K. E. (1968). Cattell and Eysenck factor scores related to Comrey personality factors. *Multivariate Behavioral Research, 3*, 379-392.

Comrey, A. L., Jamison, K., & King, N. (1968). Integration of two personality factor systems. *Multivariate Behavioral Research, 3*, 147-160.

Costello, R. M., Schoenfeld, L. S., & Kobos, J. (1982). Police applicant screening: An analogue study. *Journal of Clinical Psychology, 38*, 216-221.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cross, D. T., Barclay, A., & Burger, G. K. (1978). Differential effects of ethnic membership, sex, and occupation on the California Psychological Inventory. *Journal of Personality Assessment, 42*, 597-603.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook, Volume I: Clinical interpretation*. Minneapolis: University of Minnesota Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook, Volume II: Research applications*. Minneapolis: University of Minnesota Press.

Dailey, R. C. (1979). Locus of control, task attributes, and job performance. *Perceptual and Motor Skills, 49*, 489-490.

Datel, W. E. (1962). Socialization scale norms on military samples. *Military Medicine, 127*, 740-744.

Datel, W. E., Hall, F. D., & Rufe, C. P. (1965). Measurement of achievement motivation in Army Security Agency foreign language candidates. *Educational and Psychological Measurement, 25*, 539-545.

Drucker, E. H., & Schwartz, S. (1973, January). *The prediction of AWOL, military skills, and leadership potential* (HumRRO-TR-73-1). Alexandria, VA: Human Resources Research Organization.

Duff, F. L. (1965). Item subtlety in personality inventory scales. *Journal of Consulting Psychology, 29*, 565-570.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychology, 15*, 13-24.

Dunnette, M. D., Peterson, N. G., Houston, J. S., Rosse, R. L., Bosshardt, M. J., & Lammlein, S. E. (1980). *Causes and consequences of adolescent drug experiences: A final report* (PDRI Technical Report 58). Minneapolis: Personnel Decisions Research Institute.

Edwards, A. L. (1959). *Edwards Personal Preference Schedule Manual (Rev.)*. New York: Psychological Corporation.

Edwards, A. L., Diers, C. J., & Walker, J. N. (1962). Response sets and factor loadings on sixty-one personality scales. *Journal of Applied Psychology, 46*, 220-225.

Egbert, R. L., Meeland, T., Cline, V. B., Forgy, E. W., Spickler, M. W., & Brown, C. (1958, March). *Fighter I: A study of effective and ineffective combat performers* (HumRRO SR-13). Washington, DC: Human Resources Research Office.

Ekehammar, B. (1974). Interactionism in personality from a historical perspective. *Psychological Bulletin, 81*, 1026-1048.

Ellis, A., & Conrad, H. S. (1948). The validity of personality inventories in military practice. *Psychological Bulletin, 45*, 385-426.

Endler, N. S., & Hunt, J. McV. (1966). Sources of behavioral variance as measured by the S-R Inventory of Anxiousness. *Psychological Bulletin, 65*, 336-346.

Endler, N. S., & Hunt, J. McV. (1969). Generalizability of contributions from sources of variance in the S-R Inventories of Anxiousness. *Journal of Personality, 37*, 1-24.

Endler, N. S., & Okada, M. (1975). A multidimensional measure of trait anxiety: The S-R Inventory of General Trait Anxiousness. *Journal of Consulting and Clinical Psychology, 43*, 319-329.

Epstein, S. (1977). Traits are alive and well. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, NJ: Erlbaum.

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*, 1097-1126.

Epstein, S. (1980). The stability of behavior: II. Implications for psychological research. *American Psychologist, 35*, 790-806.

Eysenck, H. J., & Eysenck, S. B. G. (1969). *Personality structure and measurement*. San Diego, CA: Knapp.

Eysenck, H. J., & Eysenck, S. B. G. (1975). *Manual for the Eysenck Personality Questionnaire*. San Diego, CA: Educational and Industrial Testing Service.

Farr, J. L., O'Leary, B. S., & Bartlett, C. J. (1971). Ethnic group membership as a moderator of the prediction of job performance. *Personnel Psychology, 24*, 609-636.

Farr, J. L., O'Leary, B. S., Pfeiffer, C. M., Goldstein, I. L., & Bartlett, C. J. (1971, October). *Ethnic group membership as a moderator in the prediction of job performance: An examination of some less traditional predictors* (AIR-753-10/71-TR-2). Washington, DC: American Institutes for Research.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461-477.

Ghiselli, E. E., & Barthol, R. P. (1953). The validity of personality inventories in selecting employees. *Journal of Applied Psychology, 37*, 18-20.

Gilbert, J. G., & Lombardi, D. N. (1967). Personality characteristics of young male narcotic addicts. *Journal of Consulting Psychology, 31*, 536-538.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5*, 351-379.

Goldberg, L. R. (1972). [Review of the California Psychological Inventory]. In O. K. Buros (Ed.), *The seventh mental measurements yearbook* (Vol. 1). Highland Park, NJ: Gryphon Press.

Goldberg, L. R. (1981). Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Personality and social psychology review* (Vol.2) Beverly Hills, CA: Sage.

Golding, S. L. (1975). Flies in the ointment: Methodological problems in the analysis of the percentge of variance due to persons and situations. *Psychological Bulletin, 82*, 278-288.

Gordon, L. V. (1978). *Gordon Personal Profile - Inventory manual*. New York: Psychological Corporation.

Gough, H. G. (1964). Academic achievement in high school as predicted from the California Psychological Inventory. *Journal of Educational Psychology, 55*, 174-180.

Gough, H. G. (1965). Conceptual analysis of psychological test scores and other diagnostic variables. *Journal of Abnormal Psychology, 70*, 294-302.

Gough, H. G. (1966). Appraisal of social maturity by means of the CPI. *Journal of Abnormal Psychology, 71*, 189-195.

Gough, H. G. (1971). The assessment of wayward impulse by means of the Personnel Reaction Blank. *Personnel Psychology, 24*, 669-677.

Gough, H. G. (1975). *Manual for the California Psychological Inventory.* Palo Alto, CA: Consulting Psychologists Press.

Gough, H. G., & Peterson, D. R. (1952). The identification and measurement of predispositional factors in crime and delinquency. *Journal of Consulting Psychology, 16*, 207-212.

Graham, W. K., & Calendo, J. T. (1969). Personality correlates of supervisory ratings. *Personnel Psychology, 22*, 483-487.

Green, L. L., & Haymes, M. (1973). Value orientation and psychosocial adjustment at various levels of marihuana use. *Journal of Youth and Adolescence, 2*, 213-231.

Green, R. F. (1951). Does a selection situation induce testees to bias their answers on interest and temperament tests? *Educational and Psychological Measurement, 11*, 503-515.

Griffin, G. R., & Hopson, J. A. (1978, November). *An evaluation of the Omnibus Personality Inventory in the prediction of attrition in Naval aviation training* (NAMRL-1253). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Griffin, G. R., & Mosko, J. D. (1977, August). *Naval aviation attrition 1950-1976: Implications for the development of future research and evaluation* (NAMRL-1237). Pensacola, FL: Naval Aerospace Medical Research Laboratory.

Griffin, M. L., & Flaherty, M. R. (1964). Correlation of CPI traits with academic achievement. *Educational and Psychological Measurement, 24*, 369-372.

Grow, R., McVaugh, W., & Eno, T. D. (1980). Faking and the MMPI. *Journal of Clinical Psychology, 36*, 910-917.

Guilford, J. P. (1959). *Personality.* New York: McGraw-Hill.

Guilford, J. P. (1975). Factors and factors of personality. *Psycholgical Bulletin, 82*, 802-814.

Guilford, J. P., & Zimmerman, W. S. (1949). *The Guilford-Zimmerman Temperament Survey: Manual.* Beverly Hills, CA: Sheridan Supply.

Guilford, J. S., Zimmerman, W. S., & Guilford, J. P. (1976). *The Guilford-Zimmerman Temperament Survey handbook*. San Diego, CA: Educational and Industrial Testing Service.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18*, 135-164.

Gulas, I., & King, F. W. (1976). On the question of pre-existing personality differences between users and non-users of drugs. *Journal of Psychology, 92*, 65-69.

Gynther, M. D. (1972). White norms and black MMPIs: A prescription for discrimination? *Psychological Bulletin, 78*, 386-402.

Gynther, M. D. (1979). Ethnicity and personality: An update. In J. N. Butcher (Ed.), *New developments in the use of the MMPI*. Minneapolis: University of Minnesota Press.

Gynther, M. D., & Burkhart, B. R. (1983). Are subtle MMPI items expendable? In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol.2). Hillsdale, NJ: Erlbaum.

Hampton, P. J. (1953). The development of a personality questionnaire for drinkers. *Genetic Psychology Monographs, 48*, 55-115.

Harper, F. B. W. (1975). The validity of some alternative measures of achievement motivation. *Educational and Psychological Measurement, 35*, 905-909.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67*, 231-248.

Heisler, W. J. (1974). A performance correlate of personal control beliefs in an organizational context. *Journal of Applied Psychology, 59*, 504-506.

Heist, P., & Yonge, G. (1968). *Manual for the Omnibus Personality Inventory, Form F*. New York: Psychological Corporation.

Heron, A. (1956). The effects of real-life motivation on questionnaire response. *Journal of Applied Psychology, 40*, 65-68.

Hersch, P. D., & Scheibe, K. E. (1967). Reliability and validity of internal-external control as a personality dimension. *Journal of Consulting Psychology, 31*, 609-613.

Hill, H. E., Haertzen, C. A., & Davis, H. (1962). An MMPI factor analytic study of alcoholics, narcotic addicts and criminals. *Quarterly Journal of Studies on Alcohol, 23*, 411-431.

Hodo, G. L., & Fowler, R. D. (1976). Frequency of MMPI two-point codes in a large alcoholic sample. *Journal of Clinical Psychology, 32*, 487-489.

Hoffman, H., Loper, R. G., & Kammeier, M. L. (1974). Identifying future alcoholics with MMPI alcoholism scales. *Quarterly Journal of Studies on Alcohol, 35*, 490-498.

Hogan, J., Hogan, R., & Busch, C. M. (1984). How to measure service orientation. *Journal of Applied Psychology, 69*, 167-173.

Hogan, R. (1971). Personality characteristics of highly rated policemen. *Personnel Psychology, 24*, 679-686.

Hogan, R. (1978). [Review of the Personality Research Form]. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (Vol. 1). Highland Park, NJ: Gryphon Press.

Hogan, R. (1983a). A socioanalytic theory of personality. In M. M. Page (Ed.), *1982 Nebraska Symposium on Motivation*. Lincoln, NE: University of Nebraska Press.

Hogan, R. (1983b). *Personality theory, personality assessment, and Army recruit selection*. Concept paper submitted to U.S. Army Research Institute for the Behavioral and Social Sciences, Project A, Task 2.

Hogan, R., Carpenter, B. N., Briggs, S. R., & Hansson, R. O. (In Press). Personality assessment and personnel selection. In H. J. Bernardin & D. A. Bownas (Eds.), *Personality assessment in organizations*. New York: Praeger.

Hogan, R., DeSoto, C. B., & Solano, C. (1977). Traits, tests, and personality research. *American Psychologist, 32*, 255-264.

Hogan, R., Mankin, D., Conway, J., & Fox, S. (1970). Personality correlates of undergraduate marijuana use. *Journal of Consulting and Clinical Psychology, 35*, 58-63.

Hoiberg, A., Hysham, C. J., & Berry, N. H. (1973). *Predictors related to premature attrition of Navy recruits* (Report No. 73-48). San Diego, CA: Naval Health Research Center.

Hoiberg, A., & Pugh, W. M. (1978). Predicting Navy effectiveness: Expectations, motivation, personality, aptitude, and background variables. *Personnel Psychology, 31*, 841-852.

Holden, R. R., & Jackson, D. N. (1979). Item subtlety and face validity in personality assessment. *Journal of Consulting and Clinical Psychology, 47*, 459-468.

Holden, R. R., & Jackson, D. N. (1981). Subtlety, information, and faking effects in personality assessment. *Journal of Clinical Psychology*, 37, 379-386.

Hoyt, D. P., & Sedlacek, G. M. (1958). Differentiating alcoholics from normals and abnormals with the MMPI. *Journal of Clinical Psychology, 14*, 69-74.

Jackson, D. N. (1960). Stylistic response determinants in the California Psychological Inventory. *Educational and Psychological Measurement, 20*, 339-346.

Jackson, D. N. (1967). *Personality Research Form manual*. Goshen, NY: Research Psychologists Press.

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychological Review, 78*, 229-248.

Jackson, D. N. (1975). The relative validity of scales prepared by naive item writers and those based on empirical methods of personality scale construction. *Educational and Psychological Measurement*, *35*, 361-370.

Jackson, D. N. (1976). *Jackson Personality Inventory manual*. Goshen, NY: Research Psychologists Press.

Jackson, D. N., & Messick, S. (1961). Acquiescence and desirability as response determinants in the MMPI. *Educational and Psychological Measurement, 21*, 771-790.

Jackson, D. N., & Messick, S. (1962). Response styles on the MMPI: Comparison of clinical and normal samples. *Journal of Abnormal and Social Psychology, 65*, 285-299.

Jackson, D. N., & Paunonen, S. V. (1980). Personality structure and assessment. *Annual Review of Psychology, 31*, 503-551.

Johnson, C. D., & Kotula, L. J. (1958, August). *Validation of experimental self-description materials for general and differential classification* (PRB Technical Research Note 95). Washington, DC: Adjutant General's Office, Personnel Research Branch.

Johnson, D. M. (1973). Relationships between selected cognitive and noncognitive variables and practical nursing achievement. *Nursing Research, 22*, 148-153.

Jones, E. E. (1978). Black-white personality differences: Another look. *Journal of Personality Assessment, 42*, 244-252.

Jones, J. W. (1980). Attitudinal correlates of employees' deviance: Theft, alcohol use, and nonprescribed drug use. *Psychological Reports, 47*, 71-77.

Kenrick, D. T., & Stringfield, D. O. (1980). Personality traits and the eye of the beholder: Crossing some traditional philosophical boundaries in the search for consistency in all of the people. *Psychological Review, 87*, 88-104.

King, H. F., Carroll, J. L., & Fuller, G. B. (1977). Comparison of nonpsychiatric blacks and whites on the MMPI. *Journal of Clinical Psychology, 33*, 725-728.

Kirchner, W. K. (1962). "Real-life" faking on the Edwards Personal Preference Schedule by sales applicants. *Journal of Applied Psychology, 46*, 128-130.

Knapp, R. R. (1963). Personality correlates of delinquency rate in a Navy sample. *Journal of Applied Psychology, 47*, 68-71.

Knapp, R. R. (1964). Value and personality differences between offenders and nonoffenders. *Journal of Applied Psychology, 48*, 59-62.

Knecht, S. D., Cundick, B. P., Edwards, D., & Gunderson, E. K. (1972). The prediction of marijuana use from personality scales. *Educational and Psychological Measurement, 32*, 1111-1117.

Kranitz, L. (1972). Alcoholics, heroin addicts and non-addicts: Comparison on the MacAndrew Alcoholism Scale of the MMPI. *Quarterly Journal of Studies on Alcohol, 33*, 807-809.

Lanyon, R. I. (1984). Personality assessment. *Annual Review of Psychology, 35*, 667-701.

Lied, T. R., & Pritchard, R. D. (1976). Relationships between personality variables and components of the expectancy-valence model. *Journal of Applied Psychology, 61*, 463-467.

Linn, R. L. (1978). Single-group validity, differential validity, and differential prediction. *Journal of Applied Psychology, 63*, 507-512.

Loper, R. G., Kammeier, M. L., & Hoffman, H. (1973). MMPI characteristics of college freshman males who later became alcoholics. *Journal of Abnormal Psychology, 82*, 159-162.

Loudermilk, K. M. (1966). Prediction of efficiency of lumber and paper mill employees. *Personnel Psychology, 19*, 301-310.

Lowe, J. D., & Hildman, L. K. (1972). EPI scores as a function of race. *British Journal of Social and Clinical Psychology, 11*, 191-192.

Lykken, D. T. (1978). [Review of the Jackson Personality Inventory]. In O. K. Buros (Ed.), *The eighth mental measurements yearbook* (Vol. 1). Highland Park, NJ: Gryphon Press.

MacAndrew, C. (1965). The differentiation of male alcoholic outpatients from nonalcoholic psychiatric outpatients by means of the MMPI. *Quarterly Journal of Studies on Alcohol, 26*, 238-246.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.

Majumder, R. K., MacDonald, A. P., & Greever, K. B. (1977). A study of rehabilitation counselors: Locus of control and attitudes toward the poor. *Journal of Counseling Psychology, 24*, 137-141.

Manson, M. P. (1949). A psychometric analysis of psychopathic characteristics of alcoholics. *Journal of Consulting Psychology, 13*, 111-118.

McCall, R. J. (1958). Face validity in the D scale of the MMPI. *Journal of Clinical Psychology, 14*, 77-80.

McClelland, J. N., & Rhodes, F. (1969). Prediction of job success for hospital aides and orderlies from MMPI scores and personal history data. *Journal of Applied Psychology, 53*, 49-54.

McCrae, R. R. (1982). Consensual validation of personality traits: Evidence from self-reports and ratings. *Journal of Personality and Social Psychology, 43*, 293-303.

Meehl, P. E. (1945). The dynamics of "structured" personality tests. *Journal of Clinical Psychology, 1*, 296-303.

Megargee, E. I. (1972). *The California Psychological Inventory handbook*. San Francisco: Jossey-Bass.

Mills, C. J., & Bohannon, W. E. (1980). Personality characteristics of effective state police officers. *Journal of Applied Psychology, 65*, 680-684.

Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

Mischel, W., & Peake, P. K. (1982). Beyond deja vu in the search for cross-situational consistency. *Psychological Review, 89*, 730-755.

Monson, T. C., Hesley, J. W., & Chernick, L. (1982). Specifying when personality traits can and cannot predict behavior: An alternative to abandoning the attempt to predict single-act criteria. *Journal of Personality and Social Psychology, 43*, 385-399.

Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.

Nord, W. R., Connelly, F., & Daignault, G. (1974). Locus of control and aptitude test scores as predictors of academic achievement. *Journal of Educational Psychology, 66*, 956-961.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66*, 574-583.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

O'Dell, J. W. (1971). Method for detecting random answers on personality questionnaires. *Journal of Applied Psychology, 55*, 380-383.

Olweus, D. (1977). A critical analysis of the "modern" interactionist position. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology*. Hillsdale, NJ: Erlbaum.

Pervin, L. A. (1978). *Current controversies and issues in personality.* New York: Wiley.

Porter, L. W., & Steers, R. M. (1973). Organizational, work and personal factors in employee turnover and absenteeism. *Psychological Bulletin, 80*, 151-176.

Rathus, S. A., Fox, J. A., & Ortins, J. B. (1980). The MacAndrew scale as a measure of substance abuse and delinquency among adolescents. *Journal of Clinical Psychology, 36*, 579-583.

Rich, C. C., & Davis, H. G. (1969). Concurrent validity of MMPI alcoholism scales. *Journal of Clinical Psychology, 25*, 425-426.

Rorer, L. G. (1965). The great response-style myth. *Psychological Bulletin, 63*, 129-156.

Rosenberg, L. A., McHenry, T. B., Rosenberg, A. M., & Nichols, R. C. (1962). The prediction of academic achievement with the California Psychological Inventory. *Journal of Applied Psychology, 46*, 385-388.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80* (1, Whole No. 609).

Ruch, F. L., & Ruch, W. W. (1967). The K factor as a (validity) suppressor variable in predicting success in selling. *Journal of Applied Psychology, 51*, 201-204.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94*, 18-38.

Rushton, J. P., Jackson, D. N., & Paunonen, S. V. (1981). Personality: Nomothetic or idiographic? A response to Kenrick and Stringfield. *Psychological Review, 88*, 582-589.

Schmidt, H. D. (1945). Test profiles as a diagnostic aid: The Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology, 29*, 115-131.

Schmitt, N., Mellon, P. M., & Bylenga, C. (1978). Sex differences in validity for academic and employment criteria, and different types of predictors. *Journal of Applied Psychology, 63*, 145-150.

Schuh, A. (1967). The predictability of employee tenure: A review of the literature. *Personnel Psychology, 20*, 133-152.

Schwab, D. P. (1971). Issues in response distortion studies of personality inventories: A critique and replicated study. *Personnel Psychology, 24*, 637-647.

Schwab, D. P., & Packard, G. L. (1973). Response distortion on the Gordon Personal Inventory and the Gordon Personal Profile in a selection context: Some implications for predicting employee tenure. *Journal of Applied Psychology, 58*, 372-374.

Shenk, F., Watson, T. W., & Hazel, J. T. (1973, May). *Relationship between personality traits and officer performance and retention criteria* (AFHRL-TR-73-4). Lackland Air Force Base, TX: Air Force Human Resources Laboratory, Personnel Research Division.

Sinha, A. K. P. (1963). Manifest anxiety affecting industrial absenteeism. *Psychological Reports, 13*, 258.

Smart, R. G., & Fejer, D. (1969). Illicit LSD users: Their social backgrounds, drug use and psychopathology. *Journal of Health and Social Behavior, 10*, 297-308.

Spector, P. E. (1982). Behavior in organizations as a function of employee's locus of control. *Psychological Bulletin, 91*, 482-497.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology, 48*, 285-290.

Tellegen, A. (1964). The Minnesota Multiphasic Personality Inventory. In L. E. Abt & B. F. Riess (Eds.), *Progress in clinical psychology* (Vol.6). New York: Grune & Stratton.

Tellegen, A. (1965). Direction of measurement: A source of misinterpretation. *Psychological Bulletin, 63*, 233-243.

Tellegen, A. (1981). Practicing the two disciplines for relaxation and enlightenment: Comment on "Role of the feedback signal in electromyograph biofeedback: The relevance of attention" by Qualls and Sheehan. *Journal of Experimental Psychology: General, 110*, 217-226.

Tellegen, A. (1982). *Brief manual for the Differential Personality Questionnaire*. Unpublished manuscript, University of Minnesota.

Tellegen, A., Kamp, J., & Watson, D. (1982). Recognizing individual differences in predictive structure. *Psychological Review, 89*, 95-105.

Toole, D. L., Gavin, J. F., Murdy, L. B., & Sells, S. B. (1972). The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology, 25*, 661-672.

Tseng, M. S. (1970). Locus of control as a determinant of job proficiency, employability, and training satisfaction of vocational rehabilitation clients. *Journal of Counseling Psychology, 17*, 487-491.

Tubiana, J. H., & Ben-Shakhar, G. (1982). An objective group questionnaire as a substitute for a personal interview in the prediction of success in military training in Israel. *Personnel Psychology, 35*, 349-357.

Tupes, E. C., & Christal, R. E. (1961, May). *Recurrent personality factors based on trait ratings* (ASD-TR-61-97). Lackland Air Force Base, TX: Aeronautical Systems Division, Personnel Laboratory.

Vega, A. (1971). Cross-validation of four MMPI scales for alcoholism. *Quarterly Journal of Studies on Alcohol, 32*, 791-797.

Voas, R. B. (1957). Validity of personality scales for the prediction of success in naval aviation training. *American Psychologist, 12*, 465.

Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.

Webster, E. G., Booth, R. F., Graham, W. K., & Alf, E. F. (1978). A sex comparison of factors related to success in Naval Hospital Corps school. *Personnel Psychology, 31*, 95-106.

Weckowicz, T. E., & Janssen, D. V. (1973). Cognitive functions, personality traits, and social values in heavy marijuana smokers and nonsmoker controls. *Journal of Abnormal Psychology, 81*, 264-269.

Weiss, P., Wertheimer, M., & Groesbeck, B. (1959). Achievement motivation, academic aptitude, and college grades. *Educational and Psychological Measurement, 19*, 663-666.

Wiener, D. N. (1948). Subtle and obvious keys for the Minnesota Multiphasic Personality Inventory. *Journal of Consulting Psychology, 12*, 164-170.

John M. Kamp and Leatta M. Hough

# SECTION 2

# UTILITY OF BIOGRAPHICAL DATA FOR PREDICTING JOB PERFORMANCE

## Overview

Perhaps the most common axiom applied to attempts to understand human behavior is that past behavior is the best predictor of future behavior. Examples of this are so commonplace in everyday life that it is difficult to envision a situation in which behavior would be predicted that had no analogue in previous behavior. Further, the quality of this prediction is thought to be a function of the degree of correspondence between the past behavior and the future behavior.

An excellent example of this rationale is the relationship between grade point average in high school (past behavior) and in college (future behavior). Since both behaviors share many common components such as study habits, aptitude and academic interest, there is a high degree of correspondence between the predictor space and the criterion space. Thus, one would expect to observe a high degree of relationship in the two grade point averages, an expectation that has been repeatedly confirmed (Freeberg, 1967). Also, to the extent that the predictor space differs from the criterion space, the observed relationship is expected to be lower.

A similar argument has been advanced by Wernimont and Campbell (1968), building on the earlier work of Goodenough (1949) and Cronbach (1960). A distinction was drawn between signs, or indicators of predispositions to behave in certain ways, and samples of the characteristic behaviors of individuals. Wernimont and Campbell pointed out that measures of temperament, aptitude, and interests have come to be used as signs of future behavior, with the rationale that criterion behavior is determined by these somewhat generalized predispositions. They argued that it may be more fruitful to focus on meaningful samples of behavior that are consistent with, or have many common components with, the desired future behavior.

Thus, the rationale for the use of autobiographical information (biodata) as a predictor of behavior in the employment setting is based first on the axiom that past behavior is the best predictor of future behavior. Secondly, it specifies that a sample of past behavior can be elicited through biodata items that shares many common components with the criterion behavior. Increasing the number and importance of these common components is expected to increase the validity of the biodata predictor. In addition, the biodata approach emphasizes measures of behavior rather than predispositions, thus eliminating the intermediary steps of the "sign" approach and limiting criticism of the relevance of test content.

<u>History</u>. Among the first systematic work with biographical data was the objective scoring of the application blank, the immediate forebear of the scored biodata form. Owens (1976) wrote that at the 1894 Chicago Underwriters' meeting, Colonel Thomas Peters proposed that managers should require all applicants to answer a list of standardized questions in order to improve the selection of life insurance agents. Sample items included age, length of residence, marital status, and experience with selling life insurance.

Research in the life insurance industry continued with the development by Woods in 1915 of empirical analysis of the responses of good and poor salesmen to application blank items. In 1922, Dorothy Goldsmith published the first journal article that involved biographical data, and in which the procedures of empirical item analysis and weighting were made quite explicit. Also in 1922, Gerirude Cope analyzed biodata collected at the time of employment from more than 400 men hired as life insurance agents in 1919, 1920, and 1921. She found the following items to be related to success and failure: age, number of dependents, marital status, education, years since leaving school, selling experience, membership in social organizations, offices held in social organizations, home ownership, number of investments, and life insurance ownership (Thayer, 1977).

Only three years later, Manson (1925) reported on combining items for sales selection via multiple correlation analysis, Kenagy and Yoakum (1925) examined background factors and personal data in relation to general sales success, and Cope (see Anon, 1925) suggested their potential value in selecting public employees (Owens, 1976). Research in the life insurance industry also continued with the work of Albert Kurtz and Arthur Kornhauser, who analyzed biodata from thousands of agents. Kurtz identified age as a moderator variable, and together both men developed the Aptitude Index, a biodata instrument from which a currently used inventory (the Aptitude Index Battery) was derived (Thayer, 1977).

Biodata was also put to use in military applications. For example, Henry (1966) reported that the single item "Did you ever build a model airplane that flew?" was almost as good a predictor of success in flight training during World War II as the entire Air Force Battery. Investigators such as Guilford and Lacey (1947) and Levine and Zachert (1951) reported that validities for biodata in predicting military pilot and navigator success were around .30 during World War II. Following World War II, the use of biodata became more widespread, ranging from early work in classification for the Air Force (Levine & Zachert) to the extensive criterion-related validity research undertaken by all branches of the military. In the civilian sector, the 1950s and 1960s saw the weighted application blank increase in popularity (England, 1961, 1971).

Despite the increase in the demonstrated utility of biodata, however, little effort had been devoted to understanding the underlying meaning of this information. Consequently, a conference in 1965 entitled "Research on the Use and Meaning of Autobiographical Data as Psychological Predictors" assembled many of the leading researchers in America to address the problem. Among their conclusions was the following: "Aside from theoretical academic interest there were no very persuasive reasons for tackling the program until a 'prediction plateau' developed. It seems apparent now that increased efficiency will occur only when we learn more about the causal relationships underlying predictive items" (Henry, 1966, p. 248).

In accord with this observation, more recent effort in the domain has increased the emphasis on conceptual understanding of biographical information. Several factor analyses have been performed (Eberhardt & Muchinsky, 1982; Owens, 1971) and homogeneous scales have been developed for biodata items (Matteson, 1978). In addition, Owens (1968, 1971) and Owens and Schoenfeldt (1979) have proposed a model for classification of individuals that uses biographical data to form subgroups, the characteristics of which may be further elucidated by the results of other measures. These efforts have been an attempt to counteract the criticisms of biodata as achieving quite good prediction that is nevertheless accompanied by very little gain in understanding. Each topic mentioned in this overview is discussed in more detail later in this section.

## Measurement

As previously noted, the information recorded on an application blank constituted the first systematic method of biodata measurement. A more sophisticated form of the approach is the weighted application blank (WAB) in which weights are assigned in accordance with the predictive power of each item. The other major type of measurement instrument is the biographical information blank (BIB) which normally includes a larger and broader sample of items and a multiple-choice format.

While it is difficult to make absolute distinctions between the two methods, the WAB is typically shorter and more narrow in focus, with items such as "marital status" and "age" that are quite clearly associated with the application blank. By contrast, a BIB often includes items related to an individual's childhood and development and focuses on obtaining a broader assessment of background. BIBs are typically more conducive to conceptual analysis while WABs are aimed primarily at strictly empirical prediction. The two techniques are similar in that both collect self-report data and may have many overlapping items with similar content.

### Weighted Application Blanks

Construction. Although numerous WAB construction techniques are available, the approach of England (1971) is discussed because it is both representative and the most widely cited. England specified seven major steps in developing a WAB, each of which is discussed below. The first step is selecting the criterion that the blank will serve to predict. Choice of this criterion is critical because the construction of the WAB depends upon an item analysis of the responses of different criterion groups. The effectiveness of the WAB depends upon the criterion used in its development; if the criterion is inadequate or contaminated, or its measurement is inaccurate or unreliable, prediction will suffer. Also, a criterion chosen for convenience rather than importance and representativeness is likely to be of little utility.

The next step involves identifying the members of both high and low criterion groups. For example, employees who quit their job within three months of hiring might comprise the low criterion group and employees with more than one year of service could constitute the high criterion group. The exact criterion levels used to establish these cutoffs will vary with the organization, the criterion, and the number of employees available for classification into criterion groups. In order to assure adequate statistical power, Cascio (1982) recommends that criterion groups be as large as possible, with no fewer than 125 individuals in each group. Group size is particularly important because the groups must be further subdivided into validation and cross-validation groups. England (1971) recommends the inclusion of approximately one-third more individuals in the validation group than in the cross-validation group.

The selection of items is the next consideration, and England (1971) advocates using as many items as possible in the initial tabulations because a large proportion will not differentiate between the criterion groups. WABs tend to include items of a very heterogeneous nature, for example: "number of children," "previous military record," "educational level," "number of previous jobs," and "number of character references." Since only those items that differentiate between the two groups are ultimately included, this procedure represents raw empiricism at its extreme and maximizes the probability that an item could be selected based on chance alone. For this reason, the use of large samples and cross-validation is critical to lessening the likelihood of spurious results. In addition, such a blindly empirical approach often results in selected items that are criticized for lacking relevance and for contributing to adverse impact (Pace & Schoenfeldt, 1977). Consideration of these issues is important in item selection and will be discussed in more detail later.

Following initial item selection, response options or categories must be selected. For some items, suitable response categories will be obvious, as in the case of "marital status." Continuous variables, however, require somewhat arbitrary categories, perhaps using equal frequencies or equal intervals within the category to divide the responses. Another method is to use trial and error to determine the categorization that best brings out the differences between the groups. This approach capitalizes on chance differences that may result in instability, and England (1971) recommends the method of equal frequency classes. Also, some items may require the clustering of responses into groups, as would be necessary for the item "previous occupation." If the investigator wishes to specify response categories in advance, a rational approach might be used, perhaps guided by the results of previous research.

Once the responses of the criterion groups are categorized, the percentages of each group in a particular category are compared and weights are assigned on the basis of group differences. These weights may then be transformed into smaller, positive values for ease in future scoring based on tables originally developed by E. K. Strong. For example, assume that for the item "marital status," 25% of the high criterion group and only 4% of the low criterion group belong to the response category "widowed." Based on this 21% difference, an initial weight of +6 is assigned, which is then transformed to an assigned weight of +2. In England's (1971) approach, those items that do not discriminate between the two groups (in terms of percent difference) are dropped, and responses that are associated more with the low than the high group are included but receive a low or zero assigned weight.

At this point, a total WAB score can be obtained for each individual and a point-biserial correlation can be computed between criterion group membership and WAB score. It is critical, however, that the correlation be cross-validated on the holdout criterion groups that were not used in deriving item response weights. Again, cross-validation is especially necessary because WAB procedures are blindly empirical and many of the observed differences in weights may reflect chance fluctuations rather than true differences (Cascio, 1982).

The final step in the WAB procedure is to establish the cutoff score that attempts to place the maximum number of persons, according to their total scores, in the proper criterion group. In other words, the cutoff score should separate the high from the low criterion groups with a minimum of overlap (England, 1971). In some cases, this may be done by visual inspection of the two group distributions. When questionable, however, a procedure called maximum differentiation can be implemented through a table that compares the percentage of each group scoring at or above each possible total score on the WAB. The point at which the difference between the percentages is greatest is the optimal cutting score. Cascio (1982) advocates taking other factors, such as market conditions and the cost of erroneous acceptances or rejections, into account as well.

Before we turn to a discussion of the advantages and disadvantages of WABs, it is again noted that construction methods in addition to the one discussed do exist and may be more appropriate for certain purposes. The approach of England (1971), however, appears to be the most widely used for construction of WABs.

Advantages and Disadvantages. Perhaps the primary advantage of the WAB procedure is that this procedure, or a similar form, has worked well in the past. Research that will be discussed in more detail later has demonstrated substantial validity with the WAB, a conclusion that also has intuitive and anecdotal support. Because of the wide use of and confidence in application blanks, there is often less applicant resistance to them than to temperament inventories or cognitive tests. WABs can be completed quickly, are verifiable, and combine easily with other sources of information. In addition, the cost and time involved in developing a WAB can be nominal.

The disadvantages of WABs pertain mainly to the limitations of any purely empirical scales. As pointed out by Weiss (1976), WABs are developed in much the same way as were the occupational scales of the Strong-Campbell Interest Inventory. Like those scales, the WAB yields what might be termed a "blind" empirical prediction that lacks broad interpretability and impedes gains in conceptual understanding. For example, the finding that marital status, education, and height all are related to length of employment does little to explain why this occurs or why other variables may be unrelated.

A second disadvantage is somewhat related. WABs are usually developed for a specified and often narrow criterion (e.g., absenteeism) and a specified and often narrow population (e.g., Acme truck drivers). Because of the empirical construction of the WAB, it is not likely to generalize well to other organizations, types of people, or related criteria. Empirical construction techniques may also lead to adverse impact and legal problems (Pace & Schoenfeldt, 1977). Further, the validity of a WAB may decrease rapidly as a function of time (Dunnette, Kirchner, Erickson, & Banas, 1960; Wernimont, 1962), although Brown (1978) showed this is not always the case.

## Biographical Information Blanks

As mentioned, the biographical information blank (BIB) is somewhat akin to the WAB, although there are important differences. Owens (1976) defines a BIB as a standardized data form, commonly composed of multiple-choice items, that permits the respondent to describe him- or herself in terms of demographic, experiential, or attitudinal variables. These variables are presumed or demonstrated to be related to temperament variables, personal adjustment, or success in social, educational, or occupational pursuits. The items attempt to elicit factual data and the attitudes, feelings, and value judgments that result from prior experience. Like the WAB, the emphasis is on past, not present, behavior. BIBs are broader in item content, however, and emphasize self-description over demographic characteristics.

Item Construction. Perhaps the most important consideration in constructing and/or selecting items is maximizing validity. As would be expected, Williams (1961) found that items which he hypothesized would be related to a criterion of enlistment or nonenlistment were later found to be valid a much higher proportion of the time than items for which a hypothesis was not developed. Buel (in a 1971 personal communication cited in Owens, 1976) showed that items of a strictly historical nature were not valid as often as those in which opinion, attitude, and value dimensions were implied. Owens (1976) wrote that, based on his experience, items in which the response options lie along either an apparent or a demonstrated continuum are preferable to noncontinuum items in terms of validation probability.

In an investigation of reliability, Owens, Glennon, and Albright (1962) found that four item-construction rules were related to later test-retest consistency. These are: (a) Brevity is desirable; items with fewer words were found to elicit responses that were more consistent across administrations. (b) Whenever possible, numbers should be used to graduate or scale, and to define options or alternatives. (c) Either all response options should be covered or an "escape' option should be provided. (d) Items, particularly item stems, should carry a neutral or a pleasant connotation for the respondent.

Key (Scale) Construction. Traditionally, the BIB has shared many of the empirical key construction techniques that have been used in WAB construction. Those items that were related to the chosen criterion (differentiated between high and low groups) were retained and those items found to be unrelated were discarded. These empirical keying methods became so commonplace that Long and Sandiford (1935) were able to review 23 different methods (Mitchell & Klimoski, 1982).

As was mentioned for the WAB, however, a blindly empirical approach has been criticized for its lack of interpretability and advancement in conceptual understanding. Thus, as has been the case in other noncognitive areas (Clark, 1961; Hase & Goldberg, 1967), attempts have been made to develop "rational" scoring systems. Among the first of these systems was the work of Loevinger, Gleser, and DuBois (1953) who demonstrated that by maximizing the homogeneity of each subtest (scale) and minimizing the correlations between subtests, the discriminating power of a biographical inventory was maximized. Morrison, Owens, Glennon, and Albright (1962) showed that factorially derived biographical data dimensions could be used to interpret the differing profiles of three criterion groups. In a followup study, Baehr and Williams (1967) obtained 15 factors, nearly all of which discriminated among the 10 occupations included in the sample at the .0001 level of significance. Similar results were obtained by Klimoski

(1973) with various types of engineers.

Factor-analytic techniques, however, may not lend themselves to easy development of scoring keys for predictive purposes, and so Matteson (1978) adapted the procedure of DuBois, Loevinger, and Gleser (1952) in constructing homogeneous biodata keys. This technique involves grouping items into clusters, on the basis of subjective judgment of item content, and then generating a covariance matrix. From each cluster, a nucleus of three items is chosen that has a high internal covariance value. Then, items from the cluster are added singly to the key and retained only if they increase the saturation of the key. Results of the study indicated that the predictive power of the homogeneous keys was approximately equal to that found with the empirical key. Similar results have been found in the areas of temperament (Hase & Goldberg, 1967) and interests (Reilly & Echternacht, 1979).

Mitchell and Klimoski (1982) set out to test three hypotheses regarding the keying of biodata items via England's (1971) empirical approach (that often results in heterogeneous keys) versus a factor-analytic approach (that results in homogeneous or "rational" scoring keys). The hypotheses were: "(a) the derivation validity would be higher for the empirical approach, (b) the rational approach would produce a validity that would suffer less shrinkage on cross-validation, and (c) the cross-validities of the rational versus empirical methods would not be substantially different" (p. 412). The empirical key was constructed based on item-criterion relationships while the rational key was developed from factor analyses of the items. Derivation validity coefficients were .59 for the empirical approach and .36 for the rational approach. Upon cross-validation, the empirically derived validity fell to .46 while the rationally developed validity stayed at .36. Thus, the first two hypotheses were confirmed, but the empirical approach was found to be superior in the test of the third hypothesis. In their discussion of these results, however, Mitchell and Klimoski stated that the difference in validity between the two approaches was not practically significant.

To summarize, the construction of BIBs has historically been directed more toward achieving empirical prediction than producing a rational scoring system that would allow conceptual advancement. A number of studies have attempted to address this imbalance and their results suggest that such effort will be profitable in advancing the understanding and utility of the domain.

While the increased conceptual orientation is laudable, it is noted that such effort accrues much of its value through previous empirically demonstrated validity. Meehl (1954) pointed out that there is the discriminative (or validating) use of statistics as well as the structural (or analytic) use. The two uses correspond to an empirical approach that makes few assumptions in testing for relationships (in search of basic findings), and a factor-analytic approach that presupposes certain empirically based assumptions (from previous research). The two methods differ in both aim and assumptions, and yet are complementary in the development of the optimal biodata instrument. Further, the paucity of studies using a rational keying approach clearly warrants additional research, but acquires much of its direction from the wealth of previous empirical findings.

## Additional Measurement Concerns

Accuracy/Faking. As stated earlier, the use of biographical data is based on the "past predicts future" axiom that presumably affords it an advantage over other types of measures. If, however, biodata is to capitalize on the axiomatic relationship, it presumably must reflect an accurate picture of what the past actually was. Inaccurate information would eliminate this advantage and might result in lower obtained validities. Surprisingly, there have been relatively few published investigations of the level of accuracy of self-reported biographical data, and its effect on validity.

Among the first relevant findings was that of Moore (1942) who reported that, for Army personnel in World War I, only 6% of those claiming to be skilled in a particular trade actually had adequate skill, while over 30% were totally inexperienced. Lipsett (1946) and Goheen and Mosel (1950) found considerable discrepancies between information on civil service applications and the results of later investigations. On the other hand, Harris (1946) reported that military personnel rarely falsified previous psychiatric experiences, a conclusion supported by Conrad and Ellis (1948). Correlations ranging from .90 to .98 were obtained by Keating, Paterson, and Stone (1950) between interview information and previous employers' records. Similar correlations were obtained by Mosel and Cozan (1952) with application blank data.

Perhaps the most serious questions regarding biodata accuracy have been raised by Goldstein (1971). Using a sample of 94 job applicants, he found substantial discrepancies between the reported information and that supplied by previous employers. These discrepancies varied by item, but for the item "previous job duration," 57% of the applicants provided discordant data that were almost always an overestimate and by an average of 16 months. By contrast, Cascio (1975) found a median accuracy correlation of .94 for BIB information, although this was not in a selection setting.

Schrader and Osburn (1977) investigated the problem from the perspective of the faking literature by comparing the responses of students to a BIB under instructions to fake (Time 1) and to respond honestly (Time 2). Results indicated that subjects were able to substantially distort their scores in a positive direction under faking instructions, but a warning of a lie scale was an effective deterrent to faking during the honest (Time 2) administration. Haymaker and Erwin (1980) showed that it is possible to construct a faking key that can be effectively used to adjust BIB scores. Such adjustments acted to decrease mean scores, increase score variances, and had either no effect or a positive effect on validity.

Conclusions are difficult to draw from this relatively small number of somewhat heterogeneous studies concerning the accuracy or fakeability of biodata instruments. It has been definitely shown that responses can be distorted, either under instructions to do so or in a selection setting. Nevertheless, the prevalence of this state of affairs is questionable. The finding that a faking scale or faking scale warning is effective in reducing distortion suggests that users of biodata instruments would be well advised to use one of these procedures. Finally, the assumption that accurate life history is a better predictor than an individual's perception of that life history, even if such perception is inaccurate, is basically untested.

Reliability. The majority of studies investigating the reliability of biodata keys have used the test-retest method, and have obtained quite respectable coefficients. Reliabilities of .82 to .88 were obtained for ninth-grade students responding to the empirically keyed biodata form of the Institute for Behavioral Research in Creativity (IBRIC, 1968) which includes in excess of 100 items. Chaney and Owens (1964) obtained a figure of .85 with an 82-item form administered to college students with a 19-month interval. Erwin and Herring (1977) reported that correlations between the biodata scores of Army recruits obtained at administration during initial enlistment processing and several weeks later after entering the Army were .91 and .85, using BIBs of 67 and 36 items respectively. Similarly, in the manual for the 278-item Aptitude Index Battery (a BIB used in the life insurance industry), the Life Insurance Marketing and Research Association reported a test-retest reliability of .90, based on samples of more than 5,000 job applicants with an interval of up to 5 days (LIMRA, 1979). Mumford and Owens (1982) employed a factorially derived, 389-item battery and obtained test-retest coefficients ranging from .91 to .97 for males and from .77 to .97 for females.

In addition, one study was located that obtained estimates of internal consistency reliability. Baehr and Williams (1967), using factorially developed scales of only 4 to 13 items, reported KR-20 reliability estimates ranging from .43 to .76 with a vocationally heterogeneous sample of 680 male subjects.

Legal Issues. The *Uniform Guidelines on Employee Selection Procedures* (Equal Employment Opportunity Commission, 1978) specify that any selection procedure that has an adverse impact on members of a subgroup is illegal unless its job relatedness can be demonstrated (see also *Griggs v. Duke Power*, 1971). Since a number of biodata items ask explicitly for the individual's subgroup membership (e.g., marital status), they by definition could have adverse impact if used as one factor in employment decisions. Items may also have more subtle adverse impact. For example, Rosenbaum (1976) found that individuals with Detroit addresses were more likely to have been caught stealing than persons who lived outside the city limits. As Pace and Schoenfeldt (1977) pointed out, this item has adverse impact on black applicants to the extent that they are more likely than white applicants to live within city limits.

An issue related to adverse impact is that of item objectionability. While objectionable items may not always be discriminatory, they certainly will contribute to a larger number of complaints, and are often associated with adverse impact. It is interesting that of the eight WAB items that England's (1971) review found to be valid in three or more occupational areas, all would probably be considered either objectionable or discriminatory today. This is also true of some items commonly used in military research (e.g., number of arrests or school suspensions).

Since it is clear that a number of biodata items may be problematic in this respect, three possible solutions have been suggested. The first is to demonstrate job relatedness, as is typically done in a criterion-related validity study. Since only those items that are empirically related to the criterion are retained, any adverse impact might be assumed to be justifiable. Based on *Albemarle Paper Co. v. Moody* (1975), however, the *Uniform Guidelines* (1978) state that: "In conducting a validation study, the employer should consider available alternatives which will achieve its legitimate business purpose with lesser adverse impact" (p. 38291).

The second option, therefore, is to screen out those items with adverse impact and identify new items of equal or higher validity that are not discriminatory. One way to do this is to write items, based on job analysis information, that tap the experiential or success dimensions of the job. Not only would such an approach contribute to high criterion-related validity, it is also more defensible legally and more likely to produce a form to which applicants are receptive.

A third technique has been suggested by researchers such as Rosenbaum (1976), who advocated the use of different scoring keys for each race-sex group. Thayer (1977) reported that the score distribution on the Aptitude Index Battery, a life insurance biodata inventory, is slightly lower for blacks and Hispanics and lower still for women but, through scaling adjustments, adverse impact is reduced or eliminated with little effect on validity. Finally, Reilly and Chao (1982) concluded that the validity and fairness of biodata can be expected to hold for minority and majority groups, although different keys may be needed for males and females.

## Structural and Conceptual Issues

Historically, much of the effort within the biodata domain has proceeded with little or no consideration of the basic dimensions of biographical information. In fact, this lack of conceptual orientation has been the major criticism of the area: "Most users of the (biodata) method have been more intent on achieving statistical prediction than on gaining an understanding of the dynamics of success which may be suggested by the data" (Dunnette, 1962, p. 299). In attempting to correct this problem, more recent research has been directed at conceptual groupings, both of items and of persons. Grouping of items has usually taken the form of factor analysis.

### Factor Analysis

Among the first systematic attempts to group biodata items into meaningful conceptual dimensions was that of Morrison et al. (1962). A principal components factor analysis of 75 criterion-developed life history items and three job criteria was performed, yielding five factors that accounted for 23% of the total variance. The authors wrote that the relatively small proportion of variance accounted for is to be expected with heterogeneous life history data, but that the dimensions uncovered were nevertheless useful in examining the differential profiles of the criterion groups.

In a similar study, Baehr and Williams (1967) factored 150 multiple-choice biodata items, dropped 45 because of low or ambiguous loadings, and then rotated to both oblique and orthogonal factorial structures. Fifteen primary factors were obtained that accounted for 43% of the variance, as well as five second-order factors that were largely uncorrelated. The factors were found to be useful in discriminating among occupational groups, and as is typically found in biodata factoring, the orthogonal and oblique solutions were quite similar.

The generalizability of the factorial approach for biodata has been supported by Cassens (1966) who showed that the factor structure of a BIB was virtually the same for individuals in both North and South America. Similar findings for various age groups were obtained by Schmuckler (1966) in factoring a BIB.

Perhaps the most notable results of the factor analytic studies in the biodata area were reported by Owens and Schoenfeldt (1979) in summarizing well over a decade of research. The first series of factor analyses they reported was performed on the responses of 1,700 male freshmen at Purdue University to 659 rationally chosen items (from an original pool of 2,000 items). Eight factors were ultimately obtained, and a total of 389 items retained. Next, the number of items was reduced to 275 because of computer limitations, and the responses of 1,037 male and 897 female freshmen at the University of Georgia were factored. Rotation to an orthogonal varimax criterion of simple structure resulted in 13 male data factors and 15 female data factors that were sizable and interpretable.

Owens and Schoenfeldt (1979) reported that these analyses led directly to the development of a 118-item short form, the *Biographical Questionnaire* (BQ), composed of the items with the highest loadings. This form was administered to virtually all incoming University of Georgia freshmen from 1970 through 1973, with results supporting the factor structure obtained from the 275-item form. The factors obtained are as follows:

Males

Warmth of Parental Relationship
Intellectualism
Academic Achievement
Social Introversion
Scientific Interest
Socioeconomic Status
Aggressiveness/Independence
Parental Control vs. Freedom
Positive Academic Attitude
Sibling Friction
Religious Activity
Athletic Interest
Social Desirability

Females

Warmth of Maternal Relationship
Social Leadership
Academic Achievement
Parental Control vs. Freedom
Cultural-Literary Interests
Scientific Interest
Socioeconomic Status
Expression of Negative Emotions
Athletic Participation
Feelings of Social Inadequacy
Adjustment
Popularity With Opposite Sex
Positive Academic Attitude
Warmth of Paternal Relationship
Social Maturity

More recently, Neiner and Owens (1982) examined the temporal stability of the factor structure by collecting information in 1975 on the 1968 freshman class sample. Twenty-five percent of that class had both completed their bachelor's degree in four years and returned a 97-item BIB about their post-college experiences. To compare the factor structures resulting from the two inventories, canonical correlations were computed for both males and females. The obtained correlations for the same set of subjects between the underlying dimensions in 1968 and in 1975 ranged from .56 to .64, thus supporting fairly good stability of biodata dimensions over a 7-year period.

The Owens and Schoenfeldt (1979) factors, although constituting perhaps the most thoroughly replicated structure in the biodata domain, have been criticized as being characteristic only of University of Georgia college students. Thus the investigation by Eberhardt and Muchinsky (1982b), which also was based on a large sample size (816) but from a different sample of undergraduates at Iowa State University, represented an important test of generality. The factors obtained were virtually identical to those of Owens and Schoenfeldt for males, but differed on 6 of the 15 factors for females. These results suggest that the basic dimensions of biographical data as assessed by the BQ are remarkably stable for males, both across samples and over a period in excess of 10 years. Similar but weaker conclusions may also be drawn for females, perhaps due to the factor structure reflecting the changing life experiences of women in the last decade.

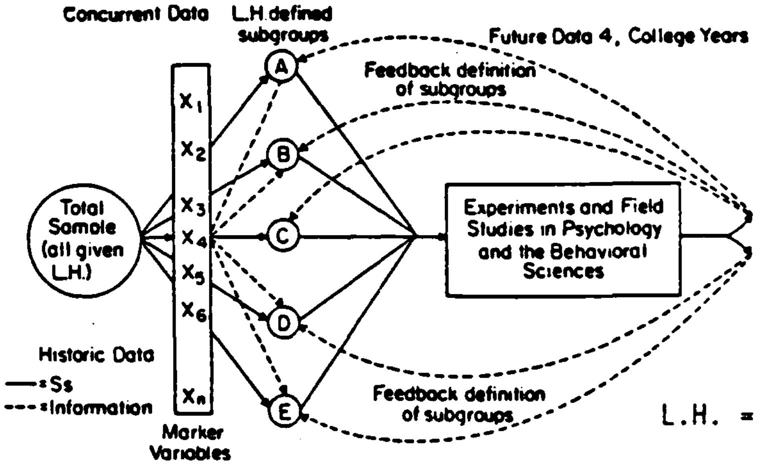
Since very few standardized BIBs are in wide use, and the number of large-scale factor analyses performed has been relatively small, it is difficult to say whether the factor structure reported by Owens and Schoenfeldt (1979) and Eberhardt and Muchinsky (1982) is similar to what would be obtained with biodata instruments quite different from the BQ. There appear to be some corollaries with the results of Morrison, Owens, Glennon, and Albright (1962) and particularly with Baehr and Williams (1967). This is clearly an area that requires further explication, however, since it influences not only how we think about biographical data, but also how measurement instruments might best be constructed and used. A stable, replicable, and generalizable structure of biodata dimensions is also important in another usage of biographical information--that of classifying persons into homogeneous subgroups that differ on important characteristics.

## Developmental-Integrative Model

Following the lead of Cronbach (1957) in calling for increased links between the experimental and correlational disciplines of psychology, Owens (1968) proposed a linking vehicle that he entitled the developmental-integrative (D-I) model. This model was designed to provide knowledge of behavior antecedents, a major weakness of the correlational approach, while also incorporating measures of individual differences, the major deficiency of the experimental method. The model is depicted graphically in Figure 1.

Each subject completed a life history (L.H.) or biodata form designed to assess the major dimensions of his or her development. The matrix of item intercorrelations on this inventory is then factored and each subject is characterized by a profile of his or her component scores. The $D^2$ statistic (Cronbach & Gleser, 1953; Osgood & Suci, 1952) is then employed to express interprofile similarity in terms of a matrix of distances between subjects. Finally, the Ward and Hook (1963) hierarchical grouping program is used to cluster profiles into the optimal number of subgroups or families (A through E in the diagram).

After the subgroups are defined, it is necessary to determine whether their characteristics are stable when assessed by methods other than self-report biodata. Thus the marker variables ($X_1$ through $X_n$) represent a broad spectrum of tests, such as personality and interest inventories, that aid in explicating subgroup characteristics as well as determining whether


Concurrent Data
L.H.defined subgroups
Future Data 4, College Years
Feedback definition of subgroups
A
B
C
D
E
$X_1$
$X_2$
$X_3$
$X_4$
$X_5$
$X_6$
$X_n$
Total Sample (all given L.H.)
Experiments and Field Studies in Psychology and the Behavioral Sciences
Historic Data
—=Ss
---=Information
Marker Variables
Feedback definition of subgroups

L.H. = Life History

Figure 1. The developmental-integrative model (From "Toward One Discipline of Scientific Psychology" by W.A. Owens, 1968, *American Psychologist, 23*, p. 783. Copyright 1968 by *American Psychologist*. Reprinted by permission.)

there are real differences between them. In addition, the experiments and field studies shown at the far right in the model serve a similar purpose in providing feedback of the effectiveness of the subgrouping. Except that the primary unit of study is the subgroup, this procedure is highly similar to that described by Cronbach and Meehl (1955) for construct validity, involving an accumulation of evidence from a number of different sources.

Owens and Schoenfeldt (1979) listed three reasons why biodata appears to be an optimal vehicle for forming the homogeneous subgroups. First, biodata instruments tend to have high criterion-related validity; the extensive evidence for this conclusion is documented later in this report. Second, since biodata factors tend to be orthogonal, their validity is largely differential validity which is very helpful in classification. Third, since the model is essentially a formalization of the "past predicts present" axiom, biodata devices provide broad, economical, and standardized measures of the past that provide data amenable to subgrouping.

Implications and Applications. Owens and Schoenfeldt (1979) listed several possible applications for a well-validated D-I model, largely resulting from the efficiency of making numerous attributions about an individual after completion of just one brief inventory. They note that such attributions would be probabilistic, depending on the accuracy of subgroup classification and the generalizability of previous findings about subgroup membership to new and different settings. With respect to the accuracy of classification, the authors reported that shrinkage in the fit of subjects to subgroups amounted to less than 8% in slotting the University of Georgia freshmen of subsequent years into the subgroups as defined in 1968. Given that some of this shrinkage may be due to true yearly variation in the student population, subgroup classification appears quite accurate.

The generalizability question may also be less intractable than it might appear, assuming, of course, that previous findings may be considered reliable. Much of the work with validity generalization (Schmidt & Hunter, 1977) suggests that previous assumptions about the situational specificity of validity results are more the result of statistical artifacts and sampling error than real differences in predictive effectiveness across settings. Brown (1981), however, using a BIB rather than the cognitive tests normally employed in these studies, showed that the overall validity of a BIB can be moderated by company differences.

Thus, it may be that stable and well-validated subgroups of people can be formed on the basis of biographical data. Such procedures would have wide applications and implications in the study of human behavior. In the area of personnel research, obtaining relationships of subgroups with criteria would be very helpful for selection purposes, and of even more utility in classification.

An earlier study supporting this assertion was conducted by Lunneborg (1968), who found that biodata's effectiveness in absolute prediction was greatly outweighed by its utility for differential prediction, the opposite of what was found for cognitive variables. Since classification largely involves differential prediction, Lunneborg's results anticipate later work with the D-I model's classification of subgroups.

Among the first practical applications of the subgrouping methodology was the work of Strimbu and Schoenfeldt (1973), who investigated its utility in predicting drug usage patterns and attitudes. Drawing on the data base accumulated by Owens and his colleagues at the University of Georgia, these authors selected a sample of 1,004 males and administered a drug usage and attitude questionnaire. Each member of the sample clearly belonged to only one subgroup, thus allowing subgroup and biodata factor comparisons of users versus non-users. Results showed that subgroup membership was significantly related to the criterion, and that the biodata factors associated with subgroups were meaningful in interpreting drug abuse corollaries.

Schoenfeldt (1974) proposed a systems model of classification that takes into account both individual and job characteristics. As with the D-I model, individuals are placed into subgroups that are homogeneous with respect to past behavior. Similarly, jobs are classified into families that are homogeneous with respect to their psychological requisites. To validate the model, Schoenfeldt used the subgroups available from the 1968 University of Georgia freshman class and formed job families based on the number of college courses completed in the ensuing 4 years in each of 12 homogeneous areas. The subgroups were compared across each of the job family profiles and on other academic criteria, such as grade point average.

Results showed that the subgroups differed both significantly and meaningfully in the great majority of cases. In addition, the findings were cross-validated with the 1970 class and held up well. Schoenfeldt remarked that the importance of this approach is that it acknowledges that most people can do many (if not most) of the jobs that need doing; classification is simplified and enhanced, however, by using homogeneous subgroups and job families that have much in common.

Morrison (1977) tested a simpler version of the Schoenfeldt procedure in an industrial setting by factor analyzing the responses of workers to a BIB. He formed two job families and employed discriminant analysis to identify the factors that best differentiated between the families. The linear function derived from the validation group was then used to predict job family membership in the cross-validation sample. Results showed that three of the eight obtained factors significantly differentiated between the job families, correctly classifying 70% of the workers. Although this was only a 10% improvement over base rate, Morrison concluded that the discriminant analysis appeared to be useful for decision making.

The next development was that of Brush and Owens (1979), who blended the more elaborate design used by Schoenfeldt (1974) with an industrial setting similar to that of Morrison (1977). Homogeneous subgroups of employees were formed based on a factor analysis of responses to a BIB, and the assignment of each individual to his or her most appropriate subgroup. Cross-validation did not shrink the percentage of subjects fitting the structure. Jobs were clustered based on the similarity of profiles that were deriveo from an examination of the distributions of job analysis variables. Finally, the subgroups were compared on six demographic and industrial criteria, and the overlap between subgroups and job families was computed.

Significant relationships were found with all criteria except race, and the subgroups frequently exceeded the base rate of membership in a job family by 15 to 25%. Base rate in a job family was equal to the percentage of the total sample expected to belong to a particular family irrespective of such factors as selection and classification. Thus, for 11 of the 18 subgroups, a consideration of subgroup membership yielded a significantly different prediction of job family membership than base rate alone. Further, an examination of the subgrouping classification showed that it was meaningful and suggested important ways in which the match between persons and jobs could be improved.

Brush and Owens argued that the combination of subgrouping persons and clustering jobs has demonstrated utility for job descrip'ion and classification, personnel selection and placement, manpower allocation and planning, and management information systems. They wrote: "Rather than selecting the right person for the right job vis-a-vis a traditional approach, emphasis is placed upon classifying applicants to that job family for which they have the necessary psychological requisites" (p. 381). Lastly, they saw the approach as more humanistic, by reducing the possibility of keeping capable people out of the organization and maximizing opportunities for available personnel.

Two additional studies have attempted to increase conceptual understanding of biodata through examining its relationship with vocational interests. Eberhardt and Muchinsky (1982a) investigated the differences on biodata factors of groups formed on the basis of Holland's hexagonal interest classification. Their results showed significant analysis of variance group effects for approximately one-third of the biodata factors for both sexes. In addition, results from regression analyses and multiple discriminant analyses of the data supported the hypothesis that a person's optimal vocational type is shaped to some extent by his or her life-history experiences.

In a related study, Mumford and Owens (1982) computed correlations between the items of the *Biographical Questionnaire* and total score on the scales of the 1968 version of the *Strong-Vocational Interest Blank*. Significant item-scale correlations were clustered on the basis of similarity in item content, after which the nonsignificant items were similarly assigned to a cluster. Finally, the percentage of "significant" items within a cluster was computed to determine the significance of the results for each cluster. Although the authors conceded that interpretation of these data is complex, they believed their results support the generalization that life history is closely, and perhaps causally, related to vocational interests.

Thus, both of these studies suggest that biodata may be conceptualized in a way similar to that of Owens (1976) who ". . . proposes that biodata be regarded as providing a postmortem view of the development of the individual--an inverted pyramid of a series of life events" (p. 625). Hence, a conceptualization of biodata is valuable not only in the "past predicts present" sense, but also in gaining a fairly standardized understanding of how an individual has developed and of the occupations and activities for which he or she is best suited.

Summary. Although a lack of conceptual development has been the primary criticism of the biodata area, some progress has been made, particularly by Owens, Schoenfeldt, and colleagues at the University of Georgia. There appears to be good agreement on the factor structure of the *Biographical Questionnaire* with college students, but whether similar results would be obtained with other instruments and populations is largely unresearched. The developmental-integrative model appears promising for both selection and classification, especially when combined with job family grouping, but it also is relatively unresearched. Given the known predictive validity of biodata and its intuitive appeal as a tool for greater understanding of individuals, further conceptually oriented research would seem to be of high potential in explicating how life history is related to current behaviors and capacities.

## Validity Research

In comparison with other measures of individual differences in human attributes, research on biodata validity has been somewhat narrow in emphasis. The overwhelming proportion of studies have been of criterion-related validity, with relatively little research on construct validity. This research emphasis may well be related to the heterogeneous constructs implicit within many biodata items, since life-history events often involve the confluence of several attributes that are conceptually distinct yet difficult to disentangle.

While this heterogeneity may impede explication of one specific construct, it is likely related to the efficacy of biodata in criterion-related validity studies. For example, an individual's high school grade point average is an excellent predictor of later performance in college, probably because the item implicitly includes a simultaneous assessment of motivation, emotional stability, academic interest, and cognitive ability, as well as other quite distinct constructs. Since each of these attributes is likely related to later academic performance, one might expect high criterion-related validity as well as problems in assigning the item to only one construct of interest. The nature of the domain's content may thus have much to do with the relative preponderance of the criterion-related study, and the ensuing discussion is a reflection of this state of affairs. We begin with a discussion of content validity.

## Content Validity

Content validity is generally thought of as the degree to which the content and format of a test correspond to a domain of relatively clear-cut knowledge or behavior (Guion, 1976). An assessment of the content validity of a BIB primarily involves three separate questions (Lammlein, 1982). First, are all or most of the important areas of knowledge or behavior in the domain represented on the test? Second, given that they are represented, do these important areas receive an appropriate amount of emphasis on the test? These first two questions involve, respectively, the ideas of representativeness and fidelity (Fitzpatrick & Morrison, 1971). Finally, is the item endorsement rate of the sample (or test) comparable to that of the domain?

It can be seen that an evaluation of content validity is a rational process directed primarily toward inventory construction. The author of each biographical inventory must determine the representativeness, fidelity, and endorsement rates of the form's items in depicting the desired domain. While these are difficult concerns for any inventory, they may be especially problematic for biodata since it is less well understood conceptually and defining the domain is more difficult. Thus, a BIB author who wishes to claim content validity must carefully document the steps taken in inventory construction to assure that the sample of selected items meets the appropriate content validity criteria. For a given job or job family, a job analysis would be instrumental in this process, since it defines the domain and establishes the importance of each aspect of the job. Unfortunately, a systematic description of how a biographical inventory has been developed with content validity concerns in mind has rarely, if ever, been reported in the literature.

An aspect of content validity that has received more attention in the literature is the identification of items that tap things outside the domain of interest--that is, are content invalid. As alluded to earlier, this has been of special concern when a blindly empirical approach was used in test construction, sometimes resulting in inventory items that do not appear to be representative of or even related to the domain in question. If such items are to be included in the inventory, they must be justified on grounds other than content validity. It may be preferable to develop alternate items that contribute to both content and criterion-related validity.

## Criterion-Related Validity

As noted earlier, criterion-related validity has been the forte of the biodata area, with numerous studies of both predictive and concurrent text presentation loses a degree of precision and detail in exchange for the simplicity of an overall statistic. Thus, the interested reader may wish to examine the studies individually in the appendix for more complete information. Also, it is noted that for the 70% of the investigations that included cross-validation, only the cross-validated correlations are reported.

The summary correlations obtained in this review are somewhat lower than those found in previous reviews. Three aspects of the present data summarization procedure may account for this state of affairs. First, when information was initially assembled, investigations with negative or modest results were included even if they were preliminary studies from which overall high validities might not be expected.

Second, since a number of investigations reported correlations between single items and the criterion, the findings could be reported based on all of the obtained correlations or on only those for which a reasonable relationship might be expected. Ghiselli (1966, p. 21) stated that for non-cognitive predictors, the only correlations reviewed were those in which the trait tapped by the predictor appeared relevant to the job. In the present review, however, all correlations were considered rather than allow the possibility that the results might influence selection of the "relevant" items.

Third, unlike reviews such as those by Ghiselli (1966) and Reilly and Chao (1982) that reported an average correlation, this review reports medians. Since the highest correlations typically depart more from the median than the lowest, an overall median is often slightly lower than an overall average.

It is recognized that the summary correlations reported will be somewhat attenuated as a function of the above procedures, and that they may well be spuriously low. On the other hand, these rather conservative estimates are perhaps more typical of biodata validity as it is found across the spectrum of its utilization. The summary correlation does not reflect the importance of carefully completed, iterative research that builds on previous findings and often yields substantially higher relationships with criteria. In fact, as Ghiselli (1966, 1973) has argued, the maximal validity of tests is of great interest since it illustrates what an inventory can do under optimal conditions. Maximal validities are listed in the "Correlation Range" category of the summary tables and discussed in the text as well.

Training Performance. This category includes validity studies that assessed the relationship between biodata and success in a training program. Measurement of training performance includes such subcategories as (a) objective measures such as test scores, (b) subjective measures such as instructor ratings, (c) a combination of these two, (d) completion versus noncompletion of the course (go/no-go), and (e) hands-on or work sample measures.

The results of the 18 correlational studies are summarized in Table 1; the studies and the criteria they address are listed in Appendix B. It can be seen that the overall median validity coefficient is about .25 across criterion subcategories. The overwhelming majority of these studies (16)

Table 1

Criterion-Related Validity: Training

| | Criterion Subcategory | | | | |
|---|---|---|---|---|---|
| | Objective Measures | Subjective Measures | Combination | Go/No-Go Course Completion | Hands-On Measures |
| Number of Studies | 6 | 2 | 2 | 8 | 0 |
| Median Correlation Overall | .23 | .33 | .11 | .25 | -- |
| Predictive | .38 | .33 | .11 | .28 | -- |
| Concurrent | .07 (1 study) | | | .25 | -- |
| Correlation Range | .00 - .59 | .11 - .47 | .01 - .47 | .00 - .55 | -- |
| *N* Range | 114 - 7,929 | 140 - 2,212 | 134 - 168 | 115 - 4,502 | -- |
| Median *N* | 519 | 1,036 | 151 | 438 | -- |

were conducted with military personnel, and the overall median validity is about the same (.25) when the military studies are considered separately. More than half of these coefficients were cross-validated and most were based on substantial sample sizes. None of these studies used hands-on measures.

In terms of the maximal validity that can be expected from biodata for training criteria, it is noted that six of the 20 studies reported median correlations above .38. Four of these were cross-validated (including one of .57), the median sample size was just under 400, and all six studies were predictive rather than concurrent. Five of the six studies were conducted with military samples. Thus, it is quite well replicated that very good validities can be obtained by using background information to predict future training performance, particularly in the military.

In addition to the studies that reported results in terms of a correlation coefficient, several other investigations appear relevant. Eaton, Weltin, and Wing (1982) used a chi-square analysis to show the significant relationship between scores on the *Military Applicant Profile* and early attrition from the Army (while the recruit is undergoing training). Similar results have been found in the Navy (Booth & Berry, 1978) and in the Air Force (Lachar, Sparks, Larsen & Bisbee, 1974). Also, a study by Webster, Booth, Graham, and Alf (1978) is a good example of biodata combining with other types of measures, in this case producing a cross-validated coefficient of .47 with training completion in a multiple correlation of biodata, temperament, interests, and aptitude variables.

Another body of literature that is related to that of training performance is the research with educational or academic performance. Biodata has been found to be an outstanding predictor in this area; in fact, high school grade point average is consistently the best single indicator of later college performance (Freeberg, 1967). A number of studies (Farr, O'Leary, Pfeiffer, Goldstein, & Bartlett, 1971; IBRIC, 1968) have obtained median validity coefficients in the .60s between a BIB and educational criteria. As noted by Asher (1972), the fact that previous academic performance is highly related to future academic performance is intuitive, and probably reflects a greater point-to-point correspondence between predictor space and criterion space than is typically found in validity studies.

Job Proficiency. This category includes studies of the relationship between biodata and measures of job performance, such as supervisory or peer rankings or ratings, job knowledge tests, and such archival production indicators as units produced and salary or organizational level. A total of 36 correlational studies related to job proficiency were located and are summarized in Table 2 and listed individually in Appendix B. None of these studies utilized job knowledge tests.

Again, similar results are observed across subcategories with the overall median validity at about .32. Validities were notably higher for predictive than concurrent studies, which is noteworthy since numbers of each type of investigation were approximately equal. Eight of these studies employed military samples and yielded results very similar to the overall predictive validity value (.20). About two-thirds of the studies were cross-validated, and again, most studies included fairly large samples.

Table 2

Criterion-Related Validity: Job Proficiency

| | Criterion Subcategory | | |
|---|---|---|---|
| | Ratings & Rankings | Archival Production | Job Knowledge Tests |
| Number of Studies | 26 | 10 | 0 |
| Median Correlation Overall | .32 | .31 | -- |
| Predictive | .18 | .20 | -- |
| Concurrent | .38 | .42 | -- |
| Correlation Range | .00 - .60 | .00 - .70 | -- |
| *N* Range | 100 - 3,964 | 100 - 12,453 | -- |
| Median *N* | 294 | 800 | -- |

Nearly half of the reviewed investigations yielded median values at or above .40, demonstrating extensive evidence of the potential validity of biodata instruments for measuring job proficiency. The largest validity obtained was a cross-validated, predictive coefficient of .70 achieved with NASA scientists (Taylor & Ellison, 1967), and seven studies reported coefficients in excess of .50. Ten of the 14 studies with validities above .40 included cross-validation and their median *N* was 226, again adding confidence to the robustness of the findings.

In research other than that reported by correlations, Hinrichs (1960) found that high performers scored higher on the application blank than other employees. Tanofsky, Shepps, and O'Neil (1969) showed via pattern analysis that biodata was predictive of insurance sales performance, a finding that has been extensively documented by some 50 years of LIMRA research (LIMRA, 1979; Thayer, 1977). Kavanaugh and York (1972) used analysis of variance and found 24 of 41 biodata items to be significantly related to job performance ratings. Also, Manyak (1975) demonstrated that profiles developed on the basis of biographical information successfully discriminated between police officers with high and low supervisory ratings.

As was mentioned earlier, a handful of biodata items are often collected along with another measure and will in many cases obtain quite high relationships with job proficiency. Among the predictors that have often demonstrated this relationship are level of education (cf. Loudermilk, 1966), length of service or experience (cf. Shanthemai, 1978), and previous academic achievement (cf. Wise, 1975). Although these items are nearly ubiquitous in the selection process, a review by Caplan and Schmidt (1977) concluded that education and experience have zero or near-zero validity overall. This conclusion notwithstanding, the overwhelming weight of evidence in military research shows such application blank-type items to be highly valid (cf. Hoiberg & Pugh, 1978).

Job Involvement. The job involvement category includes those studies that have examined biodata's relationship with job satisfaction, absenteeism and turnover, and tenure. Results of the 33 studies in this category are summarized in Table 3 and listed in detail in Appendix B.

Table 3

Criterion-Related Validity: Job Involvement

| | Criterion Subcategory | | |
|---|---|---|---|
| | Job Satisfaction | Absenteeism & Turnover | Tenure |
| Number of Studies | 0 | 15 | 18 |
| Median Correlation Overall | -- | .25 | .32 |
| Predictive | -- | .25 | .31 |
| Concurrent | -- | .33 | .56 |
| Correlation Range | -- | .00 - .60 | .05 - .74 |
| *N* Range | -- | 105 - 20,000+ | 74 - 14,738 |
| Median *N* | -- | 702 | 150 |

The overall median validity coefficient is about .30. Again, the concurrent studies obtained validities that are higher than those of predictive investigations. Eighty-four percent of the studies in this category were cross-validated, although the sample sizes, particularly for the tenure subcategory, were somewhat smaller. The nine studies that employed military samples all demonstrated good prediction, and validities ranged from .15 to .41 with a median value of .21.

In considering maximal validity, again nearly half of the studies reported median values above .40. An investigation by Federico, Federico and Lundquist (1976) obtained the highest validity, a cross-validated, predictive correlation of .74 with clerical workers. Nine studies reported values in excess of .50, and all but one of these "maximal validity" studies was cross-validated, as well as having a respectable median sample of 218.

In other research, a number of early investigations demonstrated that educational level is highly related to attrition from the military (Gordon & Bottenberg, 1962; Plag & Hardacre, 1964), a finding that has been widely replicated. More recently, a large-scale, chi-square analysis of the *Military Applicant Profile* (MAP), a 60-item biodata form, showed that MAP scores are highly related to Army attrition, and the relationship is moderated by whether the individual is a high school graduate or not (Eaton et al., 1982). Robinson's (1972) cross-validated investigation found that biodata identified 80% of the short-tenure employees at the expense of rejecting 30% of the long-tenure workers. Other studies demonstrating the validity of biodata with tenure include Driscoll (1974), and the extensive research with the *Aptitude Index Battery* in the life insurance industry (LIMRA, 1979).

Adjustment. The adjustment category includes studies with the criteria of substance abuse and various other delinquent behaviors related to disciplinary action or unfavorable discharge from the military. Only six correlational studies (listed in Appendix B) were located for these criteria, with the results summarized in Table 4.

Table 4

Criterion-Related Validity: Adjustment

| | Criterion Subcategory | | |
|---|---|---|---|
| | Substance Abuse | Delinquency | Unfavorable Discharge |
| Number of Studies | 1 | 3 | 2 |
| Median Correlation Overall | .26 | .20 | .27 |
| Predictive | -- | .14 | .27 |
| Concurrent | .26 | .33 | -- |
| Correlation Range | .16 - .39 | .00 - .63 | .07 - .36 |
| *N* Range | 2,043 | 100 - 6,185 | 6,455 - 15,252 |
| Median *N* | 2,043 | 134 | 10,854 |

The overall median coefficient is .26, three of the studies are cross-validated, and four studies had large samples. Of special note in these investigations is the work done in the Air Force with the *History and Opinion Inventory* (HOI) and its drug abuse and emotional instability subscales. As elaborated by Bloom (1977), the HOI has been used as part of a screening procedure that first identifies those recruits who are at risk for emotional problems in basic training, and then subjects them to further evaluation. Bloom (personal communication, 1983) reported that the procedure has been effective, not only in identifying those with significant adjustment problems, but in reducing delinquency problems and suicide.

In other studies, Plag (1962) found that a number of biodata items were effective in discriminating those with poor potential for adjustment to the Navy, a finding he later replicated (1969). Plag and Goffman (1973) obtained similar findings for drug abuse in the Navy, and Jessor, Jessor, and Finney (1973) found biodata items to be related to marijuana use among high school and college students. Life history events were shown to be important indicators of the transition from non-use to use of marijuana in a longitudinal study by Jessor (1976), with similar results reported for alcohol abuse by Jessor and Jessor (1975). Also, Kantor and Guinn (1975) found that, in their sample of 20,000 Air Force enlistees, high school graduates had significantly fewer disciplinary actions and unsuitability discharges than nongraduates.

Typical of these studies' findings is the following conclusion from Nail, Gunderson, Kolb, and Butler (1975, p.177) about Navy enlisted drug abusers: "Many in this drug abuser population reported problems with school and legal authorities prior to entering Naval service. Of the total sample, 30% had been expelled or suspended from high school more than once, and 38% had been arrested for other than traffic offenses. Only about half of the sample had completed high school." Finally, Kamp's (1983) review found that 17 of 20 reported relationships in 10 studies showed a significant negative relationship between religiosity (e.g., church attendance per year) and amount of alcohol and drug use in high school and college samples.

### Validity Issues

One of the principal criticisms of the use of biodata as a predictor has been the tendency for its validity to erode over time. Hughes, Dunn, and Baxter (1956) were among the first to document this phenomenon, observing a steady decline in the validity of a WAB over a 3-year period. Similarly, Dunnette et al. (1960) witnessed a decline in a WAB's validity from .74 to .38 over a 2-year period. Wernimont (1962) reported that the same form's validity had dropped further to .07 after 5 years, although reweighting it resulted in a correlation of .39. More recently, Brown (1978) reported the validity, for life insurance agents hired in 1939 and also between 1969 and 1971, of a biodata form originally validated in 1933. Despite these intervals of 6 and 38 years and a probable increase in restriction of range, little, if any, validity was lost over that originally obtained.

Brown (1978) offered two explanations to account for the discrepancies between his results and those found earlier. First, the contents of the scoring were confidential in the Brown study, thus preventing managers from manipulating the scoring keys such that applicants would pile up in the favorable categories. Hughes et al. (1956) considered this manipulation to be the major reason for the drop in validity in their study. Somewhat related is the "in-use" phenomenon described by Peterson and Wallace (1966) in which the use of the BIB as part of the selection procedure resulted in restriction of range and reduction in validity that would not occur if that information was unavailable for hiring decisions.

A second possible explanation stems from the very large development sample size (10,111) in the Brown study that resulted in a 421:1 subject to variable ratio. Such ratios are much more likely to yield stable results than those found with the smaller samples of earlier studies. Another possible source of validity distortion is that managers may make a special effort in training and supervising those new employees with relatively low scores on the inventory, resulting in what appears to be either zero validity or a curvilinear relationship (Brown, 1979).

To summarize, it seems clear that there are several threats to obtaining stable validities with biodata instruments, and that large development sample sizes and confidentiality of scoring offer partial solutions. All writers seem to agree, however, that periodic reweighting and rescaling of items is necessary in addition (cf. Thayer, 1977). Only through such scoring key maintenance and attention to other distorting influences can optimal validities be maintained.

Another issue related to the stability of biodata is its validity after cross-validation, or the amount of shrinkage typically found. Schwab and Oliver (1974) reported that, after they cross-validated four unpublished studies with validities ranging from .37 to .66, the correlations shrank to near-zero levels. They argued that this type of finding is actually fairly common but is suppressed because only those studies obtaining significant cross-validities are published in the literature. Although Schwab and Oliver raise an important point, there is reason to believe the situation at present is much less serious than they depict. Cross-validation has come to be expected of all correlational studies, and reviews of biodata validity such as England (1971), Reilly and Chao (1982), and the present one all have found extensive evidence of validity despite the near ubiquity of cross-validation. A more plausible explanation for the findings of Schwab and Oliver and others, is that often the sample sizes were quite small. Small samples, combined with a blindly empirical construction technique, are likely to yield substantial shrinkage. Those studies with larger samples often see very little loss in validity even with a double cross-validation approach (cf. Webster et al., 1978).

Several investigations have examined the robustness of biodata across national borders. Cassens (1966) concluded that although specific life history behaviors differ for workers in North and Latin America, the underlying factor dimensions were similar across cultures. Laurent (1970) computed the correlations between scores on a BIB and a composite criterion of management success for managers in Norway, Denmark, the Netherlands, and the United States and reported that validities in all samples were similar, with a median value of .60. Similar results were reported by Hinrichs, Haanpera, and Sonkin (1976), who found a median validity coefficient of .42 and a moderate degree of consistency across samples of salesmen from five European countries and America. Finally, Nevo (1976) found a brief BIB valid in predicting military rank for Israeli soldiers.

A final issue to be discussed is that of differential biodata validity for members of different subgroups. With respect to race, there appears to be little difference in the validities obtained between whites and minority group members. For example, Cascio (1976) found coefficients of .58 for minorities and .56 for nonminorities in a study of employee turnover. Frank and Erwin's (1978) large-scale investigation of early attrition from the Army showed no important differences in validity between blacks and whites. Similarly, the research at LIMRA has consistently demonstrated comparable validity for all ethnic groups (LIMRA, 1979). Thus, although group differences may be found on some items (cf. Booth & Berry, 1978), the validity of biodata appears quite comparable for minority and majority groups.

Results are similar in studies that compared biodata validity for males and females, although the group differences may be somewhat larger. Nevo (1976) found that different items were valid for males and females and obtained a validity of .36 for males and .18 for females, but these estimates may be spurious because of differential restriction of range in the criterion. Webster et al. (1978) found coefficients of .53 for men and .41 for women, but again this was not a fair comparison since aptitude and temperament variables were also included. Higher validities were obtained for women (.40) than men (.35) in a study by Ritchie and Boehn (1977) of ratings of management potential. Booth et al. (1978) reported that neither sex nor minority status affected validity in their study of training performance. LIMRA research (1979) also reported little or no differential validity by sex although separate sex scales are used. Finally, a study by Wilcove, Thomas, and Blankenship (1979) of the attrition of female personnel from the Navy obtained results comparable to those found with males. Since factor analysis has demonstrated stable differences between the sexes on biodata items (Eberhardt & Muchinsky, 1982b; Owens & Schoenfeldt, 1979), it may be that comparable validities for the sexes will often require different scales. This is also the conclusion reached by Reilly and Chao (1982) in their review.

Summary. Validity evidence reviewed in this paper strongly supports the ability of biodata measures to predict important job criteria. Median validities in the .20s and .30s were obtained for training, job performance, job involvement, and adjustment criteria. Results are based on over 100 studies, many of which involved large samples and most of which were cross-validated. The findings of this review thus support previous reviews that showed biodata to be among the best predictors available of a range of criteria.

Relatively little evidence is available concerning the content and construct validity of biodata measures, and research in these areas is clearly needed. The tendency for biodata validities to erode over time appears to be related to blind empirical weighting, small sample sizes, and manipulation of scoring keys, although periodic reweighting is advisable--even when these factors are controlled. Finally, biodata validity appears to be robust across settings and populations; separate scales may, however, be necessary for males and females.

## *Section 2 References*

Albemarle Paper Company v. Moody. *U.S. Reports, 422*, 405 (1975).

Anonymous. (1925). A method of rating the history and achievements of applicants for positions. *Public Personnel Studies, 3* (7), 202-209. (Methods and data credited to Gertrude V. Cope).

Asher, J. J. (1972). The biographical item: Can it be improved? *Personnel Psychology, 25*, 251-269.

Baehr, M., & Williams, G. B. (1967). Underlying dimensions of personal background data and their relationship to occupational classification. *Journal of Applied Psychology, 51*, 481-90.

Baehr, M. E., & Williams, G. B. (1968). Prediction of sales success from factorially determined dimensions of personal background data. *Journal of Applied Psychology, 52* (2), 98-103.

Berkhouse, R. G., & Cook, K. G. (1961). *Development of preliminary screening measures for special forces trainees* (Research Memorandum 61-7). Washington, DC: Army Behavioral Evaluation Research Laboratory.

Biersner, R. J., & Ryman, D. H. (1974). Prediction of scuba training performance. *Journal of Applied Psychology*, 59 (4), 519-521.

Black, B. B., & McKinney, A. C. (1963). Validity Information Exchange, No. 16-03. *Personnel Psychology, 16*, 173-180.

Bloom, W. (1977). Air Force Medical Evaluation Test. *Medical Service Digest, USAF, 28* (2), 1-3.

Booth, R. F., & Berry, N. H. (1978). Minority group differences in the background, personality and performance of Navy paramedical personnel. *Journal of Community Psychology, 6*, 60-72.

Booth, R. E., McNally, M. S., & Berry, N. H. (1978). Predicting performance effectiveness in paramedical occupations. *Personnel Psychology, 31*, 581-593.

Brown, S. H. (1978). Long-term validity of a personal history item scoring procedure. *Journal of Applied Psychology, 63* (6), 673-676.

Brown, S. H. (1979). Validity distortions associated with a test in use. *Journal of Applied Psychology, 64* (4), 460-462.

Brown, S. H. (1981). Validity generalization and situational moderation in the life insurance industry. *Journal of Applied Psychology, 66* (6), 664-670.

Brush, D. H., & Owens, W. A. (1979). Implementation and evaluation of an assessment classification model for manpower utilization. *Personnel Psychology, 32*, 369-383.

Bucky, S. F., Edwards, D., & Thomas, E. D. (1974). Intensity: The description of a realistic measure of drug use. *Journal of Clinical Psychology, 30*, 161-163.

Buel, W. D. (1964). Voluntary female clerical turnover: The concurrent and predictive validity of a weighted application blank. *Journal of Applied Psychology, 48*, 180-182.

Buel, W. D. (1965). Biographical data and the identification of creative research personnel. *Journal of Applied Psychology, 49*, 318-321.

Buel, W. D. (1966). A note on the generality and cross-validity of personal history for identifying creative research scientists. *Journal of Applied Psychology, 50*, 217-219.

Caplan, J. R., & Schmidt, F. L. (1977, April). *The validity of education and experience ratings.* Presented to International Personnel Management Association Assessment Council.

Cascio, W. F. (1975). Accuracy of verifiable biographical information blank responses. *Journal of Applied Psychology, 60* (6), 767-769.

Cascio, W. F. (1976). Turnover, biographical data, and fair employment practice. *Journal of Applied Psychology, 61* (5), 576-580.

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

Cassens, F. P. (1966). *Cross-cultural dimensions of executive life history antecedents.* Greensboro, NC: The Creativity Research Institute, The Richardson Foundation.

Chaney, F. B., & Owens, W. A. (1964). Life history antecedents of sales, research, and general engineering interest. *Journal of Applied Psychology, 48*, 101-105.

Clark, K. E. (1961). *Vocational interests of non-professional men.* Minneapolis: University of Minnesota Press.

Colson, K. R. (1977). *he predictive validity of a life experience inventory for the identification of creative scientists and engineers.* Unpublished doctoral dissertation, University of Southern California.

Conrad, H. S., & Ellis, A. (1948). The validity of personality inventories in military practice. *Psychological Bulletin, 45*, 385-426.

Cornelius, E. T. III. (1977, October). *The development and validation of a foreman selection test battery.* Unpublished report, Ethel Corporation.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671-684.

Cronbach, L. J., & Gleser, G. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456-473.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Dann, J. E., & Abrahams, N. M. (1969). *Validation of a biographical information blank as a predictor of retention among mechanical and electrical-electronics enlisted personnel* (SRM 69-21). San Diego, CA: Navy Personnel Research and Development Center.

Dann, J. E., & Abrahams, N. M. (1970). *Use of biographical and interview information in predicting Naval Academy disenrollment* (Research Report SRR 71-7). San Diego, CA: Naval Personnel and Training Research Laboratory.

Driscoll, P. J. (1974, May-June). Employee turnover in the die casting industry . . . A case history. *Die Casting Engineer*, 38-39.

DuBois, P. H., Loevinger, J., & Gleser, G. C. (1952). The construction of homogeneous keys for a biographical inventory. *USAF Human Resources Research Bulletin*, 52-58.

Dunnette, M. D. Personnel Management. (1962). *Annual Review of Psychology, 13*, 285-314.

Dunnette, M. D., Kirchner, W. K., Erickson, J., & Banas, P. (1960). Predicting turnover among female office workers. *Personnel Administration, 23*, 45-50.

Dyer, E. D., Cope, M. J., Monson, M. A., & Drimmelen, J. B. V. (1972). Can job performance be predicted from biographical, personality, and administrative climate inventories? *Nursing Research, 21* (4), 294-304.

Eaton, N. K., Weltin, M., & Wing, H. (1982, December). *Validity of the Military Applicant Profile (MAP) for predicting early attrition in different educational, age and racial groups.* Army Research Institute for the Behavioral and Social Sciences, Technical Report, Alexandria, VA.

Eberhardt, B. J., & Muchinsky, P. M. (1982a). Biodata determinants of vocational typology: An integration of two paradigms. *Journal of Applied Psychology, 67* (6), 714-727.

Eberhardt, B. J., & Muchinsky, P. M. (1982b). An empirical investigation of the factor stability of Owens' Biographical Questionnaire. *Journal of Applied Psychology, 67* (2), 138-145.

Ehrle, R. A. (1964). Quantification of biographical data for predicting vocational rehabilitation success. *Journal of Applied Psychology, 48*, 171-174.

Ellison, R. L., James, L. R., & Carron, T. J. (1970). Prediction of R & D performance criteria with biographical information. *Journal of Industrial Psychology, 5*, 37-57.

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

England, G. W. (1971). *Development and use of weighted application blanks*. Bulletin 55. Minneapolis, MN: Industrial Relations Center, University of Minnesota.

Erwin, F. W., & Herring, J. W. (1977, August). *The feasibility of the use of autobiographical information as a predictor of early Army attrition* (TR-77-A6). Alexandria, VA: U.S. Army Research Institute for Behavioral and Social Sciences.

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employmee selection procedures. *Federal Register, 43*, 38290-38315.

Farr, J. L., O'Leary, B. S., Pfeiffer, C. M., Goldstein, I. L., & Bartlett, D. J. (1971, September). *Ethnic group membership as a moderator in the prediction of job performance: An examination of some less traditional predictors* (Technical Report No. II). Washington DC: American Institutes for Research.

Federico, S. M., Federico, P. A., & Lundquist, G. W. (1976). Predicting women's turnover as a function of net salary expectation and biodemographic data. *Personnel Psychology, 29*, 559-566.

Fitzpatrick, R., & Morrison, E. J. (1971). Performance and product evaluation. In R. L. Thorndike (Ed.), *Educational Measurement*. Washington, DC: American Council on Education.

Fleishman, E. A., & Berniger, J. (1960). One way to reduce office turnover. *Personnel, 37*, 63-69.

Frank, B. A., & Erwin, F. W. (1978). *The prediction of early Army attrition through the use of autobiographical information questionnaires* (Technical Report TR-78-A11). Alexandria, VA: U.S. Army Research Institute.

Freeberg, N. E. (1967). The biographical information blank as a predictor of student achievement: A review. *Psychological Reports, 20*, 911-925.

Gardner, E. E., & Williams, A. P. O. (1973). A twenty-five year follow-up of an extended interview selection procedure in the Royal Navy. *Occupational Psychology, 47*, 1-13.

Gebhardt, G. M. (1979, May). *An evaluation of the validity of a weighted application blank for predicting tenure of minority and nonminority employees*. Unpublished master's thesis, Southern Illinois University.

Ghiselli, E. E. (1966, 1973). *The measurement of occupational aptitude*. Berkeley, CA: University of California Press.

Goheen, H. W., & Mosel, J. N. (1950). *The accuracy of applications for civil service employment*. Unpublished study.

Goldsmith, D. B. (1922). The use of the personal history blank as a salesmanship test. *Journal of Applied Psychology, 6*, 149-155.

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

Goodenough, F. (1949). *Mental testing: Its history, principles, and applications*. New York: Holt, Rinehart and Winston.

Gordon, M. A., & Bottenberg, R. A. (1962, April). *Prediction of unfavorable discharge by separate educational levels* (Technical Report 62-5). Lackland Air Force Base, TX: 6570th Personnel Research Laboratory.

Griggs vs. Duke Power Company. *U.S. Reports, 401*, 424 (1971).

Guilford, J. P., & Lacey, J. I. (Eds.) (1947). Printed classification tests. *AAF Aviation Psychology Research Program Reports*. No. 5. Washington, DC: Government Printing Office.

Guinn, N., Johnson, A. L., & Kantor, J. E. (1975). *Screening for adaptability to military service* (TR 75-30). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Guinn, N., Vitola, B. M., & Leisey, S. A. (1976). *Background and interest measures as predictors of success in undergraduate pilot training* (Technical Report 76-9). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Guinn, N., Wilbourn, J. M., & Kantor, J. E. (1977). Preliminary development and validation of a screening technique for entry into the security police career field. *Catalog of Selected Documents in Psychology, 7*, 122-123.

Guion, R. M. (1976). Recruiting, selection, and placement. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago, IL: Rand McNally.

Harrell, T. W., & Harrell, M. S. (1975, July). *A scale for high earners* (Technical Report N00014-67-A-0112-0073). Arlington, VA: Office of Naval Research.

Harrell, T. W., & Harrell, M. S. (1976, May). *Predictors of business managers' success at 10 years out of MBA* (Technical Report N00014-76-C-0009). Arlington, VA: Office of Naval Research.

Harrell, M. S., Harrell, T. W., McIntyre, S. H., & Weinberg, C. B. (1977). Predicting compensation among MBA graduates five and ten years after graduation. *Journal of Applied Psychology, 62* (5), 636-640.

Harris, H. J. (1946). The Cornell Selectee Index--an aid in psychiatric diagnosis. *Annals New York Academy of Science, 46*, 594-603.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different methods of deriving personality inventory scales. *Psychological Bulletin, 67*, 231-248.

Haymaker, J. C., & Erwin, F. W. (1980). *Investigation of applicant responses and falsification detection procedures for the MAP* (TR 1-80). Washington, DC: Richardson, Bellows and Henry, Inc.

Helmreich, R., Bakeman, R., & Radloff, R. (1973). The life history questionnaire as a predictor of performance in Navy diver training. *Journal of Applied Psychology, 57*, 148-153.

Henry, E. R. (Chrmn.) (1966). *Research conference on the use of autobiographical data as psychological predictors*. Greensboro, NC: The Creativity Research Institute, The Richardson Foundation.

Hinrichs, J. R. (1960, March-April). Technical selection: How to improve your batting average. *Personnel*, 56-60.

Hinrichs, J. R., Haanpera, S., & Sonkin, L. (1976). Validity of a biographical information blank across national borders. *Personnel Psychology*, 29, 417-421.

Hoiberg, A., Booth, R. F., & Berry, N. H. (1977). Non-cognitive variables related to performance in Navy "A" schools. *Psychological Reports*, 41, 647-655.

Hoiberg, A., & Pugh, W. M. (1978). Predicting Navy effectiveness: Expectations, motivation, personality, aptitude and background variables. *Personnel Psychology, 31*, 841-52.

Hughes, J. F., Dunn, J. F., & Baxter, B. (1956). The validity of selection instruments under operating conditions. *Personnel Psychology, 9*, 321-324.

Human Resources Research Organization. (1976). *Selection of qualified Army enlistees: Early development of the Military Aptitude Predictor*. U. S. Army Research Institute for the Behavioral and Social Sciences, SR-ED-76-19-2, Washington, DC.

Institute for Behavioral Research in Creativity (IBRIC). (1968). *Manual for Alpha biographical inventory*. Greensboro, NC: Predictions Press.

Jackson, D. N. (1971). The dynamics of structured personality tests: 1971. *Psychlogical Review, 78*, 229-248.

Jessor, R. (1976). Predicting time of onset of marijuana use: A developmental study of high school youth. *Journal of Consulting and Clinical Psychology, 44*, 125-134.

Jessor, R., & Jessor, S. L. (1975). Adolescent development and the onset of drinking: A longitudinal study. *Journal of Studies on Alcohol*, 36, 27-51.

Jessor, R., Jessor, S., & Finney, J. (1973). A social psychology of marijuana use: Longitudinal studies of high school and college youth. *Journal of Personality and Social Psychology, 26*, 1-15.

Johnson, D. A., Newton, N. W., & Peek, L. A. (1979, May-June). Predicting tenure of municipal clerical employees: A multiple regression analysis. *Public Personnel Management, 44*, 182-190.

Kamp, J. (1983, November). *The prediction of substance abuse*. In-house report. Minneapolis, MN: Personnel Decisions Research Institute.

Kantor, J. E., & Guinn, N. (1975). *Comparison of performance and career progression of high school graduates and non-graduates in the Air Force* (Technical Report 75-73). Lackland Air Force Base, TX: Air Force Human Resources Laboratory.

Kavanaugh, M. J., & York, D. R. (1972). Biographical correlates of middle managers' performance. *Personnel Psychology, 25*, 319-332.

Keating, E., Paterson, D. G., & Stone, C. H. (1950). Validity of work histories obtained by interview. *Journal of Applied Psychology, 34*, 1-5.

Kenagy, H. G., & Yoakum, C. S. (1925). *The selection and training of salesmen.* New York: McGraw-Hill.

Klimoski, R. J. (1973). A biographical data analysis of career patterns in engineering. *Journal of Vocational Behavior, 3*, 103-113.

Korman, A. K. (1968). The prediction of managerial performance: A review. *Personnel Psychology, 21*, 295-322.

Lachar, D., Sparks, J. C., Larsen, R. M., & Bisbee, C. J. (1974). Psychometric prediction of behavioral criteria of adaptation for U. S. Air Force basic trainees. *Journal of Community Psychology, 2* (3), 268-277.

Lammlein, S. (1982, June). *A proposal for the administration of the fundamentals of an engineering examination program.* Minneapolis, MN: Personnel Decisions Research Institute.

Laurent, H. (1970). Cross-cultural cross-validation of empirically validated tests. *Journal of Applied Psychology, 54*, 417-423.

Lee, R., & Booth, J. M. (1974). A utility analysis of a weighted application blank designed to predict turnover for clerical employees. *Journal of Applied Psychology, 59* (4), 516-518.

Lefkowitz, J. (1972). Differential validity: Ethnic group as a moderator in predicting tenure. *Personnel Psychology, 25*, 225-240.

Levine, A. S., & Zachert, V. (1951). Use of a biographical inventory in the Air Force classification program. *Journal of Applied Psychology, 35*, 241-244.

Life Insurance Marketing and Research Association, Inc. (1979). *Agent Selection Questionnaire Research.* Hartford, CT: LIMRA.

Lipsett, L. (1946). The personal investigation in the selection of employees. *Personnel Administration, 9*, 23-29.

Loevinger, J., Gleser, G. C., & DuBois, P. H. (1953). Maximizing the discriminating power of a multiple-score set. *Psychometrika, 18*, 309-317.

Long, J. A., & Sandiford, P. (1935). *The validation of test items* (Research Bulletin 3). Toronto, Canada: University of Toronto, Department of Education.

Loudermilk, K. M. (1966). Prediction of efficiency of lumber and paper mill employees. *Personnel Psychology, 19*, 301-310.

Lunneborg, C. E. (1968). Biographical variables in differential vs. absolute prediction. *Journal of Educational Measurement, 5* (3), 207-210.

Manese, W. R., Skrobiszewski, M. F., & Abrahams, N. M. (1976). *Selection criteria for recruit company commanders: Development and validation* (Research Report TR 77-0). San Diego, CA: Navy Personnel Research and Development Center.

Manson, G. E. (1925). What can the application blank tell? *Journal of Personnel Research, 4*, 73-99.

Manyak, T. C. (1975, April). *The use of background information in the police selection process*. Paper given at the 1975 National Conference on Public Administration, Chicago, IL.

Matteson, M. T. (1978). An alternative approach to using biographical data for predicting job success. *Journal of Occupational Psychology, 51*, 155-162.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Minneapolis, MN: University of Minnesota Press.

Miner, J. B. (1971). Success in management consulting and the concept of eliteness motivation. *Academy of Management Journal, 14* (3), 367-378.

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* (4), 411-418.

Moore, H. (1942). *Psychology for business and industry*. New York: McGraw-Hill.

Morrison, R. F. (1977). A multivariate model for the occupational placement decision. *Journal of Applied Psychology, 62*, 271-277.

Morrison, R. F., Owens, W. A., Glennon, J. R., & Albright, L. E. (1962). Factored life history antecedents of industrial research performance. *Journal of Applied Psychology, 46*, 281-284.

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

Mumford, M. D., & Owens, W. A. (1982). Life history and vocational interests. *Journal of Vocational Behavior, 21*, 330-348.

Nail, R. L., Gunderson, E. K., Kolb, D., & Butler, M. (1975). Drug histories of Navy amnesty cases. *Military Medicine, 140*, 172-178.

Neiner, A. G., & Owens, W. A. (1982). Relationships between two sets of biodata with seven years separation. *Journal of Applied Psychology, 67* (2), 146-150.

Neumann, I., Githens, W. H., & Abrahams, N. M. (1967). *The development of the U. S. Navy background questionnaire for NROTC (Regular) selection* (Research Report SRR 68-3). San Diego, CA: U.S. Naval Personnel Research Activity.

Nevo, B. (1976). Using biographical information to predict the success of men and women in the Army. *Journal of Applied Psychology, 61* (1), 106-108.

Osgood, C. E., & Suci, G. J. (1952). A measure of relation determined by both difference and profile information. *Psychological Bulletin, 49*, 251-262.

Owens, W. A. (1968). Toward one discipline of scientific psychology. *American Psychologist, 23*, 782-785.

Owens, W. A. (1971). A quasi-actuarial basis for individual assessment. *American Psychologist, 26*, 992-999.

Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally.

Owens, W. A., Glennon, J. R., & Albright, L. E. (1962). Retest consistency and the writing of life history items: A first step. *Journal of Applied Psychology, 46*, 329-332.

Owens, W. A., & Schoenfeldt, L. F. (1979). Towards a classification of persons. *Journal of Applied Psychology Monograph, 64*, 569-607.

Pace, L. A., & Schoenfeldt, L. F. (1977). Legal concerns in the use of weighted application blanks. *Personnel Psychology, 30*, 159-166.

Peterson, D. A., & Wallace, S. R. (1966). Validation and revision of a test in use. *Journal of Applied Psychology, 50*, 13-17.

Plag, J. A. (1962). Pre-enlistment variables related to the performance and adjustment of Navy recruits. *Journal of Clinical Psychology, 18* (2), 168-170.

Plag, J. A. (1969). *Predicting the military effectiveness of enlistees in the U. S. Navy* (Report No. 69-23). San Diego, CA: Navy Neuropsychiatric Research Unit.

Plag, J. A., & Goffman, J. M. (1966). *The prediction of four-year military effectiveness from characteristics of Naval recruits* (Report No. 66-8). San Diego, CA: Navy Neuropsychiatric Research Unit.

Plag, J. A., & Goffman, J. M. (1973). Characteristics of Naval recruits with histories of drug abuse. *Military Medicine, 138*, 354-359.

Plag, J. A., Goffman, J. M., & Phelan, J. D. (1967). *The adaptation of Naval enlistees scoring in mental group IV on the Armed Forces Qualification Test* (Report No. 68-23). San Diego, CA: Navy Neuropsychiatric Research Unit.

Plag, J. A., Goffman, J. M., & Phelan, J. D. (1971). *Predicting the effectiveness of new mental standards enlistees in the U. S. Marine Corps* (Report No. 71-42). San Diego, CA: Navy Neuropsychiatric Research Unit.

Plag, J. A., & Hardacre, L. E. (1964). *The validity of age, education and GCT score as predictors of two-years attrition among naval enlistees* (Report No. 64-15). San Diego, CA: Navy Neuropsychiatric Research Unit.

Porter, A. (1965). Validity of socio-economic origin as a predictor of executive success. *Journal of Applied Psychology, 49* (1), 11-13.

Prediger, D. J. (1969). New procedures for scoring psychological measurements: Development of moderated scoring keys for psychological inventories. *Research in Education (ERIC), 4* (05), ED 024887.

Reilly, R., & Echternacht, G. (1979). Some problems with the criterion-keying approach to occupational interest scale development. *Educational and Psychological Measurement, 39*, 85-94.

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

Rhea, B. D. (1966). *Validation of OCS selection instruments: The relation of OCS selection measures to OCS performance*. (Technical Bulletin STB 66-18). San Diego, CA: U. S. Naval Personnel Research Activity.

Richardson, Bellows, Henry, & Co. (1971). *Predicting job tenure among ES applicants and program tenure among WIN clients through the use of biographical information.* Washington, DC.

Ritchie, R. J., & Boehm, V. R. (1977). Biographical data as a predictor of women's and men's management potential. *Journal of Vocational Behavior, 11*, 363-368.

Robinson, D. D. (1972). Prediction of clerical turnover in banks by means of a weighted application blank. *Journal of Applied Psychology, 56* (3), 282.

Ronan, W. W. (1964). Evaluation of skilled trades performance predictors. *Educational and Psychological Measurement, 24* (3), 601-607.

Rosenbaum, R. W. (1976). Predictability of employee theft using weighted application blanks. *Journal of Applied Psychology, 61*, 94-98.

Sands, W. A. (1978). Enlisted personnel selection for the U.S. Navy. *Personnel Psychology, 31*, 63-70.

Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.

Schmuckler, E. (1966). *Age differences in biographical inventories: A factor analytic study*. Greensboro, NC: The Creativity Research Institute, The Richardson Foundation.

Schoenfeldt, L. F. (1974). Utilization of manpower: Development and evaluation of an assessment-classification model for matching individuals with jobs. *Journal of Applied Psychology, 59*, 583-595.

Schrader, A. D., & Osburn, H. G. (1977). Biodata faking: Effects of induced subtlety and position specificity. *Personnel Psychology, 30*, 395-404.

Schuh, A. J. (1967). The predictability of employee tenure: A review of the literature. *Personnel Psychology, 20*, 133-152.

Schwab, D. P., & Oliver, R. L. (1974). Predicting tenure with biographical data: Exhuming buried evidence. *Personnel Psychology, 27*, 125-128.

Schwartz, S. P. (No official date, but 7/75 noted by author). *Cross-validation of a weighted application blank for predicting asbenteeism of school custodial workers*. Unpublished manuscript, Atlanta Public Schools Personnel Division.

Scott, R. D., & Johnson, R. W. (1967). Use of the weighted application blank in selecting unskilled employees. *Journal of Applied Psychology, 51*, 393-395.

Seeley, L. C., Rosen, T., & Stroad, K. (1978). *Early development of the Military Aptitude Predictor* (Technical Paper 2-88). Washington, DC: U. S. Army Research Institute for the Behavioral and Social Sciences.

Shanthemai, V. S. (1978). Industrial use of dexterity tests. *Journal of Psychological Researches, 23* (3), 200-205.

Shott, G. L., Albright, L. E., & Glennon, J. R. (1963). Predicting turnover in an automated office situation. *Personnel Psychology, 16*, 213-220.

Smith, W. J., Albright, L. E., Glennon, J. R., & Owens, W. A. (1961). The prediction of research competence and creativity from personal history. *Journal of Applied Psychology, 45*, 59-62.

Standard Oil Company (New Jersey). (1962). *Social science research reports: Selection and placement, Vol. II*. New York: Author.

Standlee, L. S., & Abrahams, N. M. (1980). *Selection of Marine Corps Drill Instructors* (Technical Report 80-17). `San Diego, CA: Navy Personnel Research and Development Center.

Strimbu, J. L., & Schoenfeldt, L. F. (1973). Life history subgroups in the prediction of drug usage patterns and attitudes. *Selected Documents in Psychology, 3*, 1-15.

Tarofsky, R., Shepps, R. R., & O'Neill, P. S. (1969). Pattern analysis of biographical predictors of success as an insurance salesman. *Journal of Applied Psychology, 53*, 136-139.

Taylor, C. W. (1962). *Explorations in the measurement and prediction of contributions of one sample of scientists*. USAF Technical Report No. 61-96, San Antonio, TX.

Taylor, C. W., & Ellison, R. L. (1967). Biographical predictors of scientific performance. *Science, 155*, 1075-1080.

Thayer, P. W. (1977). Somethings old, somethings new. *Personnel Psychology, 30*, 513-524.

Toole, D. L., Gavin, J. F., Murdy, L. B., & Sells, S. B. (1972). The differential validity of personality, personal history, and aptitude data for minority and nonminority employees. *Personnel Psychology, 25*, 661-672.

Tucker, M. F., Cline, V. B., & Schmitt, J. R. (1967). Prediction of creativity and other performance measures from biographical information among pharmaceutical scientists. *Journal of Applied Psychology, 51* (2), 131-138.

Ward, J. H., & Hook, N. E. (1963). Application of a hierarchical grouping procedure to a problem of grouping profiles. *Educational and Psychological Measurement, 23*, 69-81.

Waters, L. K., Roach, D., & Waters, C. W. (1976). Estimates of future tenure, satisfaction, and biographical variables as predictors of termination. *Personnel Psychology, 29* (1), 57-60.

Webster, E. G., Booth, R. F., Graham, W. K., & Alf, E. F. (1978). A sex comparison of factors related to in Naval Hospital Corps School. *Personnel Psychology, 31* (1), 95-106.

Weiss, D. J. (1976). Multivariate procedures. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally.

Wernimont, P. F. (1962). Reevaluation of a weighted application blank for office personnel. *Journal of Applied Psychology, 46*, 417-419.

Wernimont, P. F., & Campbell, J. P. (1968). Signs, samples and criteria. *Journal of Applied Psychology, 52* (5), 372-376.

Wilcove, G. L., Thomas, P. J., & Blankenship, C. (1979). *The use of preenlistment variables to predict the attrition of Navy female enlistees* (SR 79-25). San Diego, CA: Navy Personnel Research and Development Center.

Williams, W. E. (1961). *Life history antecedents of volunteers versus nonvolunteers for an AFROTC program.* Paper presented at Midwestern Psychological Association, Chicago, IL.

Wise, D. A. (1975). Academic achievement and job performance. *American Economic Review, 65* (3), 350-366.

Wiskoff, M. F. (1980). *Selection of Marine Corps drill instructors* (Technical Report 80-17). San Diego, CA: Navy Personnel Research and Development Center.

Yellen, T. M. I. (1975). *Validation of the Delinquent Behavior Inventory as a predictor of basic training attrition* (Technical Report 76-3). San Diego, CA: Navy Personnel Research and Development Center.

# Section 3

# Utility of Interest Assessment for Predicting Job Performance

Bruce N. Barge and Leaetta M. Hough

# SECTION 3

# UTILITY OF INTEREST ASSESSMENT FOR PREDICTING JOB PERFORMANCE

## Overview

The British poet, Stephen Spender, once told about his desire to be a naturalist. As a youth, he pictured himself years later, with a flowing white beard, sitting in the park and studying the flowers. Asked what made him change his mind, Spender replied: "A course in botany" (Kuder & Diamond, 1979).

Additional information can often aid an individual or an employing organization in determining whether they are well suited for each other. Avoiding a poor match can save time and self-confidence for the individual, and selection and training costs for the organization. This additional information is of increased importance today, given the tremendous diversification of occupational opportunities available and the rapid technological evolution within given jobs. The most recent edition of the Department of Labor's *Dictionary of Occupational Titles* (1977) lists 20,000 different job titles. Also, the increasing complexity of jobs often requires that an individual be extensively trained prior to beginning work.

This section describes assessment of the interests of the individual. This section also describes how the resulting information is of value in optimizing the critical match between person and job.

## Definition and History of Interest Assessment

E. K. Strong, perhaps the most recognized researcher in interest assessment, has defined interest in the following ways: "Interest is a response of liking . . . . an aspect of behavior" (1943, pp. 6, 8). Interests "point to what the individual wants to do, they are reflections of what he considers satisfying . . .," "Interests may be conceived of as a single expression, a general tendency or the score on an interest inventory" (p. 19). Darley and Hagenah (1955) wrote that "occupational interests reflect, in the vocabulary of the world of work, the value systems, the needs, and the motivations of individuals" (p. 191). Holland (1973) added that vocational interests "are simply another aspect of personality" (p. 7).

The first real attempt to systematically assess interests was probably that of E. L. Thorndike, who in 1912 asked 100 college students to rank order their interests as they remembered them in elementary school, high school, and currently. The next significant step occurred during a seminar conducted by C. S. Yoakum at Carnegie Institute of Technology in Pittsburgh in 1919. Graduate students in the seminar wrote approximately 1,000 interest items which formed the basis of many future interest inventories. Included among these inventories was the *Strong Vocational Interest Blank* (SVIB) which was pioneered by E. K. Strong over a 40-year period beginning in the 1920s.

Strong's approach to the problem of interest measurement is summarized in the following sentence (1943, p. 107): "Men engaged in occupations so far studied have a characteristic pattern of likes and dislikes that differentiates them from men in other occupations." By using a general reference sample ("men-in-general") as the basis for weighting items to differentiate between occupational groups, Strong successfully developed numerous occupational scales with impressive validity and reliability statistics. The resulting *Strong Vocational Interest Blank* (SVIB) was a strictly empirical, atheoretical instrument which, nevertheless, worked very well. Strong's empirical approach continues to be used in its essential form in the construction of occupational scales today.

Although the empirical approach is powerful in prediction and differentiation, the resulting scaled items have sometimes been quite dissimilar in content and caused considerable difficulty in interpretation. There was no consensus on the conceptual meaning of an interest score and thus no way of integrating interests into a greater understanding of human behavior. Three historical developments have attempted to address this need: factor analysis of scale intercorrelations, construction of basic interest scales, and formulation of theories of interests.

Factor analysis has been helpful in gaining conceptual understanding of the basis of the interest domain. Beginning with early work by Strong and Thurstone in 1929, this research has investigated the underlying, basic dimensions of interests with results ranging from four factors (Thurstone, 1931) to six (Guilford, Christensen, Bond, & Sutton, 1954) to approximately 12 (Rounds & Dawis, 1979). Holland (1976) reported basic consensus on four to eight factors.

Basic interests scales are developed by statistically identifying clusters of items with homogeneous content and then interpreting and meaningfully labeling the cluster. Clark (1961) was among the first to make use of this technique, and the *Minnesota Vocational Interest Inventory* (Clark & Campbell, 1965) was the first interest inventory to include both empirical occupational scales and homogeneous basic interest scales, a practice now widely followed.

The third major area contributing to greater understanding of interests has been the theoretical work, primarily of Roe (1956) and Holland (1966, 1973). These theories have formed the basis for the integration of interests with broad conceptions of the individual's development, personality, and vocational behavior. They have also been valuable in focusing and generating further research into the role and meaning of interest measurement. All of these approaches will be discussed in more detail later in this section.

## Use of Interest Assessment

The applications of assessed interests may be considered as falling in two major categories: career assistance and vocational research (Holland, Magoon, & Spokane, 1981). While these categories overlap, career assistance is directed toward a particular individual or group and includes such activities as occupational instruction and career counseling. By contrast, vocational research addresses issues of more general concern, such as the origins and structure of interests, group differences in occupational choice, and the effectiveness of various types of vocational assessment.

Career Assistance. Career assistance, or as Tuckman (1974) called it, career development, is defined as a process that enhances a person's ability to (a) develop and become aware of concepts about oneself (self-awareness); (b) develop and become aware of one's environment, including occupations; and (c) make career choices. This process is normally an individual activity but is often aided by a career counselor and/or an exploratory and assessment tool such as an interest inventory. Another popular method of career assistance is through group treatment, such as career planning seminars which offer additional career and decision-making information.

The need for career assistance has been increasingly observed in the workplace. Hall (1976) attributed this to (a) rising concerns for quality of work life and personal life planning, (b) equal employment opportunity (EEO) legislation and affirmative action pressures, and (c) rising educational level and occupational aspirations coupled with (d) slow economic growth and reduced advancement opportunities. Another contributing factor is the changing career aspirations and opportunities for women today.

Vocational Research. In contrast to the career assistance orientation to the individual, vocational research is directed toward more general, organizational applications. Such research has investigated the role of interests in employment and educational selection and found them to be useful in choosing those individuals most likely to be satisfied in the career. Assessed interests have been found to be predictive of such criteria as length of occupational membership, job satisfaction, and performance in training and on the job.

Interests have been demonstrated to be helpful in classifying persons within an organization. This application is particularly appropriate for placing of those individuals who have the necessary aptitude for several different jobs or when the required job aptitudes are similar for all jobs. A further benefit of research on interests is in terms of better understanding of groups--why there are between-group differences and within-group changes. Similarly, this research adds to our comprehension of how individuals develop and how they differ.

This brief overview has discussed how and why the study of interests has evolved. Many of the topics have only been mentioned and will be examined in more detail later. It should be clear, however, that the study and use of measured interests has increasingly warranted a role in our understanding of human behavior and in decisions that concern behavior. Since the measurement of interests is a necessary precursor to any further discussion, issues relating to measurement will be examined next.

## Interest Measurement

There are two basic methods for measuring interests. One is simply to record the individual's expressed interest, as in the case of a college major. The other method is a systematic presentation of items for which the subjects' response indicates degree of interest or preference. This method, the interest inventory, has received the more detailed investigation within the domain and is discussed first.

## Interest Inventories--Description

Items. Inventory items may be of many content types. The *Vocational Preference Inventory* (Gottfredson, Holland, & Holland, 1978) consists entirely of occupational titles with the assumption that people hold valid occupational stereotypes and titles are the best items to elicit these stereotypes. Conversely, the Kuder *Occupational Interest Survey* (Kuder & Diamond, 1979) specifically avoids occupational titles and includes only items that specify an activity. The *Strong-Campbell Interest Inventory* (Campbell & Hansen, 1981) includes items from seven areas: occupational titles, school subjects, activities, amusements, types of people, preference between two activities, and personal characteristics. This comprehensive approach is the most common among inventory publishers.

Response Format. Responses to the items are made in two basic ways. One is forced choice, in which the respondent indicates which one of two paired items is of more interest to her or him. This is the format used in the *Vocational Interest Inventory* (Lunneborg, 1981) and the *Jackson Vocational Interest Survey* (Jackson, 1977). An alternate type of this response method is the triad system that is used in the *Kuder Occupational Interest Survey* (Kuder & Diamond, 1979) and the *Navy Vocational Interest Inventory* (Clark, 1961; Dann & Abrahams, 1973). The triad format presents three items together and the respondent indicates which item of the three is most preferred and which is least preferred.

The other major response format presents items singly and the respondent indicates his/her liking, usually in terms of "Like," "Indifferent," or "Dislike" (L-I-D). Variations of this method include the use of a 5-point preference format from "Strongly Like" to "Strongly Dislike' or a 2-point format the deletes the "Indifferent" option. Inventories that use the liking rating include the *Army Research Institute Interest Survey* (Claudy, Caylor, & Kass, 1981), the *Career Assessment Inventory* (Johansson, 1982), the *Strong-Campbell Interest Inventory* (Campbell & Hansen, 1981), the *Self-Directed Search* (Holland, 1979), the *Vocational Interest Career Examination* (Alley & Matthews, 1981), and the *Vocational Preference Inventory* (Gottfredson, Holland, & Holland, 1978).

The proponents of forced-choice formats argue that their approach helps to "pull interests apart" and reduces the problem of flat profiles (Lunneborg, 1979). This might be especially appropriate for young people who, through lack of experience, do not yet have well-differentiated interests. A second argument is that the L-I-D rating format is more susceptible to response bias because of its fixed-response format (Cronbach, 1950). Different people may respond from different reference points for the same terms (i.e., a "Like" response may mean different things to different people), and some have a tendency to like or dislike all items independently of content.

Jackson (1977) has taken a direct approach to the assessment of this bias. Through factor analysis, he identified a large first factor that accounted for approximately one-third of the total score variance and was interpreted as involving a general, like-dislike bias in responses. In developing his own interest scales, Jackson partialed out this effect and found that this procedure resulted in markedly lower scale intercorrelations and consequently more meaningful scales.

The findings with the *Strong-Campbell Interest Inventory*, however, suggest that response bias is not a serious problem for the L-I-D rating format (Campbell & Hansen, 1981). Item correlations for type of response show only a mild tendency for people to adopt the same response strategy over various sections of the inventory. It might also be argued that this tendency to generally like or dislike the activities and occupations presented as items reflects an actual behavioral tendency that is of interest and importance in itself. Further, it has been widely reported that forced-choice formats often provoke reactivity in respondents. This occurs when subjects do not greatly prefer one item over another, or when they like or dislike both items and resent having to choose between them.

Perry (1955), who was the first to compare the forced-choice (ranking) format and the rating format, found that forced-choice items differentiated groups better than L-I-D rating items, although not to a practically significant extent. Similarly, Fisher, Weiss, and Dawis (1968) found a slight psychometric superiority for ranking over rating, but at the expense of administration time. Dawis (1980) concluded that the choice of format depends upon the purpose of the interest assessment. The forced-choice method may be most appropriate for examining intraindividual differences, as in an individual counseling setting, while the rating format is well suited for assessing interindividual differences, for comparison among many members of a group (e.g., recruits).

An issue related to response format is that of ipsativity. Cattell (1944) described ipsativity as occurring when the ranking of one item over another results in a relative measure rather than an absolute one. This outcome is normally associated with the forced-choice format (Knapp, 1966). Clemans (1968) argued that the forced-choice format does not necessarily result in an ipsative measurement and that an instrument using a rating format may yield scores with ipsative characteristics. An ipsative set of scores is one that always adds up to a constant, thus making it impossible for a subject to receive the highest scores on all scales.

In this sense, many inventories using a rating format can be considered at least partially ipsative since a given individual will almost always have some low scores as well as high. This is a result not of the response format but rather the scoring of the items. Clemans (1966) noted that, when absolute measures can be obtained, they normally produce a higher multiple correlation with a criterion than the same set of ipsatized variables. In addition, ipsative scoring reports the interests of an individual relative to his or her other interests rather than on an absolute level suitable for peer comparison. Reliabilities can be hard to interpret with ipsative scoring and understanding the construct tapped by the scale can be difficult.

## Occupational Scales

Rationale. Occupational scales grew out of the need to be able to discriminate those persons who are satisfied with their employment in a certain occupation from those otherwise employed. Once the interests of satisfied persons in each occupation have been determined, this information can be used to predict the occupation in which a new worker might be most satisfied. Each scale is directed only toward a single occupation and represents a very specific prediction of satisfaction for that occupation alone.

Construction. As discussed earlier, occupational scales were originally developed through the empirical approach of E. K. Strong. Under this approach, an occupational scale is composed of the items that differentiate members of an occupational sample from the people-in-general sample. Items that are preferred more often by satisfied members of the occupation than by people-in-general receive a positive weight and items that are disliked more often are assigned a negative weight. For example, both men-in-general and engineers were asked to give a "Like," "Indifferent," or "Dislike" response to the occupation of "Actor." Twenty-one percent of the men-in-general and 9% of the engineers responded "Like" to this item. On the basic of this difference, the item is included on the engineer occupational scale with a weight of -1 for a "Like" response and a +1 for a "Dislike" response (Strong, 1943, p. 75).

The items with the largest difference between the samples above some minimum level (usually around 15% difference) are selected for inclusion in the scale. The scale constructor must decide how many of these items are necessary for adequate reliability and at what point additional items cease to increase the scale's validity significantly (Cronbach & Gleser, 1965). Abrahams (1965) showed that, when various item characteristics are held constant, there is no single, optimal upper limit for the number of items included on a scale. It should also be noted that the items included on these scales are not necessarily the ones chosen most or least often by members of an occupation. Often, items that are most or least popular are answered similarly by all and do not produce good differentiation between occupational groups and people in general; therefore, they are not included in the scale.

Those items that meet the criterion level of difference are identified and given unit weights in the appropriate direction. Several examples of this procedure are given by Campbell (1971):

| | | L | I | D |
|---|---|---|---|---|
| Item A | Criterion sample percent response | 60 | 30 | 10 |
| | Men-in-General percent response | 40 | 38 | 22 |
| | Difference | +20 | - 8 | -12 |
| | Assigned weight | + 1 | 0 | - 1 |
| Item B | Criterion sample percent response | 10 | 40 | 50 |
| | Men-in-General percent response | 3 | 28 | 69 |
| | Difference | + 7 | +12 | -19 |
| | Assigned weight | + 1 | + 1 | - 1 |
| Item C | Criterion sample percent response | 30 | 40 | 30 |
| | Men-in-General percent response | 40 | 20 | 40 |
| | Difference | -10 | +20 | -10 |
| | Assigned weight | 0 | 0 | 0 |

Once the items are selected and weighted, norms are developed. The distribution of raw scores of the criterion group is converted to a distribution of standard scores. This distribution can then be used to convert all future scores to a common numerical scale useful for comparisons across scales.

Sample Selection. Campbell (1971) reported the following desiderata for the people-in-general (general reference) sample:

1. The general reference sample should include the same general types of people as do the criterion groups.

2. The general reference sample should be highly diverse, so that most interest patterns will have some chance of being represented there.

3. Similarly, heterogeneity should be addressed to avoid any single dominant bias.

4. Finally, the group should be large enough to ensure the stability of the item statistics.

Thus, a short description of a desirable general reference sample would be "a large, diverse, heterogeneous sample of the same sorts of people used to develop the Occupational Scales."

Similarly, Campbell and Hansen (1981) listed the following member characteristics for the occupational (criterion) sample: (a) satisfied in their job, (b) not unsuccessful at work, (c) aged 25-55, (d) at least three years' experience, (e) performing the occupation in the typical manner, and (f) willing to take part in the research project. These characteristics are particularly important since, as mentioned, the criterion sample is typically used to norm the occupational scales.

An additional issue that relates to sample selection is that of sex fairness. When general reference samples are formed for scale construction, there are normally separate samples for men and women. These separate samples are necessary because men and women differ considerably in their base rate of responding to various items, even within the same occupation. The final result of this process is different scale weights, norms, and, in some cases, separate scales for the same occupation for men and women. The procedure allows males and females to assess their interests on an equal basis within their sex, but is more cumbersome and has been criticized as serving to maintain occupational inequality. Sex bias will be discussed separately and in more detail later in the section.

Alternate Approaches to Occupational Scale Construction. The method of scale construction pioneered by Strong contrasted an occupational sample with a general reference (men-in-general) sample. Kuder and Diamond (1979) criticized this approach on three grounds. First, the composition of the general reference sample is problematic. Strong (1943) discovered that scales based on a sample of the general population did not produce good differentiation between occupational groups. Second, there is no general reference group that is satisfactory for the whole range of occupations. Third, the use of a general reference group implies that the occupational choice is between a single occupation and a composite of all other occupations rather than between the occupations in each of the many possible pairs. A final problem with a general reference sample is time bias--the tendency for certain items to be more popular among a sample at one period of time than another (Campbell, 1971).

Based on these criticisms, the *Kuder Occupational Interest Survey* (Kuder & Diamond, 1979) makes use of a lambda coefficient, a statistic similar to a biserial correlation. These lambda coefficients are used as scores to express the extent to which the subject's pattern of interest resembles the pattern of interests of individual occupational groups. For example, an individual may obtain a score (lambda coefficient) of .50 on the Office Clerk scale. This score represents the degree of similarity between the individual's interests and those of office clerks. The score does not have anything to do with the interests of people-in-general.

It may appear, in view of Kuder's criticisms and alternate formulation, that scale construction using a general reference group is inappropriate. However, Strong's finding that the general reference sample must be nonrepresentative in order to differentiate between groups is probably due to the professional orientation of his occupational scales. Respondents from unskilled and semi-skilled groups are inappropriate for aiding in differentiating among business and professional occupations, and a heterogeneous sample of the same sorts of people in those occupations being studied did comprise an effective general reference sample. Secondly, with Kuder's method there is the problem of controlling for base rate popularity in which generally popular items across all occupations cause all scale correlations to be positive. Campbell (1971) pointed out that whether a scale should be constructed with criterion (occupational) and general reference samples or just criterion samples depends on whether one wishes to attend to the unique interests or the most popular interests of the criterion sample.

A third, most recent, type of occupational scale construction makes use of rational and statistical procedures. This method attempts to construct scales that are more homogeneous and psychologically meaningful based largely on statistical information. For example, the *Vocational Interest Career Examination* (VOICE) (Alley & Matthews, 1982) includes occupational scales based on a statistical analysis of basic interest scores and reported satisfaction in Air Force occupations. Each scale on the VOICE represents a prediction of job satisfaction from separate regression equations that are based on basic interest scores.

Advantages and Disadvantages. Heterogeneous occupational scales are very powerful in separating occupational groups since they have been specifically constructed in that manner. Consequently, these scales are highly effective in predicting later occupational membership. Johnson and Johansson (1972) found that, 10 years later, 75 percent of male students who had high scores on either the "Life Insurance Agent" or "Physicist" scales were in occupations related to their earlier profiles. Similarly, Strong (1955), in an 18-year followup, found that 64 percent of college students who had the highest rating on the "Physician" scale were currently employed as physicians. Also, most of those who were not working as physicians were employed in related fields such as chemistry or geology.

Occupational scales are also limited in important ways. An occupational scale score gives information only about how similar the individual's interests are to the interests of members of the occupation. A high score on the "Engineer" scale says nothing about what engineers' interests are or in what way the individual's interests are like those of engineers. Thus, the statement "Your interests are similar to those of engineers" is difficult to interpret.

A second disadvantage is that there is no limit to the potential number of occupational scales. In theory, one could be developed for each of the 20,000 jobs listed in the *Dictionary of Occupational Titles*. This narrowness of occupational scales may invite unwarranted interpretations. For example, generalizing the score on an available scale to an occupation for which there is no scale may or may not be justified and needs to be investigated empirically first. Also, for virtually any application of interest scores, having so many different scores is extremely cumbersome.

### Basic Interest Scales

Rationale. Basic interest scales were developed to address the interpretability problems of occupational scales. Thus, they are characterized by homogeneity of content, generalizability across occupations, and a relatively small number of scales representing broad areas of interest. Examples of basic interest scales are science, art, outdoors, and medical science.

Construction. Based on the responses of a general reference sample, an item intercorrelation matrix is generated. This matrix is then used to cluster related items into a scale. Campbell (1971) wrote that two standards should be followed in this process. First, each clustering decision should be based on statistical evidence (high item intercorrelations). Second, even if an item has high intercorrelations with the items of a cluster, it should not be included in the scale if its content does not fit with the cluster content (face validity). Johansson (1982) added that if an item could be included statistically in more than one cluster, it should usually be added to the cluster to which it is most similar psychologically.

A related method of constructing basic scales has been used by Jackson (1977). This procedure involves factor analyzing the intercorrelation matrix and building scales that correspond to the factors. Factor analysis is distinguished from cluster analysis, in that in the former items may contribute variance to more than one factor, while in the latter they are assigned to only one cluster. An advantage of the factor analysis is that it can produce uncorrelated factor scores for each scale, thus holding scale redundancy to a minimum while still ensuring high homogeneity.

Weighting of items for basic interest scales can be done with unit weights (as discussed for occupational scales) or a complex weighting scheme that takes into account item correlations with irrelevant scales as well as scale intercorrelations. Basic scales are normally evaluated in terms of their internal consistency and scale intercorrelations. Ideally, the former would be high and the latter low, indicating that the scales are homogeneous in content and nonredundant with each other. This combination would allow the maximum in interpretability. As with occupational scales, norms are developed from a sample representative of the targeted population by converting raw scores to standard scores.

Advantages and Disadvantages. As discussed, basic interest scales have a number of advantages over occupational scales in terms of interpretability. The scores are easy to understand and generalizable to a large number of applications and criteria. Also, these homogeneous scales are helpful in understanding the organization of interests as revealed by the patterns of intercorrelations.

The main disadvantage of basic interest scales is that they may be overly general for specific occupational choice. A high interest score in science is actually of little help in choosing a future career because of the multiplicity of scientific branches. For this application, the relatively small number of basic interest scales is actually a handicap.

## Occupational Theme Scales

Rationale. The results of both factor analytic and theoretical work suggest that there are between four and eight basic dimensions of interests (Holland, 1976). These dimensions have been incorporated into the theories of interests proposed by Holland (1966, 1973) and by Roe (1956). Briefly, Holland's theory postulates that people can be categorized in terms of six basic types--realistic, investigative, artistic, social, enterprising, and conventional--and that occupational environments can be divided in the same way. Roe's theory is similar to that of Holland, but utilizes eight dimensions of interests and occupations. Occupational theme scales are constructed to align with the basic dimensions and thus incorporate the interest assessment into a theory of careers. The score on a theme scale is an indicator of the likelihood that the individual's type will match with the environmental categories.

Construction. Item selection for theme scales is much more of a rational procedure than the atheoretical methods used with other scale types. Although statistical information such as item intercorrelations is taken into account, items are selected by rational judgment of which items would be most appropriate for each scale as delineated in the theory. As with the other scale types, items are unit weighted (+1, 0, -1) and scores are normed in the usual fashion. Campbell and Holland (1972) and Hansen and Johansson (1972) described how this process was completed for the *Strong Vocational Interest Blank*.

Advantages and Disadvantages. The main advantage to theme scales is that they allow interpretation of interest scores in terms of a theory of interests that is not restricted to a single interest or a single job. Theme scales offer an appraisal of broad aspects of an individual and also the work environment she or he plans to enter. Also, the small number of scales provides a compact overview of interests and jobs that is easy to grasp and to remember, and that can be used to focus further investigation.

The disadvantages of theme scales are similar to those of basic interest scales. With the increase in generalizability comes a lack of information about specific occupations. A compact overview is useful but cannot stand alone; other information is necessary.

## Empirical Comparison of Scale Types

Dolliver (1975) used the *Strong Vocational Interest Blank* scores obtained 12 years earlier from 163 college graduates to compare what he termed the global (Holland theme scales), the middle (basic scales), and the most specific (occupational scales). His procedure included: (a) identifying the scale of each scale type that most closely corresponds to each subject's current occupation; (b) examining the score received 12 years earlier on that scale; (c) tallying the accuracy of the inventory's prediction, while (d) controlling for the differential effect of chance. (Because there are fewer theme scales, the probability of a hit being due to chance is greater than with the more numerous occupational scales.) Results indicated that although the Holland theme scales gave the best results and occupational scales the poorest, all three scale types yielded roughly comparable hit rates.

Another study, by Dolliver, Irvin, and Bigley (1972), found that while SVIB occupational scales were fairly effective in predicting occupational membership, they had no relationship with job satisfaction. This result was followed up by another study (Kunce, Decker, & Eckleman, 1976) that compared SVIB occupational scales and basic scales in predicting job satisfaction. Job satisfaction was found to be significantly related to the level of correspondence between subject's occupation and basic interests. The authors suggested that "the correspondence between basic interest scales and jobs may have more predictive validity for job satisfaction than the correspondence between occupational scales and jobs" (p. 361).

It should be noted, however, that these three studies all compared scales that were developed with the same SVIB items, an item pool that had been empirically selected prior to homogeneous scale construction. In this situation, homogeneous scales might be expected to work as well or better than heterogeneous scales since the items making up all of the scales are the same and have been previously validated. Thus, the homogeneous scales of the SVIB have the best of both worlds, the interpretability advantage of high item intercorrelations as well as the validity stemming from previous empirical item selection. By contrast, other studies have compared empirically developed, heterogeneous scales with homogeneous scales that were constructed without the benefit of previous empirical validation.

Reilly and Echternacht (1979) constructed homogeneous, basic interest scales and heterogeneous, occupational scales with more than 3,000 Air Force personnel from eight occupational areas. The homogeneous scales were developed by assignment of rationally written items to scales that had been determined a priori. Next, correlations were computed between each item and the total score of all items assigned to the scale. These correlations were used to eliminate items with the lowest relationship to the scale. The heterogeneous, occupational scales were constructed using basically the same procedure originated by Strong with an occupational sample and a men-in-general sample.

After cross-validation, the scales of each type were compared for prediction of both occupational membership and job satisfaction. Heterogeneous scales were found to be better in predicting whether a subject belonged to a particular occupational group or to the people-in-general group. Conversely, the homogeneous scales had higher correlations with job satisfaction. The authors speculated that the empirical selection of occupational scale items causes items to be included that have idiosyncratic relationships with the occupation but, unlike homogeneous scales, are not related to a broader range of criteria, such as job satisfaction and tenure. These results were general trends, however, and did not reflect large differences in either criterion.

Both types of scales make an important contribution to interest assessment and either type is less valuable without the other. In addition, their differences in level of specificity add to their complementary nature. Thus, inventories such as the *Strong-Campbell Interest Inventory* (Campbell & Hansen, 1981), the *Career Assessment Inventory* (Johansson, 1982), and the *Vocational Interest Career Examination* (Alley & Matthews, 1982) include both occupational and basic interest scales in order to yield a profile that is well rounded and suited to the multiplicity of uses in which interest assessment is employed.

## Additional Measurement Concerns

Reliability. Although reliability levels vary by inventory and calculation method, nearly all published interest inventories have very good reliability. Campbell and Hansen (1981) reported that median test-retest values for the *Strong-Campbell Interest Inventory* are around .90 for intervals up to 30 days with all types of scales. Kuder and Diamond (1979) reported 2-week, test-retest values of .90 for individual profiles on the *Kuder Occupational Interest Survey* and correlations ranging from .84 to .92 for the differences between pairs of scale scores obtained in two administrations. Similarly, the scales of the *Vocational Interest Career Examination* have a median coefficient alpha reliability of .94 (Alley & Matthews, 1982).

Stability. In studies of the stability of interests, the test-retest method is typically used over a long interval with the focus on the interests of the group rather than evaluation of the inventory. Strong (1943), in an extensive treatment of stability, drew the following conclusions: (a) of the several different ways to measure stability such as computing a coefficient of stability, most yield similar results; (b) stability is higher with shorter test-retest intervals and older subjects; (c) different items have different stabilities associated with them; and (d) interests are more stable than is reflected by interest inventories because of the limits of a three-point response format. Campbell (1971) presented the SVIB profiles of several groups for test-retest periods of up to 38 years. The profiles showed striking similarity. He concluded that after age 25, people's interests change very little.

A number of studies have investigated the development of interests with age. Strong (1943) reported that the interests of 25- and 55-year-old men correlated .88 (test-retest), interests of 15- and 25-year-old men correlated .82, and interests of 15- and 55-year-old men correlated .73. Strong has said informally that the change in interests between the ages of 15 and 25 can be divided into thirds--the first third occurring between 15 and 16, the second between 16 and 18, and the last between 18 and 25 (Campbell, 1971). Baggaley (1974) found that the differences between the scores of adolescents when they were in grade 8 and when they were in grade 10 were small in magnitude and may have been culturally influenced toward humanistic interests. Hansen and Stocco (1980) found that a large percentage of adolescents and adults have stable interest patterns throughout their educational career, although this stability is not universal. Similarly, Hansen and Swanson (1983) found that the majority of students have very stable interests during their college careers.

Faking. The distortion of a test score in a desired direction, commonly referred to as faking, is a serious concern for any self-report instrument because the response distortion may easily yield an invalid profile and result in incorrect decision making. Campbell (1971) reported a number of studies which, over a 40-year period, have examined the effect of faking on scores on the *Strong Vocational Interest Blank*. Much of this information had earlier been reviewed by Gray (1959). The basic method used in these early studies was to have subjects complete the inventory under normal conditions, and after a period of time, complete it again but this time after being instructed to fake their responses in order to resemble the responses of a given occupation. In general, the results reviewed showed that responses could be distorted by the subjects, although the distortion varied by the occupation to be faked and the faked items tended to be those most obviously related to the occupation.

Later, Dolliver and Clark (1972) examined the problem of faking as related to occupational prestige or status. Subjects' scores were compared under normal conditions and under instructions to attempt to score high on high-status occupations. Although the results indicated that some high-status occupations were successfully faked and others were not, there were large differences in scores under the two conditions. These results were replicated both with normal instructions followed by fake instructions and the inverse, which indicates that the effect of faking is quite independent of the order of instruction presentation.

Thus, the results of several studies of the fakability of interest inventories supports findings obtained with other self-report instruments. That is, scores can generally be distorted in response to specific instructions to fake although there is variability in the level and direction of the change. As several researchers have emphasized, subjects' ability to fake under specific fake instructions does not mean that they will fake under more normal administration conditions. Investigation of this question has largely centered on selection contexts since these situations would logically provide the most incentive to fake.

Gray (1959) compared the interest scores of 278 college students seeing counselors versus the scores of the same students applying for medical school some months later. He concluded:

> Forty-seven percent of the medial applicant group did not or could not raise their physician score between testing for counseling and testing for admission to medical school. Twenty-four percent raised their physician score enough to have a serious effect on its interpretation by an admissions officer; twenty-nine percent raised their score by a less important amount (p. 296).

Gray added that variation in the elapsed time between tests did not systematically affect the physician score.

Kirchner (1961) found that sales applicants scored lower on the "sales" scale than employed salesmen; they did, however, demonstrate a greater tendency to respond "like," which may be attributable to socially desirable responding. Campbell (1971) reported constructive replications of both Gray's and Kirchner's studies, with results that generally support the idea that faking in an application setting does not have the large effect found with more artificial faking instructions.

Abrahams, Neumann, and Githens (1968, 1971) obtained interest scores for individuals who had completed the SVIB both as part of the application for a Navy ROTC scholarship and under routine conditions (either as a high school student prior to college application or one year later as college freshmen). Ninety-seven percent overlap of the testing distributions was found in both comparisons, which caused the authors to conclude that "there is neither a consistent nor significant tendency for applicants to increase their selection scores" (1971, p. 11).

This research suggests that faking in a selection setting may be less pronounced than was earlier believed. However, this is not to say that faking has no effect or that it may not have a larger effect in some settings or for some individuals. A number of procedures have been suggested to reduce this possibility. Doll (1971) listed these as: (a) the forced-choice item technique, (b) correction or suppressor scales, (c) development of items that do not lose their validity under a faking condition, and (d) identification of the fakers.

The rationale for use of the forced-choice format is that it is possible to match the paired items on either response frequency or social desirability, thus making it more difficult to fake (Norman, 1963). Suppressor scales are designed to correct for the portion of the score that is due to faking (Meehl & Hathaway, 1946). Items that do not lose their validity under a faking condition are those that are more subtly related to the criterion and thus more difficult for subjects to perceive as appropriate for faking (Gadel & Kriedt, 1952). Finally, identifying the fakers is normally done through the use of a verification scale that indicates whether the respondent is answering honestly or only in a socially desirable manner (Kuder & Diamond, 1979).

Several studies have reported other research relevant to reducing faking effects. Zalinski and Abrahams (1979) found that using the scale's items alone rather than imbedded in the context of the whole inventory contributed to more faking on the scale. Doll (1971) suggested that a fake detection key should include subjective kinds of items with continuous response formats, since these are most likely to show faking. Gordon and Gross (1978) showed statistically that each of the several operational definitions of faking yield somewhat different and incomplete information. They concluded that the best method of assessing faking depends upon the use intended for the instrument. Yet another approach to the problem comes from Kroger (1974) who hypothesized that faking may be defined as a form of role-taking. Kroger gave subjects subtle environmental cues related to a given scale (as opposed to explicit instructions) and found the cues decisively influenced interest scores. He concluded that subjects were not so much faking as role playing in accord with the earlier cues.

Sex Bias/Fairness. The problem of sex bias is especially thorny for interest inventories because researchers have consistently found differences in interests between females and males that appear early and persist through adulthood (Campbell & Hansen, 1981; Riley, 1981). Hansen (1983) suggested that these differences may be characterized as cultural--for example, when men prefer items from Holland's "Realistic" area and women prefer the "Artistic" area--or occupational, which vary from occupation to occupation. Many of the differences in interests are relatively constant across all occupations and no occupation is free of them. Further, examination of the amount of difference for the 1930s, 1960s, and 1970s has shown that only small changes have occurred over time and these have not been consistent in either direction (Campbell & Hansen).

Since it is possible that biases within interest inventories have contributed to the differences, both the Association for Measurement and Evaluation in Guidance and the National Institute of Education (NIE) organized symposia in the mid-1970s to study the issues involved. Sex bias was defined by the NIE as " . . . any factor that might influence a person to limit--or might cause others to limit--his or her consideration of a career solely on the basis of gender" (Diamond, 1975, p. xxiii). The outcomes of these efforts, such as the NIE's *Guidelines for Assessment of Sex Bias and Sex Fairness in Career Interest Inventories* have combined with federal legislation to impose a number of changes in the field of interest assessment.

Gender-neutral language, such as "police officer" rather than "policeman," is now widely used. Most inventories have made an attempt to include only those items that are equally familiar to both females and males, and the same items are answered by both sexes. Scores for all occupations and interest areas included in the inventory are reported for both sexes, and inventories are designed to encourage wider career exploration for males and females. Information relevant to the sex fairness of the inventory, such as criterion and norm group composition and rationale, is clearly reported.

Controversy continues, however, over the use or non-use of separate sex norms and separate sex scales. For occupational scales, Campbell and Hansen (1981) concluded that separate scales should be developed for men and women because same-sex scales result in better differentiation between occupational groups and the general reference sample. Hansen (1976) has shown that combined-sex scale construction is possible only if sex differences are treated as irrelevant variables and concurrent validity is sacrificed. Similarly, Kuder and Diamond (1979) reported that, for the Kuder occupational scales, norms for the two sexes cannot be combined without discriminating seriously against one sex or the other.

Both the *Strong-Campbell Interest Inventory* and the *Kuder Occupational Interest Survey* additionally report scores on occupational scales that are developed or normed for the other sex. This approach encourages both men and women to consider occupations heretofore dominated by the other sex. Dolliver (1981) has shown that this procedure usually results in higher scores for both men and women on the opposite sex scale.

Johnson (1977) criticized the approach as serving to reinforce sexual stereotypes, since both sexes score relatively high on cross-sex scales representing "traditional" occupations and low on those representing "nontraditional" occupations. He advocated the use of separate sex norms rather than separate sex scales, use of the lambda coefficient rather than a general reference sample procedure, use of Rayman's (1976) procedure of developing sex-balanced items, and most especially, emphasis on homogeneous basic and theme scales used with separate sex norms.

There is also controversy over whether separate sex norms are appropriate with homogeneous scales. Gottfredson, Holland, and Gottfredson (1975) compared separate sex-normed scores with raw scores and found that the use of these norms resulted in 16 times as many Realistic assessment outcomes for college-level women as actual Realistic employment. They argued that separate sex norms result in inappropriate assessments and also reduce predictive efficiency (Gottfredson & Holland, 1975). Conversely, a number of studies summarized by Hansen, Prediger, and Schussel (1977) found that separate sex-norm scores are at least as valid as raw scores. Further, Lamb and Prediger (1979) showed that interest scales composed of sex-balanced items can obtain criterion-related validities as high as those of raw scores or same-sex normed scores.

To summarize, it appears that significant steps have been taken to reduce overt sex bias in interest inventories. Controversy remains, however, over scaling and norming procedures and whether a trade-off is necessary between validity and the reduction of sex differences. Also, the appropriateness of certain procedures (e.g., separate-sex norms) may depend on how the inventory's scales were constructed; there may be no one "correct" procedure that is universally optimal for all inventories. For a thorough discussion of these issues, the reader is referred to Tittle and Zytowski (1978).

Racial Bias. In contrast to sex bias, the racial bias of interest inventories has received much less attention, primarily because no important and stable racial differences have been identified. Strong (1943) reported no significant differences between white Americans and blacks, second-generation Japanese, or Scotchmen. Similar findings will be discussed later for the Holland RIASEC structure of interests.

Other studies by Barnette and McCall (1964), Borgen and Harper (1973), and Lamb (1976) all found no differences between black subjects and white subjects. Berger and Berger (1977) did find black-white differences on the *Vocational Interest Career Examination*, with blacks showing higher mean scores on most VOICE scales. The differences were not considered sizable enough to warrant separate norms, however. Whetstone and Hayles (1975) also found a small black/white difference, with blacks achieving higher scores, but again the difference was not significant.

## Expressed Interests

Assessment. In contrast to the numerous items and complex psychometrics of interest inventories, an alternate measure of vocational interest is as simple as the question: What career are you most interested in? This measure, called expressed interest, has also been obtained as an individual's college major or through a short questionnaire assessing the degree of commitment to an occupation about to be entered (Reeves & Booth, 1979). Increasingly, expressed interest is seen as a valuable tool in predicting future occupation.

Comparison of Expressed and Inventoried Interests. Perhaps the most recognized review of this literature was that of Dolliver (1969b) who listed the following conclusions regarding the *Strong Vocational Interest Blank* versus expressed interest: (a) There is only a moderate degree of overlap between the results of the SVIB and the results of an expressed-interest method. (b) The reliability of the SVIB exceeds that of expressed interest; the reliability of expressed interest is moderately low. (c) The predictive validity of expressed interest is at least as great as the predictive validity of the SVIB. In no study where a direct comparison was made was the SVIB as accurate as expressed interests in predicting the occupation engaged in. (d) There is an apparent discrepancy, since expressed interests do not seem highly reliable but yet seem highly valid. This result may be due to the observation made by several authors that expressed interests that develop early are highly predictive. Thus, there appears to be a closer link between the reliability and validity for expressed interests than for the SVIB. (e) There is no evidence to show that the SVIB is superior to expressed interests.

In reaching these conclusions, Dolliver examined research by Strong (1935, 1943, 1953), Dyer (1939), Wightwick (1945), Enright and Pinneau (1955), and McArthur and Stevens (1955). He noted that there are numerous potential pitfalls in the methodology of this type of research and, for that reason, considered his conclusions to be somewhat uncertain. Dolliver's major contention was that the SVIB has been overused in comparison with expressed interest because of longstanding, but unwarranted, prejudice against expressed interest. A similar review by Whitney (1969) of nine large-sample, longitudinal studies drew conclusions nearly identical to those of Dolliver.

Following the lead of these reviewers, a number of more recent studies have employed both expressed and inventoried interests in their research in an attempt to understand the respective role of each in interest assessment. Gade and Soliah (1975) confirmed the results of Holland and Lutz (1968) that expressed choice is a better predictor of later occupational membership than the top three theme scale scores of the *Vocational Preference Inventory* (Gottfredson, Holland, & Holland, 1978). Further, they found no relationship between expressed choice and the inventory's scores. Gottfredson and Holland (1975), however, used expressed choice as a criterion in assessing the predictions of the *Self-Directed Search* (Holland, 1979) and found that the inventory's prediction and expressed choice were the same for 40 percent of the men and 66 percent of the women in their sample.

Dolliver and Will (1977) reported a 10-year followup of subjects who had completed both the SVIB and the *Tyler Vocational Card Sort* (TVCS), a task that requires subjects to sort vocations into a desirability hierarchy that is self-descriptive. Their results showed that the TVCS, which was considered a form of expressed interest, was slightly more accurate in predicting occupational membership although both methods achieved about 50 percent accuracy. Slaney (1978) found that the overlap between Holland theme scores as measured by the *Strong-Campbell Interest Inventory* and a vocational card sort of the same themes was statistically significant, but that the measures were not interchangeable. The subjects reported that the two methods were complementary.

A direct comparison of expressed choice and the SVIB was conducted with several hundred male National Merit Scholars who completed the SVIB and gave expressed choices prior to entering college (Borgen & Seling, 1978). Followed up three years later, they supplied their college major and career choices. Fifty-two percent of the men had both college majors and career plans in accord with their prior expressed choice while the comparable figures for the SVIB were 31 percent and 40 percent, respectively. When the expressed and inventoried predictions were congruent, the hit rate was 70 percent, but when they were incongruent, expressed prediction declined mildly (45% and 41%) and the SVIB markedly (14% and 23%). This study, which was predictive rather than retrospective, would appear especially damaging to the use of inventories rather than expressed interest. It should be noted, however, that the sample is highly atypical and the higher efficacy of expressed choice may be an artifact attributable to high level of intelligence.

Bartling and Hood (1980) set out to replicate Borgen and Seling's predictive design with a more typical college sample, followed up 11 years after assessment. The results again clearly support the conclusion that the predictive accuracy of expressed interest is greater than the predictive accuracy of measured interest. Women's expressed choice had a 60 percent "good hit" rate as compared to 30 percent for the SVIB, and the men's hit rate, while not as extreme, also showed a clear advantage for expressed choice. Hit rates were highest when the methods were congruent, but when they were not, expressed interest yielded an overall good hit rate of 47 percent compared to 9 percent for the SVIB. Bartling and Hood also examined the predictive accuracy of the SVIB for decided versus undecided students since undecided students have no expressed interest. The results showed little difference between the two groups, with good hit rates slightly above 30 percent.

Another study (Cairo, 1982) used both the SVIB and expressed choice to assess the vocational interests of 83 males at ages 15, 18, and 25. It then followed up 21 years after the first survey when subjects were age 36 and appeared to comprise a very representative sample in terms of socioeconomic status. "Good hit" rates are as shown:

| | Age of Assessment | | |
|---|---|---|---|
| | Age 15 | Age 18 | Age 25 |
| SVIB | 16% | 27% | 43% |
| Expressed | 27% | 30% | 55% |

Cairo also examined the results to see whether they might be attributable to chance (base rate). He concluded that chance was more likely to have increased the number of hits for measured interests than for expressed interests.

A final study to be discussed is that of Reeves and Booth (1979) who examined the predictiveness of expressed and inventoried interests for approximately 2,500 Navv enlistees in the Hospital Corpsman field. All recruits completed both the *Navy Vocational Interest Inventory* and a six-item questionnaire that attempted to measure their expressed commitment to the Hospital Corpsman field. A representative item was: "If you had the opportunity right now to change your job in the Navy, would you do it?" Two years later, information was collected to assign the recruits to either the "effective" or the "ineffective" criterion group. Those rated "effective" had completed training, remained on the job for their first two years, and had advanced in rank, while the "ineffective" did not meet all of these criteria. Analysis of the data showed that inventoried and expressed interests had very similar (.28 and .27) correlations with the criterion as well as contributing similarly to a multiple correlation (.41 and .40) which included aptitude variables. The two methods were correlated .51 with each other, and when both were entered into the multiple regression equation, the multiple correlation with the criterion rose to .42.

Conclusions. It is clear from the convergence of these studies that the more parsimonious expressed choice is at least equal to inventoried interest in predictive accuracy. Given this fact, it may seem that inventories could be abandoned without any major loss in the prediction of vocational behavior. However, many of the individuals who participate in interest assessment do so because they are undecided and have no prominent expressed choice. This is especially applicable among younger people who are beginning their first job, as is often the case in the Army.

In this light, Hansen (1983) reported the following propositions:

1. Even though a person is not knowledgeable about all jobs or occupations, she or he can give informed responses of interest (such as like, indifferent, or dislike) to items about familiar activities.

2. The interest factors underlying familiar activities are the same as those factors underlying unfamiliar activities.

3. Responses to familiar activities (items), therefore, can be used to identify unfamiliar occupational interests.

4. And, the large number of items in interest inventories can provide a thorough sampling of interests.

Thus, if an 18- or 19-year-old is somewhat naive about the world of work and undecided as to her or his role in it, an interest inventory is applicable while expressed interest is not.

Two points appear relevant at this juncture, both having to do with the validity of an inventory for an individual who is undecided about future vocation. First, as was mentioned above, Bartling and Hood (1980) found that the SVIB was equally predictive for decided and undecided college students. Second, Campbell (1971) reported that a number of studies have shown that so-called "flat profiles," which presumably reflect indecision, do not have an important effect upon validity. He concluded that the general elevation of the profile is not as important as the shape. These findings provide empirical support for Hansen's propositions and the effective use of interest inventories with young and undecided individuals.

## Structure, Models, and Theories

Measurement within any subject area is seriously impeded without a conception of the most important dimensions of the area. The formulation of theories and research on structure not only helps in interpreting the results of current measurement but also points out ways to increase the efficacy of future measurement techniques. This process also allows the formation of broad statements which can then be investigated empirically in the ongoing explication of a domain. Researchers in the area of interests have been heavily involved in this process, with both factor analyses and theory building.

### Factor Analyses

Strong and L. L. Thurstone, in the early 1930s, were the first to use factor analyses to investigate the structure of interests. Five factor analyses were conducted with the evolving SVIB, the first by Thurstone and the next four by Strong. These analyses, as well as two others by Carter, Pyles, and Bretnall (1935), found that four or five factors were sufficient to account for all or nearly all of the variation in interests. Strong (1943) believed the inclusion of the fifth factor was questionable and he was also uncomfortable with the labeling of the factors by Thurstone as science, language, people, and business. He wrote, "Naming a factor is largely guessing today" (p. 166). Strong apparently considered the factors to be of value, however, since he used them to guide an occupational grouping scheme designed to aid interpretation of occupational scales.

The next major factor analytic work was by Guilford, Christenson, Bond, and Sutton (1954) who constructed 1,000 activity, self-description, and attitude items and combined them in homogeneous, 10-item tests that were representative of "variables that had been established in previous factor analyses." This inventory was administered to 600 Air Force enlisted men and 600 officers and officer candidates. Orthogonal rotations of the data were carried out with a goal of simple structure and psychological meaningfulness. Twenty-four airman factors and 23 officer factors were identified; 17 were common to both samples.

The most important finding of the study was that the vocational interests factors obtained were directed toward broad vocational stereotypes or occupational classes. These seven factors--mechanical interest, scientific interest, social welfare interest, aesthetic expression interest, clerical interest, business interest, and outdoor work interest--are nearly identical to those incorporated in Holland's (1973) model of the structure of interests. This model has been widely researched and will be discussed in more detail later.

The conclusions of Guilford et al. (1954) are of further interest. They wrote that the structure of interests includes 19 factors that may be considered basic interest factors, upon which the seven vocational classes can be superimposed. Also, each vocational class bears some consistent relationship to the basic interests factors as well as relationships with one another. Finally, they concluded: "The results support well the belief in vocational interest factors as genuine psychological entities.

Our social culture has established firmly in the minds of men the vocational stereotypes represented (p. 29).

More recently, Droege and Padgett (1979) employed occupational analysts from the U.S. Employment Service to develop activity items that covered all worker trait groups and occupational groups in the current *Dictionary of Occupational Titles*. After the items were administered to job applicants, the data were factor analyzed and the factors were named by occupational analysts. The result from this nontraditional procedure and heterogeneous item pool was 11 factors that map well onto those found by Guilford et al. (1954).

Rounds and Dawis (1979) factor analyzed the items of the SVIB, rather than groupings of items as in previous studies. The following five same-sex samples were included: Women in general (*N*=1,000), men in general (*N*=1,000), female occupational (*N*=2,500), male occupational (*N*=3,600), and male rehabilitation client (*N*=3,600). The authors wrote:

> From 11 to 13 definable factors were found, depending on the sample. The factors were for the most part equivalent across samples; when not equivalent, factors for one sample could be mapped onto another (same-sex) sample's factors. These 11-13 factors serve to define the SVIB item domain; they identify the dimensions underlying this set of empirically derived items. (p. 142)

Hansen (1983) wrote that more factors are typically found when analysis is at the item rather than the scale level.

The results of some of the factor analyses mentioned above, as well as three similar studies, have been presented by Roe (1956) in terms of her own classification. They are reported in Table 1 (as extracted from Holland, 1976, p. 530). The similarity across investigations is striking.

As can be noted from the left side of the table, Roe's classification involves eight occupational categories, but her classification is also two-dimensional with six levels for each category. For example, level 1 of the science category includes the occupation of "osteopath," while level 3 includes "weather observer," and level 5 contains "veterinary attendant." All occupations are relevant to science, but they vary as to level of responsibility, education, and skills.

Roe's structure of interests has spawned the construction of two interest inventories, both of which provide support for her model. *The Vocational Interest Inventory* (Lunneborg, 1981) and the RAMAK (Barak & Meir, 1974) have demonstrated both concurrent validity in differentiating between groups, and predictive, criterion-related validity (Barak & Meir, 1974; Lunneborg, 1979; Mitchell et al., 1971). Perhaps the most important contribution of these inventories, however, has been in terms of understanding the structure of interests. The results of several studies (Jones, 1965; Lunneborg, 1978; Meir, 1973; Meir & Ben-Yehuda, 1976) support the idea that Roe's occupational classifications have a circular arrangement, thus adding additional convergence to its already striking similarity with Holland's hexagonal model of interests, described next.

Table 1

The Relation of Roe's Occupational Categories to the Factors From Selected Factor Analyses

| Classification | Vernon | Thurstone | Darley | Strong | Kuder | Guilford et al. |
|---|---|---|---|---|---|---|
| I. Service | Social welfare vs. administrative<br>Gregarious vs. isolated | People | Welfare uplift | People | Social service | Social welfare |
| II. Business Contact | Gregarious vs. isolated | People | Business contact | Business | Persuasive | Business |
| III. Organization | Administrative vs. social welfare | Business | Business detail<br>CPA | Business system | Clerical<br>Computational | Business<br>Clerical |
| IV. Technology | Scientific vs. display<br>Isolated vs. gregarious | Science | Technical | Things vs. people | Scientific<br>Mechanical<br>Computational | Scientific<br>Mechanical |
| V. Outdoor | Active vs. verbal | | | | Outdoor | Physical drive<br>Preference for outdoor work |
| VI. Science | Scientific vs. display<br>Isolated vs. gregarious | Science | Technical | Things vs. people | Scientific | Scientific |
| VII. General Cultural | Verbal vs. active | Language | Verbal | Language | Literary | Cultural |
| VIII. Arts and Entertainment | Display vs. scientific | Language | Verbal | Language | Artistic<br>Musical<br>Literary | Aesthetic expression<br>Aesthetic appreciation<br>Cultural<br>Physical drive in some active vs. verbal |

*Note*: From "Vocational Preferences" by J. L. Holland, in *Handbook of Industrial and Organizational Psychology*, M. D. Dunnette (Ed.). Copyright 1976 by Rand McNally College Publishing Company. Reprinted by permission.

## Holland's Hexagonal Model

Description. Holland (1973) summarized his model in terms of its four working assumptions:

1. In our culture, most persons can be categorized as one of six types: realistic, investigative, artistic, social, enterprising, or conventional.

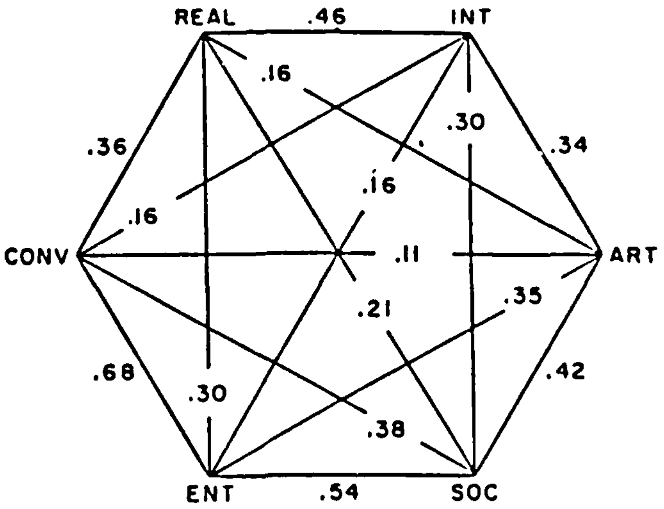
2. There are six kinds of environments: realistic, investigative, artistic, social, enterprising, or conventional.

3. People search for environments that will let them exercise their skills and abilities, express their attitudes and values, and take on agreeable problems and roles.

4. A person's behavior is determined by an interaction between his personality and the characteristics of his environment.

The relationships among the six types and environments in the R-I-A-S-E-C model can be illustrated by a hexagon as diagrammed in Figure 1.

***A Hexagonal Model for Defining the Psychological Resemblances Among Types and Environments and Their Interactions***



*Source*: Holland, J.L., Whitney, D.R., Cole, N.S., and Richards, J.M., Jr. An empirical occupational classification derived from a theory of personality and intended for practice and research. ACT Research Report No. 29. Iowa City: The American College Testing Program, 1969.

Figure 1. Hexagonal model. *From Making Vocational Choices: A Theory of Careers*, by G. L. Holland, 1973, Prentice-Hall, Englewood Cliffs, N.J., page 7. Copyright 1969 by the American College Testing Program. Reprinted by permission.

The lines between the types show both the correlation and the distance between two types. Thus, artistic and conventional are very different while enterprising and conventional are quite similar. This degree of relatedness is called consistency by Holland. Other "secondary constructs" that are important in understanding the theory are differentiation, which is the degree to which a person or environment resembles many types or only a single type; congruence, the degree to which a type is matched with its environment; and calculus, the degree to which the internal relationships of the theory fit a geometric hexagon.

Research Findings. Holland, Magoon, and Spokane (1981), in their *Annual Review of Psychology* chapter, reported that approximately 300 empirical studies regarding Holland's theory were conducted during the period 1964 to 1979. They concluded that, although the evidence for the secondary constructs is mixed, the basic person-environment typology has been strongly supported in tests of its organizing power. This is basically the same conclusion drawn by Walsh (1979) in his review of the literature. A number of selected studies are cited to represent the essence of this research.

Wall, Osipow, and Ashby (1967), using a sample of 186 male college freshmen, demonstrated that a student's ranking of six descriptions of the Holland types according to his resemblance to each was significantly related to his SVIB group score. The students tended to see themselves in ways that correspond with their interest scores. Lee and Hedahl (1973) found that basic interest scale scores were usually highest for scales associated with the individual's vocational type; for instance, enterprising students had the highest mean on public speaking.

Cole and Hansen (1971) compared the scales of several interest inventories--the SVIB, the *Kuder Occupational Interest Survey*, the *Minnesota Vocational Interest Inventory*, the *ACT Vocational Interest Profile*, and the *Vocational Preference Inventory*--by applying configural analysis to each inventory. "The configurations of the scales for all inventories were found to be similar and to conform to the circular configurations of interest proposed by Roe and Holland" (p. 485). Edwards and Whitney (1972), using the Cole configural method and factor analysis of Self-Directed Search data, found that Holland's six types were also related to a person's activities, competencies, occupations, and self-ratings.

Nafziger and Helms (1974) applied a clustering technique to the scales of the *Kuder Occupational Interest Survey*, the *Minnesota Vocational Interest Inventory*, and the *Strong Vocational Interest Blank* for both men and women. The results showed the existence of a few broad, internally consistent groups of occupations which follow the hexagonal ordering of occupational categories. The clusters for all inventories, for both sexes, possessed similar Holland codes, although the MVII, which is a "blue collar" inventory, contains primarily Realistic occupations. The Holland model was also supported in a multidimensional scaling analysis by Rounds, Davison, and Dawis (1979) but differentially for males and females. For males, the interrelationships among the SVIB scales were found to conform to Holland's hexagon, but this was much less evident for females. "The same item sets apparently do not have identical meaning for both sexes" (p. 312).

Gottfredson (1980) classified each of 437 census occupational titles in terms of Holland's typology, an occupational prestige scale, an occupational self-direction scale, the *Dictionary of Occupational Titles*, the Census Bureau classification, and, for 120 of the titles, occupational reinforcer patterns. Three main conclusions were drawn from her analyses: (a) the construct validity of Holland's occupational scheme is supported; (b) it is misleading to ignore differences in occupational level, as has generally been done in tests of Holland's typology; and (c) greater specificity of constructs (theoretical predictions) is needed.

Gati (1982) postulated that previous tests of the validity of Holland's model were not rigorous enough since the null hypothesis is set up in such a way that the proportion of correct predictions will be around .5. Therefore, he set up a null hypothesis that the hexagonal model is totally valid, that is, the proportion of correct predictions derived from it is 1.00. Using this more rigorous approach, 13 data sets of published studies were reanalyzed. Eleven of these studies supported the hexagonal model with the less rigorous approach but only two supported it when the alternate method was used. The author stated that the hexagonal model is a rough approximation of the structure of interests but that a hierarchical model (Gati, 1979) would justify stronger inferences. Gati's hierarchical model appears promising although further investigation is needed.

Generalizability. A number of studies have investigated the applicability of Holland's theory for women and minorities. Prediger and Hanson (1976) found substantial stereotypic differences in Holland's raw scores for males and females in the same occupation. As mentioned, Rounds et al., (1979) found the model fit less well for females than for males. Bingham and Walsh (1978) concluded from a review of five other studies (Harvey & Whinfield, 1973; Horton & Walsh, 1976; Matthews & Walsh, 1978; Werner, 1969; Wiggins, 1976) that the model is tentatively supported for employed women. In their own study, Bingham and Walsh found some support for Holland's theory with employed, college-degreed, black women. The women of each race were found to be far more similar than different on the scales of *Self-Directed Search* but the relationship was opposite on the scales of the *Vocational Preference Inventory*.

Wakefield, Yom, Doughtie, Chang, and Alston (1975) found that *Vocational Preference Inventory* scales for black subjects corresponded generally to Holland's model but not as well as they did for white subjects. They wrote that "what is surprising is that [black results] do approximate the [Holland] model with only a few weaknesses." Lamb (1976) reported that basic interest scale configurations of blacks correspond to those of the same-sex, white sample and follow the hexagonal model. Harrington and O'Shea (1980) assessed the interest of 267 Spanish-speaking Americans (Mexican American, Puerto Rican, Cuban, South American) with a translation of the *Career Decision Making Inventory* and found that the results were consistent with the Holland model.

Secondary Constructs. Holland's secondary constructs of congruence, consistency, differentiation, and calculus have also been examined empirically. Walsh, Howard, O'Brien, Santa-Maria, and Edmondson (1973) investigated the differences on the variables of satisfaction, self-concept, self-acceptance, and vocational maturity between college students whose occupational choices were either congruent or incongruent with their assessed Holland type. Initially, congruence was found to be unrelated to the other variables, but a more rigorous definition of congruence requiring the individual's two highest Holland types to be identical with the two highest types of the occupational choice did show a relationship with satisfaction. Thus, there may be differing levels of congruence that affect the success of the person-environment match. Mount and Muchinsky (1978) found that significantly more subjects were in congruent pairings of type and occupation than incongruent. The incongruent subjects were not, however, more concentrated in adjacent occupational typologies as was predicted by Holland.

O'Neil and Magoon (1977) used a sample of male, "Investigative" college students to explore the efficacy of Holland's three consistency levels in predicting later occupational status. According to the theory, those individuals who are more "consistent" are also more predictable in occupational membership. Differential predictability of the levels was supported in 20 of the 27 tests applied.

Peiser and Meir (1978) found that both consistency and congruence were positively correlated with occupational choice satisfaction (OCS) measured 7 years later for males and females. This result was followed up by Gati and Meir (1982) who found that congruence and consistency were of similar efficiency in predicting OCS whether used in Holland's model or a hierarchical model. Guthrie and Herman (1982) found that congruence did relate significantly to vocational maturity but that differentiation and consistency did not. Similarly, Villwock, Schnitzen, and Carbonari (1976) found that all three constructs predict stability of occupational choice but the efficiency of prediction is not improved by adding differentiation and/or consistency to congruence. Also, the three constructs in a multiple regression equation accounted for only 13.7 percent of the variance in predicting stability.

The construct of calculus had received less attention; in fact, the term is rarely used. Basically, the research suggests that, while a hexagon may be a useful approximation of the structure of interests, it is not necessarily regular or equilateral (Edwards & Whitney, 1972; Rounds et al., 1979). Holland (1979) noted that "at best, the hexagons resulting from real-world data are misshapen polygons . . . the hexagon is an ideal" (p. 43).

Summary and Integration. The core of Holland's theory--that is, the R-I-A-S-E-C model of persons and environments--has been well supported by a very large number of studies with a myriad of samples, criteria, and instruments. Of the secondary constructs, congruence, or the degree to which a person's type matches her or his environment, has received the most support. The evidence concerning all of the secondary constructs is mixed, however, implying that additional specification is necessary for these constructs to be considered integral to acceptance of the theory.

The theory has had tremendous influence on vocational interest measurement in at least four respects (Hansen, 1983). First, the theory has prompted development of inventories and sets of scales to measure the six types. Second, it has stimulated extensive research on many aspects of vocational interests. Third, it has integrated and organized the relevant information under one system. Fourth, it has provided a simple structure of the world of work which is amenable to career assistance.

Prediger (1982) has attempted to uncover the dimensions that underlie the Holland hexagon and thus explain why the link between interests and occupations exists. He reported two studies that provide strong support for the assertion that interest inventories work because they tap activity preferences that parallel work tasks. In the first study, results showed that two dimensions, things/people and data/ideas, accounted for a mean value of 60 percent of the nonresponse-set variance among a very large data set of interest scores.

In the second study, the procedure involved recording interest inventory scores and work-task scores (from job analysis) for 78 occupations in terms of the data/ideas and things/people work task dimensions. These dimensions were found to be essentially independent (correlation of -.13). Pearson product-moment correlations were then calculated to determine relationships between the vocational interests of persons and the work tasks which characterize their occupations. These correlations ranged from the upper .60s to lower .80s.

Prediger provided a valuable link in an understanding of what underlies Holland's model and how the model achieves its matching of interests and occupations. The emphasis is on activities and whether an individual's desired activities match an occupation's required activities. If indeed the reason that types and environments match is their similar activities, then such factors as prestige and security become less important to the Holland theory. These factors, and others perceived as needs and reinforcers are, however, among the main features of an alternate theory of vocational behavior which has been proposed by Dawis, Lofquist, and Weiss (1968) and is described below.

## The Theory of Work Adjustment

Description. Like Holland's theory, the Theory of Work Adjustment (TWA) (Dawis, Lofquist, & Weiss, 1968) is based on the concept of correspondence between individual and environment. There are, however, important differences that can be illustrated by referral to Figure 2.

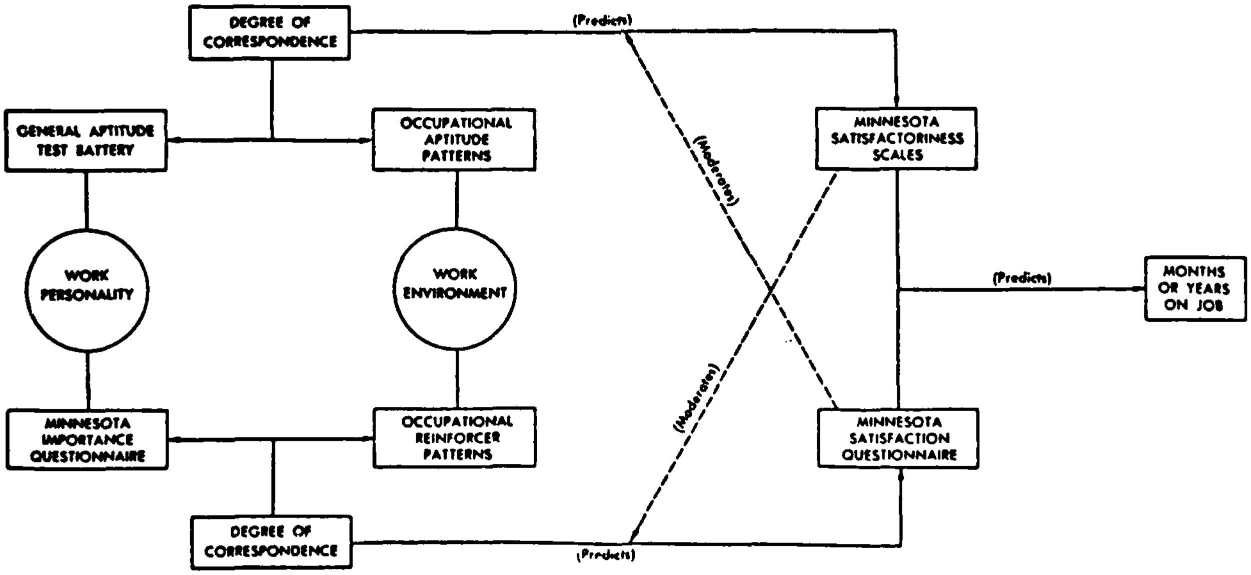

DEGREE OF CORRESPONDENCE
(Predicts)
GENERAL APTITUDE TEST BATTERY
OCCUPATIONAL APTITUDE PATTERNS
MINNESOTA SATISFACTORINESS SCALES
(Moderates)
WORK PERSONALITY
WORK ENVIRONMENT
(Predicts)
MONTHS OR YEARS ON JOB
(Moderates)
MINNESOTA IMPORTANCE QUESTIONNAIRE
OCCUPATIONAL REINFORCER PATTERNS
MINNESOTA SATISFACTION QUESTIONNAIRE
DEGREE OF CORRESPONDENCE
(Predicts)

Figure 2. The Theory of Work Adjustment in Operational Terms. From "A Theory of Work Adjustment," by R. V. Davis, L. M. Lufquist, and D. J. Weiss, *Minnesota Studies in Vocational Rehabilitation*, Bulletin 23, 1968, p. 23. Copyright 1968 by *Minnesota Studies in Vocational Rehabilitation*. Reprinted by permission.

As depicted in the far left of the diagram, an individual's work personality or suitability for a job is assessed in two parts. A cognitive test measures the individual's aptitudes and an instrument such as the *Minnesota Importance Questionnaire* (MIQ) measures his or her needs for various job rewards (or reinforcers). The work environment is also measured in two parts: the aptitudes necessary to do the work and the rewards offered by the job. The degree of correspondence can then be calculated between the individual's aptitudes and needs and the job's requirements and rewards. High correspondence between aptitudes and requirements is associated with satisfactoriness of job performance, and high correspondence between needs and reinforcers is associated with job satisfaction. Also, as shown, the degree of satisfactoriness moderates the relationship with job satisfaction and vice versa, ultimately resulting in a prediction of tenure.

Thus, the TWA differs from Holland's model in the way in which it assesses both the work environment and the individual. Unlike the hexagonal approach, the TWA model includes measures of aptitude, and assesses vocational needs rather than interests. Some researchers have hypothesized that vocational needs are determinants of vocational interests (Bordin, Nachman, & Segal, 1963; Darley & Hagenah, 1955; Roe & Siegelman, 1964; Schaeffer, 1953) while others believe that needs and interests reflect the operation of similar underlying variables (Strong, 1943; Thorndike, Weiss, & Dawis, 1968). Rounds (1981) has shown that needs and interests measure different aspects of vocational behavior.

The TWA specifies a set of 20 reinforcing conditions that can be used to assess both vocational needs and occupational rewards, and thus predict job satisfaction and satisfactoriness of performance. For example, the reinforcing condition may be security of employment and those individuals with a high need for security are best suited for an occupation that has very stable employment. These 20 reinforcers were identified through literature review and analysis of the set of items representative of the reinforcer dimensions. The following are examples of the reinforcing conditions, with brief descriptions:

Recognition    Work environment in which rewards are forthcoming for praiseworthy individual performance.

Authority    Tasks which include or consist of power to decide the methods by which a job is performed and to impose those decisions on co-workers.

Independence[1]    Tasks which can be executed from beginning to end by one individual.

[1] The other conditions are: Ability Utilization, Achievement, Activity, Advancement, Company Policies, Compensation, Co-workers, Creativity, Moral Values, Responsibility, Security, Social Service, Social Status, Supervision-Human Relations, Supervision-Technical, Variety, and Working Conditions.

The TWA also contains a set of 15 formal propositions that are general laws about an aspect of work adjustment and serve as a basis for further research. These propositions cover areas of vocational behavior ranging from turnover to personality style, and give attention to many of the process aspects of vocational behavior as it develops and is maintained. The comprehensiveness and specificity of the propositions allow a holistic appraisal of vocational behavior while simultaneously maintaining scientific rigor. Thus, the propositions form the bulwark of the TWA and represent an elaborate set of tested and testable hypotheses, the equivalent of which is seldom found in the behavioral sciences.[2]

Research Findings. Much of the research concerning the TWA has been directed toward verification of the theory's propositions. Lofquist and Dawis (1969) reported a number of studies which support the first four propositions:

| | |
|---|---|
| Proposition I: | Satisfaction and satisfactoriness are independent indicators of work adjustment. |
| Proposition II: | Satisfactoriness is a function of the correspondence between an individual's abilities and the ability requirements of the work environment. |
| Proposition III: | Satisfaction is a function of the correspondence between the reinforcer system of the work environment and the individual's needs. |
| Proposition IV: | Satisfaction moderates the functional relationship between satisfactoriness and the correspondence of the individual's abilities with the ability requirements of the work environment. |

More recently, much of the research emphasis has focused on the process aspects of work adjustment. Dawis and Lofquist (1976) extended the TWA to include the concepts of work personality style and work environment style. These concepts characterize the day-to-day interactions of individuals and their environments in terms of their mutual responsiveness. Also described was a systems-type model that integrates the early propositions of the theory with the more recent work on personality-style dimensions (Dawis & Lofquist, 1978). The research represented in these two studies constitutes a major contribution to an understanding of the dynamics of work adjustment rather than a more static model of traits and factors alone.

Rounds (1981) has reviewed 11 studies from the period 1968 to 1981 which utilize job satisfaction (Proposition III) as the dependent variable. He drew the following conclusions: (a) Need-reinforcer correspondence indices have been found to have a consistent, significant relationship with

[2] The interested reader can obtain additional information about the TWA by consulting the recent book *A Psychological Theory of Work Adjustment* by Dawis and Lofquist, 1984. Minneapolis, MN: University of Minnesota Press.

job satisfaction. (b) The relationship between correspondence and job satisfaction ranges from .18 to .42 with the use of a profile-shape index. (c) Studies using a longitudinal design have resulted in the best predictions of job satisfaction while studies with a restricted range of occupations and tenure have failed to find a relationship. It appears quite well replicated that the TWA approach will yield criterion-related validities with job satisfaction that range from moderate to quite good. Further, research reported by Scarpello and Campbell (1983) suggests that individual differences in aspiration level and different views of career progression help explain the level of job satisfaction over and above the match of needs and rewards.

Conclusions. Although the TWA is associated with approximately 200 research studies over a 20-year period, it is still relatively unknown. This may be attributable to the fact that many of the studies pertinent to the TWA have been published in research monographs and government publications not readily accessible to a wide range of professionals in the area of vocational behavior. This general lack of exposure is unfortunate since, as Holland, Magoon, and Spokane (1981) wrote:

> The Minnesota work (TWA) is one of only a few programmatic efforts to comprehend a major element of vocational life. Both practitioners and researchers can benefit from a reading of perhaps the most analytical and hard-headed analysis of person-job interactions. (p. 297)

## Validity Research

While numerous procedures are available for evaluating validity, the APA standards (1974) recognize three principal classifications: content, criterion-related (concurrent and predictive), and construct. Each of these methods has been used to examine the validity of interest assessment and each has made its own contribution within that process. In this subsection, the results of a number of studies characteristic of each procedure are reviewed and summarized.

### Content Validity

Guion (1976) described content validity as the degree to which the content and format of a test correspond to a domain of relatively clear-cut knowledge or behavior. Content validity is of special concern during the construction of an interest inventory when the form's author must determine how to achieve the desired degree of correspondence with the interests domain. As noted by Lammlein (1983), this process involves three separate questions. First, are all or most of the important areas of knowledge or behavior in the domain represented in the inventory (representativeness)? Second, given that they are represented, do these important areas receive an appropriate amount of emphasis in the inventory (fidelity)? Thirdly, is the item endorsement rate of the sample (or inventory) comparable to that of the domain?

Since the structure of the domain of interests appears quite well documented (Holland, 1976), it would seem that an inventory measuring each of these important interest dimensions would be representative. If each of the dimensions and their subareas received the appropriate amount of emphasis, it would also have good fidelity. Finally, if the various item response options of the inventory were endorsed with a frequency comparable to that expected in the population for each area of the domain, then all content validity concerns would be addressed.

Unfortunately, a systematic description of how these procedures have been followed is, for the most part, lacking from interest inventories. Several authors have reported as content validity what is more accurately labeled face validity. One study that did address the item endorsement question demonstrated that the items of the *Vocational Interest Career Examination* (VOICE) were understandable to the members of the population taking the inventory, thus allowing reasonable responding to the items (Alley & Matthews, 1982). It seems likely that many interest inventories do have good content validity and that future research could demonstrate this aspect of their utility fairly easily. Such research not only would be of value for content validity, but would likely contribute to criterion-related and construct validity as well.

## Criterion-Related Validity

Criterion-related validity, which is the assessment of the relationship between an individual difference measure and an external criterion, follows one of two designs: concurrent or predictive. In a concurrent design, criterion data are collected at the same time as test scores, while in a predictive design, criterion data are obtained at a later time. Both of these designs have been used with assessed interests, for example, in differentiating between occupational groups (concurrent) and in forecasting later occupation (predictive).

Before examining criterion-related validity results in more detail, two points are noted. First, the criteria against which interests are assessed all are related to jobs or occupations. These occupations, the way in which they are performed, and their interrelationships can vary greatly, thus defying any universal and clear-cut classification (Campbell, 1971). Also, Schmidt (1974) has shown that the validity of an interest inventory varies as a function of base rate and the relative values of true and false positives and negatives, factors that have sometimes been overlooked or misinterpreted in validity studies. Because of these effects and the various sources of error typically found in applied research, some degree of fluctuation is evident in the results reported by different investigators. The overall convergence of findings, however, testifies to the robustness of the validity associated with this domain.

### Occupational Membership.

In criterion-related validity studies using measured interests, the most popular dependent variable has been occupational membership. Concurrent validity studies with this criterion have investigated whether members of different occupations respond differently to interest items (e.g., do engineers receive their highest scores on scales related to engineering and do members of unrelated occupations receive lower scores on these scales?). It should be noted that the use of Strong's empirical method of scale construction assures a degree of concurrent validity, since only those items that discriminated between occupational members and people-in-general are included on the occupational scales.

Campbell and Hansen (1981) reported the concurrent validity of the *Strong-Campbell Interest Inventory* in terms of the percent overlap (Tilton, 1937) of the interest score distributions of occupational members and people-in-general. If a scale discriminates perfectly between the two groups there is zero overlap, and if it does not discriminate at all there is 100 percent overlap. Values for the occupational scales of the SCII range from 16 percent to 54 percent with a median of 34 percent overlap. This degree of overlap means that the scales are separating the two groups by about two standard deviations, on the average.

A second method of assessing concurrent validity is to determine how well the scales separate occupations from each other rather than from people-in-general. Campbell (1971) reported the SVIB scores of more than 50 occupations on each other's scales for both men and women. The median score for members of one occupation on the scales of other occupations was 25, as compared with a median score of 50 for members of an occupation on their own scale (standardized scores with mean of 50 and standard deviation of 10).

Kuder and Diamond (1979) reported the concurrent validity of the *Kuder Occupational Interest Survey* scales in terms of the percentage of subjects who scored higher on their own occupational scale than on each of five other occupational scales. The higher score on their own scale occurred 71 percent of the time when a general reference group had been used in scale construction and 81 percent of the time when lambda coefficients were used. In a second study that involved 30 occupational groups, the median correlation was .84 between highest score on the 30 scales and actual group membership.

In addition to the studies with occupational scales, concurrent validity has also been demonstrated with homogeneous, basic interest scales. Campbell and Hansen (1981) reported that the SCII basic interests scales spread occupations' scores over 2-2.5 standard deviations, and that the patterns of high- and low-scoring occupations are substantially related to the occupations that people pursue. For example, astronauts were among the highest scorers on the "Adventure" scale and bankers were among the lowest. Similarly, interior decorators were among the highest scorers on the "Art" scale while farmers were among the lowest.

It has been repeatedly shown that members of occupations with similar knowledge, skill, and ability requirements tend to respond more similarly to interest items than members of unrelated occupations. Kuder (1977) reported that the interest scores of architects are very similar to those of interior decorators and very dissimilar from those of truck drivers and carpenters. The largest difference score found was between machinists and college women majoring in drama, while the most similar occupations were carpenter and auto mechanic. Similar results have been reported by numerous investigators with a number of different inventories (Campbell & Hansen, 1981; Echternacht, Reilly, & McCaffrey, 1973; Gottfredson, Holland, & Holland, 1978; Holland & Holland, 1977; Jackson, 1977; Strong, 1943, 1955).

The predictive validity of measured interests in forecasting later occupational membership has also been extensively investigated. Perhaps the classic study of this type was given in Strong's 1955 book which reported in detail an 18-year followup of 663 Stanford University students who had completed the *Strong Vocational Interest Blank* while still in college. Strong found that subjects averaged 43.6 on the scale of the occupation engaged in 18 years later (a standard score of 45 or above received an "A" rating, the highest rating obtainable) as compared with 48.9 for concurrently derived criterion groups, and 30.2 for men-in-general. He also concluded that for those with an "A" rating for an occupation, the expectancy ratio or odds of entering that occupations were about 3.5 to 1. This expectancy ratio has been shown by Dolliver (1969a) to be problematic and the hit rate of the SVIB/SCII and other inventories is now reported in terms of the percentage of individuals who have a "good" [McArthur's (1954) classification system] correspondence between their interest scores and the occupation ultimately entered. Campbell and Hansen (1981) reviewed eight predictive validity studies and concluded that the "good" hit rate of the Strong inventories is approximately 50 percent, the "poor" hit rate is around 15 percent, and the remaining 35 percent of the predictions are "clean misses."

The following is a sample of hit rates obtained in a variety of additional predictive studies of occupational membership. For the *Kuder Preference Record*, hit rates were: 53% in a 25-year followup (Zytowski, 1974), 63% in a 7- to 10-year followup (McRae, 1959), and 51% in a 12- to 19-year followup (Zytowski, 1976). Lau and Abrahams (1971) reported 68% in a 6-year followup with the *Navy Vocational Interest Inventory* while Gottfredson and Holland (1975) found an average 50% hit rate in a 3-year followup with the *Self-Directed Search*. The *Strong Vocational Interest Blank* hit rates were 39% (corrected for base rate) in an 18-year followup (Worthington & Dolliver, 1977) and 50% in a 10-year followup (Campbell, 1971). Gade and Soliah (1975) reported 74% for graduation of college majors in a 4-year followup with the *Vocational Preference Inventory*. Hit rates from a sampling of studies using expressed interest are also available: 69% in a 9-year followup (Strong, 1943), 52% in a 3-year followup (Borgen & Seling, 1978), 72% in an 11-year followup (Bartling & Hood, 1980), and 46% in a 21-year followup (Cairo, 1982).

The results of these studies illustrated not only the variability normally found with different researchers and instruments but also effects of varying base rates, classification systems, and extraneous factors that were mentioned earlier. Consistent throughout the studies, however, is the demonstrated relationship between an individual's interests and the tendency to stay in a related job or occupation. This consistency is even more remarkable in light of the extensive followup periods and the fact that interests are the only variable being considered, disregarding the effects of ability and economic change.

Job Involvement/Satisfaction. Strong (1943, p. 385) wrote that he "can think of no better criterion for a vocational interest test than that of satisfaction enduring over a period of time." Thus, he and subsequent researchers of the validity of interest measurement have normally incorporated satisfaction into studies of occupational membership by assuming that continuance in an occupation is a measure of satisfactory adjustment.

The relationship between interests and job satisfaction has also been studied directly, both correlationally and through group differentiation. Since correlational studies are easily summarized in tabular form, these results and the results of correlational studies of job proficiency and training performance are summarized in Tables 2-4 and listed individually in Appendix C as well. Noncorrelational studies for each of these criteria are discussed only in the text.

Table 2

Criterion-Related Validites: Job Involvement

| | Job Satisfaction | Re-enlistment Turnover | Occupational Membership |
|---|---|---|---|
| Number of Studies | 18 | 3 | This criterion has been studied using hit rates and group comparisons and is discussed in the text, rather than tabularized. |
| Median Correlation | | | |
| Overall | .31 | .29 | |
| Predictive | .23 | .29 | |
| Concurrent | .33 | | |
| Correlation Range | .01 - .62 | .19 - .29 | |
| N Range | 25 - 18,207 | 125 - 789 | |
| Median N | 501 | 520 | |

Table 2 summarizes a total of 21 correlational studies of job involvement, the overwhelming proportion of which were directed toward job satisfaction. It can be seen that the studies generally reported correlations of around .30, with concurrent investigations obtaining somewhat higher validities, as would be expected. Six of the studies were conducted with military personnel and obtained almost exactly the same results as with civilians. Overall, the investigations reported remarkably similar results, as evidenced by almost half of the median validities falling in the range of .25 to .35. Finally, since the median sample size of these studies exceeds 500, considerable confidence can be placed in their replicability.

In noncorrelational studies, Kuder (1977) examined the differences in interests between satisfied members of occupations and their dissatisfied counterparts. He concluded that members of the dissatisfied group are much more likely to receive higher scores in other occupations than are the members of the satisfied group, and the dissatisfied group is more heterogeneous than the satisfied group in the same occupation.

A number of studies have also shown that group differences in interests are related to differences in job satisfaction. For example, McRae (1959) studied 1,164 young people from 31 states whose interests had been measured in high school and who responded 7 to 10 years later to a job satisfaction questionnaire. McRae found that of those in an occupation consistent with their earlier interests, 62 percent were satisfied as compared with 34 percent of those in an occupation inconsistent with their earlier interests. Similar results have been reported by A. M. Brayfield (1942), A. H. Brayfield (1953), DiMichael and Dabelstein (1947), Hahn and Williams (1945), Herzberg and Russell (1953), North (1958), and Trimble (1965), among others. Lastly, Arvey and Dewhirst (1979) found that general diversity of interests was related to job satisfaction.

A number of other studies, however, have found no significant relationship between interests and job satisfaction. These include Butler, Crinnion, and Martin (1972), Dolliver et al., (1972), Schletzer (1966), Trimble (1965), and Zytowski (1976).

Campbell (1971) wrote that the generally modest relationships reported between interests and job satisfaction may be due to restriction in range. Studies have generally reported a high percentage of satisfied workers (around 80%), thus resulting in very little criterion variance. Another viewpoint, expressed by Strong (1955), is that job satisfaction is such a complex and variable concept that a good measure of it is unobtainable. With these factors in mind, the smaller number of studies with negative and modest findings seem less at odds with the preponderance of the evidence, and the intuitive correspondence between interests, job satisfaction, and sustained occupational membership appears more reasonable.

Job Proficiency. Strong (1943, p. 485) wrote that "The man on the street or in the shop emphasizes interest or willingness to work as the most important of all factors explaining (job) success. Can such claims be justified?" Strong's own investigation with life insurance agents showed that successful agents did receive higher interest scores on relevant scales than unsuccessful agents and that the correlation between production and interest scores was approximately .40. He added that many of those with low scores did not stay in the occupation, thus restricting the range in the predictor and also suggesting an interaction effect.

A related study by Ferguson (1958) reported that,

> For all agents we find no difference between the termination rates of those with the interests of successful salesmen and those without. But for agents whose performance is average or above average, we find that those with interests similar to successful salesmen terminated less frequently than those without. In other words, when the ability (performance) differential drops out, the interest differential takes over. (p. 191)

This research suggests a three-way interaction among job performance, occupational membership, and interests in which each may moderate the relationship between the others. Similarly, Clark (1961) used ability as a mediator variable and found interest scores more predictive of job performance at some ability levels than others.

A number of studies have shown that measured interests can differentiate between those rated successful and unsuccessful within an occupation. For example, Abrahams, Neumann, and Rimland (1973) found that the highest interest quartile contained three times as many Navy recruiters rated effective as the lowest quartile. Similarly, Azen, Snibbe, and Montgomery (1973) showed that interest scores correctly classified 67 percent of a sample of deputy sheriffs in terms of job performance ratings.

Arvey and Dewhirst (1979) demonstrated that general diversity of interests was positively related to salary level. Also, Campbell (1965, 1971) reported that past presidents of the American Psychological Association, who have enjoyed outstanding professional success, have higher interest scores in the physical sciences than psychologists-at-large, who have higher interest scores in the social sciences. This finding might also be attributable to the importance of diversity of interests in job performance, and illustrates an important aspect of interests' validity in differentiating among individuals at differing performance levels.

In correlational research, a total of 14 studies were located and are summarized in Table 3 as well as listed individually in Appendix C. The majority of the research utilized ratings as the measure of job performance, although three studies examined interests' relationship with archival production.

Table 3

Criterion-Related Validities: Job Proficiency

| | Ratings | Job Knowledge Tests | Archival Production |
|---|---|---|---|
| Number of Studies | 11 | 0 | 3 |
| Median Correlation | | | |
| Overall | .20 | -- | .33 |
| Predictive | .20 | -- | .33 |
| Concurrent | .25 | -- | |
| Correlation Range | .01 - .40 | -- | .24 - .53 |
| N Range | 50 - 2,400 | -- | 37 - 195 |
| Median N | 464 | -- | 116 |

Median validities are .20 for studies employing ratings as criteria and .33 for the archival production investigations. These values suggest a range of .20 to .30 for the overall correlation between measured interests and job performance. Both the number of studies and the median sample size are higher within the rating criterion subcategory. Since the values obtained in this subcategory are lower than for archival production, they may represent a more stable and conservative estimate of validity for predicting job performance from measured interests.

Training Performance. The training performance criterion category includes objective tests of training knowledge, subjective ratings by instructors, completion versus noncompletion of the training program (go/no-go), and hands-on measures of learning. It is difficult in each of these subcategories to draw an absolute distinction between training and academic performance. The research to be discussed here includes some studies that might be labeled academic rather than training, but the two are combined because they are highly similar and for simplicity of discussion.

Table 4 summarizes the 13 correlational studies located that were thought to be most relevant to the training criterion category. All were predictive investigations, yet a fair amount of variability is found across subcategory medians. Ratings of training performance were predicted best (.35) and objective measures predicted least well (.17), while a median of .28 was found for studies of course completion/noncompletion. Eight investigations utilized military samples and generally obtained higher validities, with a median value of .28. In addition, the military studies included the use of much larger samples and are more consistent with a training rather than academic emphasis. Thus, it seems reasonable to expect the correlation between interest and later training performance to generally fall around .25, and perhaps higher with instruments specifically constructed for a given set of training programs or jobs (as was often the case in the military research).

Research in more academic settings has also suggested a fairly modest degree of relationship with vocational interests. Strong (1943), after reviewing the results from a large number of studies, concluded that (academic) scholarship is more closely associated with intelligence than with interests. He adds, however, that very few study an area successfully if they have low interest in it. This conclusion is supported by more recent research (Barak & Rabbi, 1982; Wiley & Magoon, 1982) that suggests the consistency of interests (Holland, 1973) is a mediator variable in the relationship between interests and academic success.

Campbell (1971) reported that the Academic Achievement scale of the SVIB yielded a cross-validated correlation of .36 with first-year college grade-point average. It also increased the multiple correlation when added to either high school rank or aptitude scores. The scale, now entitled the Academic Comfort scale on the SCII, was reported by Campbell and Hansen (1981) to obtain correlations with grades ranging from .10 to .30. Finally, in a more training-oriented, noncorrelational study, Doll, Ambler, Lane, and Bale (1972) found that interest scores could successfully differentiate between Naval aviation cadets who would voluntarily drop out of training and those who would not. A similar finding was reported earlier by Rosenberg and Izard (1954).

Table 4

Criterion-Related Validities: Training

| | Objective Measures | Subjective Measures | Go/No-Go (Course Completion) | Hands-On Measures |
|---|---|---|---|---|
| Number of Studies | 8 | 2 | 3 | 0 |
| Median Correlation | | | | |
| Overall | .17 | .35 | .28 | -- |
| Predictive | .17 | .35 | .28 | -- |
| Concurrent | -- | -- | -- | -- |
| Correlation Range | .02 - .43 | .28 - .41 | .23 - .42 | -- |
| N Range | 53 - 3,505 | 27 - 373 | 355 - 4,502 | -- |
| Median N | 751 | | 593 | -- |

Summary of Criterion-Related Validity. More than 100 studies of different aspects of the criterion-related validity of measured interests have been reported in this subsection. Among the most thoroughly replicated findings in these studies is the substantial relationship between individuals' interests and their sustained membership in an occupation. The relationship has been demonstrated for a wide variety of occupations and over lengthy periods of time. As such, this finding ranks as one of the most theoretically and practically important discoveries in the broad area of individual differences. In addition to the relationship of interests with occupational membership, research has shown these preferences to have validity for various aspects of job involvement, job proficiency, and training performance. The degree of this association generally appears to be in the .20 to .30 range, expressed in correlational terms. Much research has been conducted with military personnel, and the validities appear to be as high, or often higher, for these individuals and settings. Finally, it has also been well replicated that measures of interests combine well with other types of predictors in yielding an optimal multiple correlation.

## Construct Validity

Construct validity is a broad and abstract process that includes the accumulation of evidence from a number of different sources; it is not accomplished in a single study. Of the many types of evidence relevant to construct validity, the following have been investigated within the area of interests:

1. The internal consistency of interest scales.

2. The relationship of the interest scales of one inventory with the scales of another.

3. The factor-analytic structure of a group of interest items or scales.

4. The processes through which interests change and are expressed.

5. The relationship between interests and behavior in an experimental study.

6. The ability of interest measures to differentiate between groups (e.g., occupations).

7. The convergent and divergent validity of interest scores, that is, related on measures of the same construct and unrelated on measures of different constructs.

8. The rationally judged relationship between the content of interest inventories and relevant behavior.

Internal Consistency. Alley and Matthews (1982) have reported that the internal consistency (coefficient alpha) values of the scales of the *Vocational Interest Career Examination* range from the high .80s to mid .90s. Similarly, the coefficient theta values of the *Jackson Vocational Interest Survey* (Jackson, 1977) range from .70 to .91 with a median of .83. Gottfredson, et al. (1978) found that the KR-20 values for the scales of the *Vocational Preference Inventory* range from .85 to .91.

Relationship Between Interest Inventories. Johansson (1982) reported the construct validity of the *Career Assessment Inventory* (CAI) in terms of its correlations with the *Strong-Campbell Interest Inventory* (SCII) and the *Minnesota Vocation Interest Inventory* (MVII). Like-named scales of the SCII and CAI correlated in the high .70s and low .80s, while the correlations between MVII and CAI scales were more variable and somewhat lower but still showed meaningful relationships.

Jackson (1977) gave the correlations between like-named scales of the *Jackson Vocational Interest Survey* and the SVIB as generally being in the .40s and .50s and also demonstrating considerable divergent validity through low and negative correlations between unrelated scales. Correlations in the .50s and .60s were found by Alley, Berberich, and Wilbourn (1977) between the like-named scales of the *Vocational Interest Career Examination*, the *Navy Vocational Interest Inventory*, and the *Army Classification Inventory*. Similarly, Lunneborg (1981) reported median correlations of .53 and .51 between the like-named scales of the *Vocational Interest Inventory* and the *Vocational Preference Inventory* and *Strong-Campbell Interest Inventory*, respectively.

Factor Analyses. The results of the factor-analytic studies that were discussed in more detail earlier in this report support the idea that there are a relatively small number of general interest factors, factors that are obtained in different investigations and with various inventories and item pools. This finding adds significantly to confidence that interest inventories measure several rather basic dimensions with perhaps more specific interest constructs superimposed upon them.

Process Studies. Studies of the processes involved in interest development and change contribute to construct validity through the testing of hypotheses about how and why subjects respond to interest items in the way that they do. Strong (1943) discussed the changes in interests with age-- for example, the tendency for an increase in "like" responses up to age 25, and a decrease thereafter. Roe's theory (1956) that a child's early experiences with his or her parents create or foster an adult's pattern of interests has received only ambiguous support (Crites, 1969). Cooley's (1967) monograph on the findings of Project TALENT reported that, for 9th and 12th graders, interests both follow and affect abilities as students become more aware of their own abilities and channel those abilities toward their interests or vice versa.

Other longitudinal process studies include the *Career Pattern Study* (Super, Crites, Hummel, Moser, Overstreet, & Warnath, 1957), the *Career Development Study* (Gribbons & Lohnes, 1968), the *Youth in Transition Study* (Bachman, Kahn, Mednick, Davidson, & Johnston, 1970), and the *ROTC/Army Officer Study* (Card, Goodstadt, Gross, & Shanner, 1975). Finally, Nelson's (1978) study with young children suggested that the process of vocational development is related to the processes of cognitive development.

Experimental Studies. Dawis (1980) wrote:

> . . . the construct (interest) is notable for its absence from the literature of experimental psychology. This historical development has had two consequences for psychological research which are related: a focus on the practical use value of interest measures and a failure to explicate the construct . . . . What are badly needed are studies, especially experimental studies, into the nature of interests; for example, their relationship to more basic cognitive processes, such as attention and memory, and their role in learning and social psychological processes, such as impression formation and attitude change. (p. 81)

Stulman and Dawis (1976) illustrated how the experimental procedure might be applied to vocational needs through the use of the Creativity and Independence scales of the Minnesota Importance Questionnaire. Subjects were exposed to four conditions in the assembly of a Tinker Toy (High Creativity-High Independence, High Creativity-Low Independence, Low Creativity-High Independence, Low Creativity-Low Independence), and in 12 subsequent sessions the subjects were allowed to work in any condition they preferred. Time spent in each condition was recorded and compared across groupings which had been made on the basis of MIQ scale scores and sex. Results of the analysis of variance showed significant differences in later behavior between those scoring high and those scoring low on the two MIQ scales, thus experimentally supporting the validity of the scales. Unfortunately, the experimental study is still rarely utilized in combination with individual difference measures, and the deficiency observed by Dawis (1980) continues to warrant further research.

Group Differentiation. The validity evidence for interest inventories in differentiating between occupational groups has already been discussed. Brandt and Hood (1968) added to the generality of those findings with a study comparing the predictability of the SVIB for "normal" versus "deviant" students (based on *Minnesota Multiphasic Personality Inventory* scores). The subjects were followed up several years after testing and compared for the accuracy of occupational prediction. "Normals" were found to be significantly more predictable than the "deviants," as well as having higher levels of job satisfaction.

Convergent and Divergent Validity. Since convergent and divergent validity may be relevant to a number of sources of construct validity, they have been discussed in several places throughout this report. To summarize, these types of validity have been demonstrated by such things as:

1. Members of an occupation who score high on scale A also score high on scales correlated with A but low on uncorrelated scales.

2. Members of an occupation score high on scales to which the occupation is intuitively related and low on intuitively unrelated scales.

3. Members of an occupation who score high on scale A and low on scale B for one interest inventory will do the same for like-named scales of another interest inventory.

Content-Based Judgment. The relationship between the content of an interest inventory and the relevant construct can be rationally judged when a criterion-related study is impossible or impractical. This process is distinguished from content validity in that it is an inference about score relationships with behavior rather than a process of determining an inventory's relationship with the interest domain. While this type of evidence should not stand alone in support of a hypothesized relationship, it does constitute an important additional source of construct validity information.

## Summary and Integration

The information that has been gathered in support of the validity of assessed interests comes in many forms and from a great variety of sources. Research into the content, criterion-related, and construct validities of interest inventories has supported them as measures of a distinct and important domain of human characteristics that is also related to a number of occupational criteria.

The primary deficiency of this research, however, remains a clear understanding of the relationship between interests and other human characteristics. Holland (1966, 1973), among others, has attempted to bridge this gap by considering interests as an aspect of temperament, but large-scale cross-domain research is lacking. Similarly, the developmental work mentioned earlier has looked at analogues between interests and cognitive development, and research with biographical data has demonstrated some relationships with interests (Eberhardt & Muchinsky, 1982; Mumford & Owens, 1982).

Perhaps the most promising model, however, is that specified in the *Theory of Work Adjustment* (Dawis et al., 1968). By taking into account not only a person's vocational needs, but also his or her aptitudes and the many facets of the occupational environment, the TWA seems the most promising alternative for a theory that captures the complexity of work behavior. Preliminary research with interests (Rounds, 1981) and temperament (Dawis & Lofquist, 1976, 1978) suggests that these may be incorporated into the TWA as well, and that a thorough conceptualization of the worker and work environment can be achieved. Such a complete approach seems to have optimal opportunity for uniting the validity of interests with that of other domains, resulting in what Strong in 1943 called "each finding the place where he/she will have the best chance for happiness and success" (p. 3).

## *Section 3 References*

Abrahams, N. M. (1965). *The effect of key length and item validity on overall validity, cross-validation shrinkage, and test-retest reliability of interest keys.* Unpublished doctoral dissertation, University of Minnesota.

Abrahams, N. M., Lau, A. W., & Neumann, I. (1968). *An analysis of the Navy Vocational Interest Inventory as a predictor of school performance and rating assignment* (SRR 69-11). San Diego, CA: Naval Personnel and Training Research Laboratory.

Abrahams, N. M., & Neumann, I. (1973). Predicting the unpredictable: A validation of the Strong Vocational Interest Blank for predicting military aptitude ratings of Naval Academy midshipmen. *Proceedings of the 81st Annual Convention of the American Psychological Association, 8*(2), 747-748.

Abrahams, N. M., Neumann, I., & Githens, W. H. (1968, February). *The Strong Vocational Interest Blank in predicting NROTC officer retention. Part II, Fakability* (Technical Bulletin STB 68-9). U.S. Navy Personnel Research Activity.

Abrahams, N. M., Neumann, I., & Githens, W. H. (1971). Faking vocational interests: Simulated versus real-life motivation. *Personnel Psychology, 24,* 5-12.

Abrahams, N. M., Neumann, I., & Rimland, B. (1973). *Preliminary validation of an interest inventory for selection of Navy recruiters* (Research Memorandum SRM 73-3). San Diego, CA: Navy Personnel and Training Research Laboratory.

Alley, W. E., Berberich, G. L., & Wilbourn, J. M. (1977). *Development of factor-referenced subscales for the Vocational Interest Career Examination* (AFHRL TR 76-88). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Alley, W. E., & Matthews, M. D. (1982). The Vocational Interest Career Examination: A description of the instrument and possible applications. *The Journal of Psychology, 112,* 169-193.

Alley, W. E., Wilbourn, J. M., & Berberich, G. L. (1976). *Relationships between performance on the Vocational Interest Career Examination and reported job satisfaction* (AFHRL-TR-76-89). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (Joint Committee). (1974). *Standards for educational and psychological tests* (Rev. ed.). Washington, DC: American Psychological Association.

Arvey, R. D., & Dewhirst, H. D. (1979). Relationship between diversity of interests, age, job satisfaction, and job performance. *Journal of Occupational Psychology, 52,* 17-23.

Azen, S. P., Snibbe, H. M., & Montgomery, H. R. (1973). A longitudinal predictive study of success and performance of law enforcement officers. *Journal of Applied Psychology, 57*(2), 190-192.

Bachman, J. G., Kahn, R. L., Mednick, M., Davidson, T. N., & Johnston, L. D. (1970). *Youth in transition,* Vols. 1 and 2. Ann Arbor: University of Michigan Institute of Social Research.

Baggaley, A. R. (1974). The stability of interest variables and items during adolescence. *Multivariate Experimental Clinical Research, 1*(2), 38-45.

Barak, A., & Meir, E. I. (1974). The predictive validity of a vocational interest inventory--"RAMAK:" Seven year followup. *Journal of Vocational Behavior, 4,* 377-387.

Barak, A., & Rabbi, B. (1982). Predicting persistence, stability, and achievement in college by major choice consistency: A test of Holland's consistency hypothesis. *Journal of Vocational Behavior, 20,* 235-243.

Barnette, W. L., Jr., & McCall, J. N. (1964). Validation of the Minnesota Vocational Interest Inventory for vocational high school boys. *Journal of Applied Psychology, 48,* 378-382.

Bartling, H. C., & Hood, A. B. (1980, September). *Validity of measured interest for decided and undecided students.* Paper presented at the 88th annual convention of the American Psychological Association.

Bauer, R., Mehrens, W. A., & Vinsonhaler, J. F. (1968). Predicting performance in a computer programming course. *Educational and Psychological Measurement, 28,* 1159-1164.

Berger, F. R., & Berger, R. M. (1977, September). *Vocational Interest Career Examination: Norming and standardization on a nation-wide high school sample* (AFHRL TR 77-69). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Betz, N. E., & Taylor, K. M. (1982). Concurrent validity of the Strong-Campbell Interest Inventory for graduate students in counseling. *Journal of Counseling Psychology, 29*(6), 626-635.

Bingham, R. P., & Walsh, W. B. (1978). Concurrent validity of Holland's theory for college-degreed black women. *Journal of Vocational Behavior, 13,* 242-250.

Bordin, E. S., Nachman, B., & Segal, S. J. (1963). An articulated framework for vocational development. *Journal of Counseling Psychology, 10,* 107-116.

Borgen, F. H., & Harper, G. T. (1973). Predictive validity of measured vocational interests with black and white college men. *Measurement and Evaluation in Guidance, 6,* 19-29.

Borgen, F. H., & Seling, M. J. (1978). Expressed and inventoried interests revisited: Perspicacity in the person. *Journal of Counseling Psychology, 25*(6), 536-543.

Borman, W. C., Toquam, J. L., & Rosse, R. L. (1979). *Development and validation of an inventory battery to predict Navy and Marine Corps recruiter performance* (Technical Report No. 22). Minneapolis, MN: Personnel Decisions Research Institute.

Brandt, J. E., & Hood, A. B. (1968). Effect of personality adjustment on the predictive validity of the Strong Vocational Interest Blank. *Journal of Counseling Psychology, 15*, 547-551.

Brayfield, A. H. (1953). Clerical interest and clerical aptitude. *Personnel and Guidance Journal, 31*, 304-306.

Brayfield, A. H., & Marsh, M. M. (1957). Aptitudes, interests and personality characteristics of farmers. *Journal of Applied Psychology, 41*, 98-103.

Brayfield, A. M. (1942). Review of the Kuder Preference Record. *Occupations, 21*, 267-269.

Brokaw, L. D. (1959). *School and job validation of selection measures for air traffic control training.* Technical Report WADC-TN-59-39, Lackland Air·Force Base, TX: Wright Air Development Center Personnel Laboratory.

Burgess, M. M., Duffey, M., & Temple, F. G. (1972). Two studies of prediction of success in a collegiate program of nursing. *Nursing Research, 21*(4), 357-366.

Butler, F. J., Crinnion, J., & Martin, J. 9 (1972). The Kuder Preference Record in adult vocational guidance. *Occupational Psychology, 46*, 99-104.

Cairo, P. C. (1982). Measured interests versus expressed interests as predictors of long-term occupational membership. *Journal of Vocational Behavior, 20*, 343-353.

Campbell, D. P. (1965). The vocational interests of APA presidents. *American Psychologist, 20*, 636-644.

Campbell, D. P. (1971). *Handbook for the Strong Vocational Interest Blank.* Stanford, CA: Stanford University Press.

Campbell, D. P., & Hansen, J. C. (1981). *Manual for the SVIB-SCII.* Stanford, CA: Stanford University Press.

Campbell, D. P., & Holland, J. L. (1972). A merger in vocational interest research: Applying Holland's theory to Strong's data. *Journal of Vocational Behavior, 2*, 353-376.

Card, J. J., Goodstadt, B. C., Gross, D. E., & Shanner, T. I. M. (1975). *Development of an ROTC/Army career commitment model.* Palo Alto, CA: American Institutes for Research.

Carter, H. D., Pyles, M. K., & Bretnall, E. P. (1935). A comparative study of factors in vocational interest scores of high school boys. *Journal of Educational Psychology, 26,* 81-98.

Cattell, R. B. (1944). Psychological measurement: Ipsative, normative and interactive. *Psychological Review, 51,* 292-303.

Clark, K. E. (1961). *Vocational interests of non-professional men.* Minneapolis, MN: University of Minnesota Press.

Clark, K. E., & Campbell, D. P. (1965). *Manual for the Minnesota Vocational Interest Inventory.* New York, NY: Psychological Corporation.

Claudy, J. G., Caylor, J. S., & Kass, R. H. (1981, December). *Development of the Army Research Institute Interest Survey* (unnumbered technical report).

Clemans, W. V. (1966). An analytic and empirical examination of some properties of ipsative measures. *Psychometric Monographs, 14,* 1-56.

Clemans, W. V. (1968). Interest measurement and the concept of ipsative measures. *Measurement and Evaluation in Guidance, 1*(1), 50-55.

Cole, M. S., & Hanson, G. R. (1971). *An analysis of the structure of vocational interests* (ACT Research Report No. 40). Iowa City, IA: The American College Testing Program.

Cooley, W. W. (1967). Interactions among interests, abilities, and career plans. *Journal of Applied Psychology, 51*(5, Whole No. 640).

Crites, J. O. (1969). *Vocational Psychology.* New York: McGraw-Hill.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10,* 3-31.

Cronbach, L. J., & Gleser, G. C. (1965). *Psychological tests and personnel decisions* (2nd ed.). Urbana, IL: University of Illinois Press.

Dann, J. E., & Abrahams, N. M. (1973, October). *Occupational scales of the Navy Vocational Interest Inventory: I. Development* (Technical Report No. 74-4). San Diego, CA: Navy Personnel Research and Development Center.

Dann, J. E.. & Abrahams, N. M. (1977). *Occupational scales of the Navy Vocational Interest Inventory: III. Relationship to job satisfaction, "A"school grades, and job performance* (Technical Report No. 78-3). San Diego, CA: Navy Personnel Research and Development Center.

Darley, J. G., & Hagenah, T. (1955). *Vocational interest measurement.* Minneapolis, MN: University of Minnesota Press.

Dawis, R. V. (1980). Measuring interests. *New Directions for Testing and Measurement, 7,* 77-93.

Dawis, R. V., & Lofquist, L. H. (1976). Personality style and the process of work adjustment. *Journal of Counseling Psychology, 23,* 55-59.

Dawis, R. V., & Lofquist, L. H. (1978). A note on the dynamics of work adjustment. *Journal of Vocational Behavior, 12,* 76-79.

Dawis, R. V., Lofquist, L. H., & Weiss, D. J. (1968). A theory of work adjustment (rev.). *Minnesota Studies in Vocational Rehabilitation,* Bulletin 23.

Diamond, E. E. (Ed.). (Spring, 1975). *Issues of sex bias and sex fairness in career interest measurement.* Washington, DC: Department of Health, Education, and Welfare, National Institute of Education, Career Education Program.

DiMichael, S. G., & Dabelstein, D. H. (1947). Work satisfaction and work efficiency of vocational rehabilitation counselors as related to measured interests (Abstract). *American Psychologist, 2,* 342-343.

Doll, R. E. (1971). Item susceptibility to attempted faking as related to item characteristic and adopted fake set. *The Journal of Psychology, 77,* 9-16.

Doll, R. E., Ambler, R. K., Lane, N. E., & Bale, R. M. (1972). Vocational interest differences between students completing the Naval Aviation Training Program and students voluntarily withdrawing (Abstract). *Proceedings of the 80th Annual Convention of the American Psychological Association, 7*(2), 621-622.

Dolliver, R. H. (1969a). "3.5 to 1" on the Strong Vocational Interest Blank as a pseudo-event. *Journal of Counseling Psychology, 16,* 172-174.

Dolliver, R. H. (1969b). Strong Vocational Interest Blank versus expressed vocational interests: A review. *Psychological Bulletin, 72,* 95-107.

Dolliver, R. H. (1975). Concurrent prediction from the Strong Vocational Interest Blank. *Journal of Counseling Psychology, 22*(3), 199-203.

Dolliver, R. H. (1981). A review of female-male score differences on the Strong-Campbell twin occupational scales. *Journal of Counseling Psychology, 28*(4), 334-341.

Dolliver, R. H., & Clark, J. A. (1972). Status faking on the SVIB-M. *Journal of Vocational Behavior, 2,* 47-56.

Dolliver, R. H., Irvin, J. A., & Bigley, S. S. (1972). Twelve year followup of the Strong Vocational Interest Blank. *Journal of Counseling Psychology, 19*(3), 212-217.

Dolliver, R. H., & Will, J. A. (1977). Ten year followup of the Tyler Vocational Card Sort and the Strong Vocational Interest Blank. *Journal of Counseling Psychology, 24*, 48-54.

Dore', R. L. (1970). Self-concept and interests related to job satisfaction of managers. Doctoral thesis, University of Washington. *Dissertation Abstracts International, 31*, 2338B.

Dore', R., & Meachum, M. (1973). Self-concept and interests related to job satisfaction of managers. *Personnel Psychology, 26*(1), 49-59.

Droege, R. C., & Padgett, A. (1979). Development of an interest oriented occupational classification system. *Vocational Guidance Quarterly, 27*(4), 302-310.

Dyer, D. T. (1939). The relation between vocational interests of men in college and their subsequent histories for ten years. *Journal of Applied Psychology, 23*, 280-288.

Eberhardt, B. J., & Muchinsky, P. M. (1982). Biodata determinants of vocational typology: An integration of two paradigms. *Journal of Applied Psychology 67*, 714-727.

Echternacht, G. J., Reilly, R. R., & McCaffrey, P. J. (1973, December). *Development and validity of a vocational and occupational interest inventory* (AFHRL-TR-73-38). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Edwards, K. J., & Whitney, D. R. (1972). Structural analysis of Holland's personality types using factor and configural analysis. *Journal of Counseling Psychology, 19*(2), 136-145.

Enright, J. B., & Pinneau, S. R. (1955). Predictive value of subjective choice of occupation and of the Strong Vocational Interest Blank over fifteen years (abstract). *American Psychologist, 10*, 424-425.

Ferguson, L. W. (1958). Life insurance interest, ability and termination of employment. *Personnel Journal, 11*, 189-193.

Fisher, S. T., Weiss, D. J., & Dawis, R. V. (1968). A comparison of Likert and pair-comparisons techniques in multivariate attitude scaling. *Educational and Psychological Measurement, 28*, 81-94.

Gade, E. M., & Soliah, D. (1975). Vocational Preference Inventory high point codes versus expressed choices as predictors of college major and career entry. *Journal of Counseling Psychology, 22*, 117-121.

Gadel, M., & Kriedt, P. H. (1952). Relationships of aptitude, interest, performance, and job satisfaction of IBM operators. *Personnel Psychology, 5*, 207-212.

Gati, I. (1979). A hierarchical model for the structure of vocational interests. *Journal of Vocational Behavior, 15*, 90-106.

Gati, I. (1982). Testing models for the structure of vocational interests. *Journal of Vocational Behavior, 21*, 164-182.

Gati, I., & Meir, E. I. (1982). Congruence and consistency derived from the circular and the hierarchical models as predictors of occupational choice satisfaction. *Journal of Vocational Behavior, 20*, 354-365.

Gordon, M. E., & Gross, R. H. (1978). A critique of methods for operationalizing the concept of fakability. *Educational and Psychological Measurement, 38*, 771-782.

Gottfredson, G. D.., & Holland, J. L. (1975). Vocational choices of men and women: A comparison of predictors from the Self-Directed Search. *Journal of Counseling Psychology, 22*(1), 28-34.

Gottfredson, G. D., Holland, J. L., & Gottfredson, L. S. (1975). The relation of vocational aspirations and assessments to employment reality. *Journal of Vocational Behavior, 7*, 135-148.

Gottfredson, G. D., Holland, J. L., & Holland, J. E. (1978). The seventh revision of the Vocational Preference Inventory. *JSAS Catalog of Selected Documents in Psychology*, 8:98, No. 1783.

Gottfredson, L. S. (1980). Construct validity of Holland's occupational typology in terms of prestige, census, Department of Labor, and other classification systems. *Journal of Applied Psychology, 65*(6), 697-714.

Gray, C. W. (1959). *Detection of faking in vocational interest measurement.* Unpublished doctoral dissertation, University of Minnesota.

Gribbons, W. D., & Lohnes, P. R. (1968). *Emerging careers.* New York, NY: Teacher's College Press.

Guilford, J. P., Christenson, P. R., Bond, N. A., Jr., & Sutton, M. A. (1954). A factor analysis study of human interests. *Psychological Monographs, 68*(4, Whole No. 375).

Guinn, N., Vitola, B. M., & Leisey, S. A. (1976). *Background and interest measures as predictors of success in undergraduate pilot training* (AFHRL TR-76-9). Brooks AFB, TX: Air Force Human Resources Laboratory.

Guinn, N., Wilbourn, J. M., & Kantor, J. E. (1977). *Preliminary development and validation of a screening technique for entry into the Security Police Career Field* (AFHRL-TR-77-38). Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Guion, R. M. (1976). Recruiting, selection, and placement. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology.* Chicago: Rand McNally.

Guthrie, W. R., & Herman, A. (1982). Vocational maturity and its relationship to Holland's theory of vocational choice. *Journal of Vocational Behavior, 21*, 196-205.

Hahn, M. E., & Williams, C. T. (1945). The measured interests of Marine Corps women reservists. *Journal of Applied Psychology, 29*, 198-211.

Hall, D. T. (1976). *Careers in organizations.* Santa Monica, CA: Goodyear.

Hansen, J. C. (1976). Exploring new directions for Strong-Campbell Inventory occupational scale construction. *Journal of Vocational Behavior, 9*, 147-160.

Hansen, J. C. (1983). *Measurement of interests.* Unpublished manuscript, Minneapolis, MN: Personnel Decisions Research Institute.

Hansen, J. C., & Johansson, C. B. (1972). The application of Holland's vocational model to the Strong Vocational Interest Blank for women. *Journal of Vocational Behavior, 2*, 479-493.

Hansen, J. C., & Stocco, J. L. (1980). Stability of vocational interests of adolescents and young adults. *Measurement and Evaluation in Guidance, 13*(3), 171-176.

Hansen, J. C., & Swanson, J. L. (1983). Stability of interests and the predictive and concurrent validity of the 1981 Strong-Campbell Interest Inventory for college majors. *Journal of Counseling Psychology, 30*(2), 194-201.

Hanson, G. R., Prediger, D. J., & Schussel, R. H. (1977). *Development and validation of sex-balanced interest inventory scales* (ACT Research Report No. 78). Iowa City, IA: American College Testing Program.

Harrington, D. F., & O'Shea, A. J. (1980). Applicability of the Holland (1973) model of vocational development with Spanish-speaking clients. *Journal of Counseling Psychology, 27*(3), 246-251.

Harvey, D. W., & Whinfield, R. W. (1973). Extending Holland's theory to adult women. *Journal of Vocational Behavior, 3*, 115-129.

Hener, T., & Meir, E. I. (1981). Congruency, consistency, and differentiation as predictors of job satisfaction within the nursing occupation. *Journal of Vocational Behavior, 18*, 304-309.

Herzberg, F., & Russell, D. (1953). The effects of experience and change of job interest on the Kuder Preference Record. *Journal of Applied Psychology, 37*, 478-481.

Holland, J. L. (1966). *The psychology of vocational choice.* Waltham, MA: Blaisdell.

Holland, J. L. (1973). *Making vocational choices: A theory of careers.* Englewood Cliffs, NJ: Prentice-Hall.

Holland, J. L. (1976). Vocational preferences. In M. D. Dunnette (Ed.), *Handbook of Industrial and Organizational Psychology.* Chicago: Rand McNally.

Holland, J. L. (1979). *Manual for the Self-Directed Search.* Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L., & Lutz, S. W. (1968). The predictive value of a student's choice of vocation. *Personnel and Guidance Journal, 46*, 428-434.

Holland, J. L., Magoon, T. M., & Spokane, A. R. (1981). Counseling psychology: Career interventions, research and theory. *Annual Review of Psychology, 32*, 279-305.

Horton, J., & Walsh, W. B. (1976). Concurrent validity of Holland's theory for college degreed working women. *Journal of Vocational Behavior, 9*, 201-208.

Jackson, D. N. (1977). *Jackson Vocational Interest Survey Manual.* London, Ontario, Canada: Research Psychologists Press.

Johansson, C. B. (1982). *Manual for the Career Assessment Inventory.* Minneapolis, MN: Interpretive Scoring Systems.

Johnson, D. M. (1973). Relationships between selected cognitive and noncognitive variables and practical nursing achievement. *Nursing Research, 22*(2), 148-153.

Johnson, J. C., & Dunnette, M. D. (1968). Validity and test-retest stability of the Nash managerial effectiveness scale on the revised form of the Strong Vocational Interest Blank. *Personnel Psychology, 21*(3), 283-294.

Johnson, R. W. (1977). Relationships between female and male interest scales for the same occupations. *Journal of Vocational Behavior, 11*, 239-252.

Johnson, R. W., & Johansson, C. B. (1972). Moderating effects of basic interests on predictive validity of SVIB occupational scales. *Proceedings of the 80th Annual Convention, American Psychological Association*, 589-590.

Johnson, R. W., Kirk, K. W., & Ohvall, R. A. (1975). Predictive validity of SVIB pharmacist scales. *Educational and Psychological Measurement, 35*, 951-955.

Jones, K. J. (1965). Occupational preference and social orientation. *Personnel and Guidance Journal, 43*, 574-579.

Kates, S. L. (1950). Rorschach responses related to vocational interests and job satisfaction. *Psychological Monographs, 64*, 1-34.

Kirchner, W. K. (1961). "Real-life" faking on the Strong Vocational Interest Blank by sales applicants. *Journal of Applied Psychology, 45*, 273-276.

Knapp, T. R. (1966). Interactive versus ipsative measurement of career interests. *Personnel and Guidance Journal, 44*, 482-486.

Knauft, E. B. (1951). Vocational interests and managerial success. *Journal of Applied Psychology, 35,* 160-163.

Kroger, R. O. (1974). Faking in interest measurement: A social-psychological perspective. *Measurement and Evaluation in Guidance, 7*(2), 130-134.

Kuder, G. F. (1977). *Activity interests and occupational choice.* Chicago: Science Research Associates.

Kuder, G. F., & Diamond, E. E. (1979). *Kuder DD Occupational Interest Survey General Manual.* Chicago: Science Research Associates.

Kunce, J. T., Decker, G. L., & Eckleman, C. C. (1976). Strong Vocational Interest Blank basic interest clusters and occupational satisfaction. *Journal of Vocational Behavior, 9,* 355-362.

Lamb, R. F. (1976). *Validity of the American college Testing Interest Inventory for minority group members* (Research Report 72). Iowa City, IA: American College Testing Program.

Lamb, R. R., & Prediger, D. J. (1979). Criterion-related validity of sex-restrictive and unisex interest scales: A comparison. *Journal of Vocational Behavior, 15,* 231-246.

Lammlein, S. (1982). *A proposal for the administration of the fundamentals of an engineering examination program.* Minneapolis, MN: Personnel Decisions Research Institute.

Lau, A. W., & Abrahams, N. M. (1969). *Area scales of the Navy Vocational Interest Inventory as predictors of school performance and rating assignment* (Report SRR 70-1). San Diego, CA: Navy Personnel Research and Development Center.

Lau, A. W., & Abrahams, N. M. (1970). *The Navy Vocational Interest Inventory as a predictor of job performance* (Report SSR 70-28). San Diego, CA: Navy Personnel and Training Research Laboratory.

Lau, A. W., & Abrahams, N. M. (1971). *Reliability and predictive validity of the Navy Vocational Interest Inventory* (Research Report SRR 71-16). San Diego, CA: Navy Personnel and Training Research Laboratory.

Lau, A. W., Lacey, L. A., & Abrahams, N. M. (1969). *An analysis of the Navy Vocational Interest Inventory as a predictor of career motivation* (Technical Report 69-27). San Diego, CA: Navy Personnel and Training Research Laboratory.

Lee, D. L., & Hedahl, B. (1973). Holland's personality types applied to the SVIB basic interest scales. *Journal of Vocational Behavior, 3,* 61-68.

Lofquist, L. H., & Dawis, R. V. (1969). *Adjustment to work.* New York: Appleton-Century-Crofts.

Lunneborg, P. G. (1979). The Vocational Interest Inventory: Development and validation. *Educational and Psychological Measurement, 39*(2), 445-451.

Lunneborg, P. W. (1978). What it means to be general cultural. *Journal of Vocational Behavior, 13*(2), 222-228.

Lunneborg, P. W. (1981). *Vocational Interest Inventory Manual.* Los Angeles: Western Psychological Services.

Mandell, M. (1957). The selection of executives. In Dooher, M. J., & Marting, E. (Eds.), *The selection of management personnel, I and II.* New York: American Management Association.

Martin, G. R. (1968). Job satisfaction in practical nursing as a function of measured and expressed interest. Doctoral thesis, University of Illinois, Urbana, IL. *Dissertation Abstracts International, 30,* 266B.

Matthews, D. F., & Walsh, W. B. (1978). Concurrent validity of Holland's theory for noncollege-degreed working women. *Journal of Vocational Behavior, 12,* 371-379.

Mayeske, G. W. (1964). The validity of Kuder Preference Record scores in predicting forester turnover and advancement. *Personnel Psychology, 17,* 207-210.

McArthur, C. (1954). Long-term validity of the Strong Interest Test in two subcultures. *Journal of Applied Psychology, 38,* 346-354.

McArthur, C., & Stevens, L. B. (1955). The validation of expressed interests as compared with inventoried interests: A fourteen year followup. *Journal of Applied Psychology, 39,* 184-189.

McRae, G. G. (1959). *The relationship of job satisfaction and earlier measured interests.* Doctoral dissertation, University of Florida.

Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. *Journal of Applied Psychology, 30,* 525-564.

Meir, E. I. (1973). The structure of occupations by interests: A smallest space analysis. *Journal of Vocational Behavior, 3,* 21-31.

Meir, E. I., & Ben-Yahuda, Y. A. (1976). Inventories based on Roe and Holland yield similar results. *Journal of Vocational Behavior, 8,* 269-274.

Meir, E. I., & Erez, M. (1981). Fostering a career in engineering. *Journal of Vocational Behavior, 18,* 115-120.

Miner, J. B. (1960). The Kuder Preference Record in management appraisal. *Personnel Psychology, 13,* 187-196.

Mitchell, S. K., Lunneborg, P. W., & Lunneborg, C. E. (1971). *A vocational interest inventory based on Roe's interest areas.* Unpublished manuscript, University of Washington, Seattle.

Mount, M. K., & Muchinsky, P. M. (1978). Concurrent validation of Holland's hexagonal model with occupational workers. *Journal of Vocational Behavior, 13,* 348-354.

Mumford, M. D., & Owens, W. A. (1982). Life history and vocational interests. *Journal of Vocational Behavior, 21,* 330-348.

Nafziger, D. H., & Helms, S. T. (1974). Cluster analyses of interest inventory scales as tests of Holland's occupational classification. *Journal of Applied Psychology, 59*(3), 344-353.

Nash, A. N. (1966). Development of an SVIB key for selecting managers. *Journal of Applied Psychology, 50,* 250-254.

Nelson, J. A. N. (1978). Age and sex differences in the development of children's occupational reasoning. *Journal of Vocational Behavior, 13,* 287-297.

Norman, W. T. (1963). Personality measurement, faking and detection: An assessment method for use in personnel selection. *Journal of Applied Psychology, 48,* 225-241.

North, R. D., Jr. (1958). Tests for the accounting profession. *Educational and Psychological Measurement, 18,* 691-713.

O'Neil, J. M., & Magoon, T. M. (1977). The predictive power of Holland's Investigative personality type and consistency levels using the Self-Directed Search. *Journal of Vocational Behavior, 10,* 39-46.

Peiser, C., & Meir, E. I. (1978). Congruency, consistency, and differentiation of vocational interests as predictors of vocational satisfaction and preference stability. *Journal of Vocational Behavior, 12,* 270-278.

Perry, D. K. (1955). Forced choice versus L-I-D items in vocational interest measurement. *Journal of Applied Psychology, 39,* 256-262.

Perry, D. K. (1967). Vocational interests and success of computer programmers. *Personnel Psychology, 20*(4), 517-524.

Pickrel, E. W. (1954). *The relative predictive efficiency of three methods of utilizing scores from biographical inventories* (AFHTRC-TR-54-73). Lackland AFB, TX: Air Force Personnel and Training Research Center.

Porter, A. (1962). Effect of organization size on the validity of Masculinity-Femininity score. *Journal of Applied Psychology, 46,* 228-229.

Prediger, D. H., & Hanson, G. R. (1976). Holland's theory of careers applied to women and men: Analysis of implicit assumptions. *Journal of Vocational Behavior, 8,* 167-184.

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior, 21*, 259-287.

Rayman, J. R. (1976). Sex and the single interest inventory: The empirical validation of sex-balanced interest inventory items. *Journal of Counseling Psychology, 23*(3), 239-246.

Reeves, D. J., & Booth, R. F. (1979). Expressed versus inventoried interests as predictors of paramedical effectiveness. *Journal of Vocational Behavior, 15*, 155-163.

Reilly, R., & Echternacht, G. (1979). Some problems with the criterion-keying approach to occupational interest scale development. *Educational and Psychological Measurement, 39*, 85-94.

Riley, P. J. (1981). The influence of gender on occupational aspirations of kindergarten children. *Journal of Vocational Behavior, 19*, 244-250.

Roe, A. (1956). *The psychology of occupations.* New York: Wiley.

Roe, A., & Siegelman, M. (1964). *The origin of interests.* Washington, DC: American Personnel and Guidance Association.

Rosenberg, N., & Izard, C. E. (1954). Vocational interests of naval aviation cadets. *Journal of Applied Psychology, 38*(5), 354-358.

Rounds, J. B., Jr. (1981). *The comparative and combined utility of need and interest data in the prediction of job satisfaction.* Unpublished doctoral dissertation, University of Minnesota.

Rounds, J. B., Davison, M. L., & Dawis, R. V. (1979). The fit between Strong-Campbell Interest Inventory general occupational themes and Holland's hexagonal model. *Journal of Vocational Behavior, 15*, 303-315.

Rounds, J. B., Jr., & Dawis, R. V. (1979). Factor analysis of Strong Vocational Interest Blank items. *Journal of Applied Psychology, 64*(2), 132-143.

Scarpello, V., & Campbell, J. P. (1983). Job satisfaction and the fit between individual needs and organizational rewards. *Journal of Occupational Psychology, 56*, 315-328.

Schaeffer, R. H. (1953). Job satisfaction as related to need satisfaction in work. *Psychological Monographs, 67*(14, Whole No. 364).

Schletzer, V. M. (1966). SVIB as a predictor of job satisfaction. *Journal of Applied Psychology, 50*, 5-8.

Schmidt, F. L. (1974). Probability and utility assumptions underlying use of the Strong Vocational Interest Blank. *Journal of Applied Psychology, 59*(4), 456-464.

Schultz, I. T., & Barnabas, B. (1945). Testing for leadership in industry. *Transactions of the Kansas Academy of Science*, 160-164.

Slaney, R. B. (1978). Expressed and inventoried vocational interests: A comparison of instruments. *Journal of Counseling Psychology, 25*(6), 520-529.

Strong, E. K. (1943). *Vocational interests of men and women.* Stanford, CA: Stanford University Press.

Strong, E. K., Jr. (1935). Predictive value of the vocational interest test. *Journal of Educational Psychology, 26*, 331-349.

Strong, E. K., Jr. (1953). Validity of occupational choice. *Educational and Psychological Measurement, 13*, 110-121.

Strong, E. K., Jr. (1955). *Vocational interests 18 years after college.* Minneapolis, MN: University of Minnesota Press.

Stulman, D. A., & Dawis, R. V. (1976). Experimental validation of two MIQ scales. *Journal of Vocational Behavior, 9*, 161-167.

Super, D. E., Crites, J. O., Hummel, R. C., Moser, H. P., Overstreet, P. L., & Warnath, C. F. (1957). *Vocational development: A framework for research.* New York: Teacher's College Press.

Thorndike, R. M., Weiss, D. J., & Dawis, R. V. (1968). Canonical correlation of vocational interests and vocational needs. *Journal of Counseling Psychology, 15*(2), 101-106.

Thurstone, L. L. (1931). A multiple factor study of vocational interests. *Personnel Journal, 10*, 198-205.

Tilton, J. W. (1937). The measurement of overlapping. *Journal of Educational Psychology, 28*, 656-662.

Tittle, C. K., & Zytowski, D. G. (Eds.) (1978). *Sex fair interest measurement: Research and implications.* Washington, DC: National Institute of Education.

Trimble, J. T. (1965). *Ten-year longitudinal followup study of inventoried interests of selected high school students.* Unpublished doctoral dissertation, University of Missouri.

Tuckman, B. W. (1974). An age grade model for career development education. *Journal of Vocational Behavior, 4*, 193-212.

U.S. Department of Labor. (1977). *Dictionary of Occupational Titles.* Washington, DC: U.S. Government Printing Office.

Villwock, J. D., Schnitzen, J. P., & Carbonari, J. P. (1976). Holland's personality constructs as predictors of stability of choice. *Journal of Vocational Behavior, 9*, 77-85.

Wakefield, J. A., Jr., Yom, B. L., Doughtie, E. B., Chang, W. C., & Alston, H. L. (1975). The geometric relationship between Holland's personality typology and the Vocational Preference Inventory for blacks. *Journal of Counseling Psychology, 22,* 58-60.

Wall, H. W., Osipow, S. H., & Ashby, J. D. (1967). SVIB scores, occupational choices, and Holland personality types. *Vocational Guidance Quarterly, 15,* 201-205.

Walsh, W. B. (1979). Vocational behavior and career development; 1978: A review. *Journal of Vocational Behavior, 15,* 119-154.

Walsh, W. B., Howard, P. R., O'Brien, W. F., Santa-Maria, M. L., & Edmondson, C. J. (1973). Consistent occupational preferences and satisfaction, self-concept, self-acceptance, and vocational maturity. *Journal of Vocational Behavior, 3,* 453-463.

Werner, J. E. (1969). A study of Holland's theory of vocational choice as it applies to selected working women. *Dissertation Abstracts International, 30,* 1832A.

Whetstone, R. D., & Hayles, V. R. (1975). The SVIB and black college men. *Measurement and Evaluation in Guidance, 8,* 105-109.

Whitney, D. R. (1969). Predicting from expressed vocational choice: A review. *Personnel and Guidance Journal, 48,* 279-286.

Wiener, Y., & Klein, K. L. (1978). The relationship between vocational interests and job satisfaction. *Journal of Vocational Behavior, 13,* 298-304.

Wiggins, J. D. (1976). The relation of job satisfaction to vocational preferences among teachers of the educable mentally retarded. *Journal of Vocational Behavior, 8,* 13-19.

Wightwick, I. (1945). Vocational interest patterns. *Teacher's College Contributions to Education,* No. 900.

Wiley, M. O., & Magoon, T. M. (1982). Holland high point social types: Is consistency related to persistence and achievement? *Journal of Vocational Behavior, 20,* 14-21.

Williams, F., & Harrell, T. W. (1964). Predicting success in business. *Journal of Applied Psychology, 68,* 164-167.

Wiskoff, M. F. (1980). *Selection of Marine Corps drill instructors* (Technical Report 80-17). San Diego, CA: Navy Personnel Research and Development Center.

Worthington, E. L., & Dolliver, R. H. (1977). Validity studies of the Strong Vocational Interest Inventories. *Journal of Counseling Psychology, 24*(3), 208-216.

Zalinsky, J. S., & Abrahams, N. M. (1979). The effects of item context in faking personnel selection inventories. *Personnel Psychology, 32*(1), 161-166.

Zytowski, D. G. (1974). Predictive validity of the Kuder Preference Record, Form B, over a 25-year span. *Measurement and Evaluation in Guidance, 7*(2), 122-129.

Zytowski, D. G. (1976). Predictive validity of the Kuder Occupational Interest Survey: A 12 to 19 year followup. *Journal of Counseling Psychology, 3*, 221-233.

# APPENDIX A

## Existing Temperament Scales Classified According to Construct

## TEMPERAMENT SCALES INCLUDED IN THE TAXONOMY

### California Psychological Inventory (CPI)

Dominance (Do)
Capacity for Status (Cs)
Sociability (Sy)
Social Presence (Sp)
Self-acceptance (Sa)
Sense of Well-being (Wb)
Responsibility (Re)
Socialization (So)
Self-control (Sc)
Tolerance (To)
Achievement via Conformance (Ac)
Achievement via Independence (Ai)
Intellectual Efficiency (Ie)
Psychological-mindedness (Py)
Flexibility (Fx)
Femininity (Fe)

### Comrey Personality Scales (CPS)

Trust vs. Defensiveness (T)
Orderliness vs. Lack of Compulsion (O)
Social Conformity vs. Rebelliousness (C)
Activity vs. Lack of Energy (A)
Emotional Stability vs. Neuroticism (S)
Extraversion vs. Introversion (E)
Masculinity vs. Femininity (M)
Empathy vs. Egocentrism (P)

### Differential Personality Questionnaire (DPQ)

Wellbeing (Wb)
Social Potency (SP)
Achievement (Ach)
Social Closeness (SC)
Stress Reaction (SR)
Alienation (Al)
Aggression (Agg)
Control (Con)
Harmavoidance (Ha)
Traditionalism (Tr)
Absorption (Ab)

## Edwards Personal Preference Schedule (EPPS)

Achievement (ach)
Deference (def)
Order (ord)
Exhibition (exh)
Autonomy (aut)
Affiliation (aff)
Intraception (int)
Succorance (suc)
Dominance (dom)
Abasement (aba)
Nurturance (nur)
Change (chg)
Endurance (end)
Heterosexuality (het)
Aggression (agg)

## Eysenck Personality Questionnaire (EPQ)

Neuroticism (N)
Extraversion (E)
Psychoticism (P)

## Gordon Personal Profile-Inventory (GPPI)

Ascendancy (A)
Responsibility (R)
Emotional Stability (E)
Sociability (S)
Cautiousness (C)
Original Thinking (O)
Personal Relations (P)
Vigor (V)

## Guilford-Zimmerman Temperament Survey (GZTS)

General Activity (G)
Restraint (R)
Ascendance (A)
Sociability (S)
Emotional Stability (E)
Objectivity (O)
Friendliness (F)
Thoughtfulness (T)
Personal Relations (P)
Masculinity (M)

## Jackson Personality Inventory (JPI)

Anxiety (Anx)
Breadth of Interest (Bdi)
Complexity (Cpx)
Conformity (Cny)
Energy Level (Enl)
Innovation (Inv)
Interpersonal Affect (Iaf)
Organization (Org)
Responsibility (Rsy)
Risk Taking (Rkt)
Self Esteem (Ses)
Social Adroitness (Sca)
Social Participation (Spt)
Tolerance (Tol)
Value Orthodoxy (Vlo)

## Minnesota Multiphasic Personality Inventory (MMPI)

Subtle defensiveness (K)
Hypochondriasis (Hs)
Depression (D)
Hysteria (Hy)
Psychopathic Deviate (Pd)
Masculinity-femininity (Mf)
Paranoia (Pa)
Psychasthenia (Pt)
Schizophrenia (Sc)
Mania (Ma)
Social Introversion (Si)

## Omnibus Personality Inventory (OPI)

Thinking Introversion (TI)
Theoretical Orientation (TO)
Estheticism (Es)
Complexity (Co)
Autonomy (Au)
Religious Orientation (RO)
Social Extroversion (SE)
Impulse Expression (IE)
Personal Integration (PI)
Anxiety Level (AL)
Altruism (Am)
Practical Outlook (PO)
Masculinity-Femininity (MF)

## Personality Research Form (PRF)

Abasement (Ab)
Achievement (Ac)
Affiliation (Af)
Aggression (Ag)
Autonomy (Au)
Change (Ch)
Cognitive Structure (Cs)
Defendence (De)
Dominance (Do)
Endurance (En)
Exhibition (Ex)
Harmavoidance (Ha)
Impulsivity (Im)
Nurturance (Nu)
Order (Or)
Play (Pl)
Sentience (Se)
Social Recognition (Sr)
Succorance (Su)
Understanding (Un)

## Sixteen Personality Factor Questionnaire (16PF)

A: Reserved vs. Outgoing (A)
B: Less Intelligent vs. More Intelligent (B)
C: Affected by Feelings vs. Emotionally Stable (C)
E: Humble vs. Assertive (E)
F: Sober vs. Happy-go-lucky (F)
G: Expedient vs. Conscientious (G)
H: Shy vs. Venturesome (H)
I: Tough-minded vs. Tender-minded (I)
L: Trusting vs. Suspicious (L)
M: Practical vs. Imaginative (M)
N: Forthright vs. Shrewd (N)
O: Placid vs. Apprehensive (O)
$Q_1$: Conservative vs. Experimenting ($Q_1$)
$Q_2$: Group-dependent vs. Self-sufficient ($Q_2$)
$Q_3$: Undisciplined Self-conflict vs. Controlled ($Q_3$)
$Q_4$: Relaxed vs. Tense ($Q_4$)

# POTENCY SCALES

1. CPS Activity (CPS A)
2. DPQ Wellbeing (DPQ Wb)
3. GPPI Vigor (GPPI V)
4. GZTS General Activity (GZTS G)
5. JPI Energy Level (JPI Enl)
6. DPQ Social Potency (DPQ SP)
7. EPPS Dominance (EPPS dom)
8. GPPI Ascendancy (GPPI A)
9. GZTS Ascendance (GTZS A)
10. PRF Dominance (PRF Do)
11. 16PF E: Assertive (16PF E)
12. PRF Exhibition (PRF Ex)
13. CPS Extraversion (CPS E)
14. EPQ Extraversion (EPQ E)
15. GPPI Sociability (GPPI S)
16. GZTS Sociability (GZTS S)
17. JPI Self Esteem (JPI Ses)
18. MMPI Social Introversion (MMPI Si) - reversed
19. OPI Social Extroversion (OPI SE)
20. 16PF F: Happy-go-lucky (16PF F)
21. 16PF H: Venturesome (16PF H)
22. CPI Dominance (CPI Do)
23. CPI Capacity for Status (CPI Cs)
24. CPI Sociability (CPI Sy)
25. CPI Social Presence (CPI Sp)
26. CPI Self-acceptance (CPI Sa)

## ADJUSTMENT SCALES

1. DPQ Alienation (DPQ Al) - reversed
2. GZTS Objectivity (GTZS O)
3. CPS Emotional Stability (CPS S)
4. DPQ Stress Reaction (DPQ SR) - reversed
5. EPQ Neuroticism (EPQ N) - reversed
6. GPPI Emotional Stability (GPPI E)
7. GZTS Emotional Stability (GZTS E)
8. JPI Anxiety (JPI Anx) - reversed
9. OPI Personal Integration (OPI PI)
10. OPI (low) Anxiety Level (OPI AL)
11. 16PF C: Emotionally Stable (16PF C)
12. 16PF O: Apprehensive (16PF O) - reversed
13. 16PF $Q_4$: Tense (16PF $Q_4$) - reversed
14. CPI Sense of Well-being (CPI Wb)
15. CPI Responsibility (CPI Re)
16. CPI Self-control (CPI Sc)
17. CPI Tolerance (CPI To)
18. CPI Achievement via Conformance (CPI Ac)
19. CPI Intellectual Efficiency (CPI Ie)
20. MMPI Subtle defensiveness (MMPI K)
21. MMPI Depression (MMPI D) - reversed
22. MMPI Psychasthenia (MMPI Pt) - reversed
23. MMPI Schizophrenia (MMPI Sc) - reversed

## AGREEABLENESS SCALES

1. CPS Trust (CPS T)
2. GZTS Personal Relations (GZTS P)
3. DPQ Aggression (DPQ Agg) - reversed
4. EPPS Aggression (EPPS agg) - reversed
5. EPQ Psychoticism (EPQ P) - reversed
6. GZTS Friendliness (GZTS F)
7. PRF Aggression (PRF Ag) - reversed
8. 16PF L: Suspicious (16PF L) - reversed
9. GPPI Personal Relations (GPPI P)
10. JPI Tolerance (JPI Tol)
11. PRF Abasement (PRF Ab)
12. PRF Defendence (PRF De) - reversed
13. CPS Empathy (CPS P)
14. EPPS Nurturance (EPPS Nur)
15. JPI Interpersonal Affect (JPI Iaf)
16. PRF Nurturance (PRF Nu)

## DEPENDABILITY SCALES

1. CPI Socialization (CPI So)
2. CPS Social Conformity (CPS C)
3. MMPI Psychopathic Deviate (MMPI Pd) - reversed
4. OPI Impulse Expression (OPI IE) - reversed
5. CPS Orderliness (CPS O)
6. EPPS Order (EPPS ord)
7. JPI Organization (JPI Org)
8. PRF Order (PRF Or)
9. CPI Flexibility (CPI Fx) - reversed
10. DPQ Control (DPQ Con)
11. GPPI Responsibility (GPPI R)
12. GZTS Restraint (GZTS R)
13. PRF Cognitive Structure (PRF Cs)
14. PRF Impulsiveness (PRF Im) - reversed
15. 16PF G: Conscientious (16PF G)
16. 16PF $Q_3$: Controlled (16PF $Q_3$)
17. EPPS Change (EPPS chg) - reversed
18. GPPI Cautiousness (GPPI C)
19. JPI Risk Taking (JPI Rkt) - reversed
20. MMPI Mania (MMPI Ma) - reversed
21. PRF Change (PRF Ch) - reversed
22. DPQ Harmavoidance (DPQ Ha)
23. PRF Harmavoidance (PRF Ha)

## INTELLECTANCE SCALES

1. DPQ Traditionalism (DPQ Tr) - reversed
2. JPI Value Orthodoxy (JPI Vlo) - reversed
3. OPI Autonomy (OPI Au)
4. OPI (lack of) Religious Orientation (OPI RO)
5. 16PF $Q_1$: Experimenting (16PF $Q_1$)
6. JPI Complexity (JPI Cpx)
7. OPI Complexity (OPI Co)
8. OPI Practical Outlook (OPI PO) - reversed
9. GPPI Original Thinking (GPPI O)
10. JPI Innovation (JPI Inv)
11. JPI Breadth of Interest (JPI Bdi)
12. OPI Thinking Introversion (OPI TI)
13. OPI Theoretical Orientation (OPI TO)
14. OPI Estheticism (OPI Es)
15. PRF Understanding (PRF Un)
16. EPPS Intraception (EPPS int)
17. GZTS Thoughtfulness (GZTS T)

## AFFILIATION SCALES

1. DPQ Social Closeness (DPQ SC)
2. EPPS Affiliation (EPPS aff)
3. JPI Social Participation (JPI Spt)
4. PRF Affiliation (PRF Af)
5. 16PF A: Outgoing (16PF A)
6. EPPS Autonomy (EPPS aut) - reversed
7. PRF Autonomy (PRF Au) - reversed
8. 16PF $Q_2$: Self-sufficient (16PF $Q_2$) - reversed
9. JPI Conformity (JPI Cny)
10. PRF Social Recognition (PRF Sr)
11. EPPS Succorance (EPPS suc)
12. PRF Succorance (PRF Su)

## MISCELLANEOUS SCALES

1. CPI Achievement via Independence (CPI Ai)
2. CPI Psychological-mindedness (CPI Py)
3. CPI Feminity (CPI Fe) - reversed
4. CPS Masculinity (CPS M)
5. DPQ Achievement (DPQ Ach)
6. DPQ Absorption (DPQ Ab) - reversed
7. EPPS Achievement (EPPS ach)
8. EPPS Deference (EPPS def) - reversed
9. EPPS Exhibition (EPPS exh)
10. EPPS Abasement (EPPS aba) - reversed
11. EPPS Endurance (EPPS end)
12. EPPS Heterosexuality (EPPS het) - reversed
13. GZTS Masculinity (GZTS M)
14. JPI Responsibility (JPI Rsy)
15. JPI Social Adroitness (JPI Sca)
16. MMPI Hypochondriasis (MMPI Hs) - reversed
17. MMPI Hysteria (MMPI Hy) - reversed
18. MMPI Masculinity - femininity (MMPI Mf) - reversed
19. MMPI Paranoia (MMPI Pa) - reversed
20. OPI Altruism (OPI Am)
21. OPI Masculinity - Femininity (OPI MF)
22. PRF Achievement (PRF Ac)
23. PRF Endurance (PRF En)
24. PRF Play (PRF Pl)
25. PRF Sentience (PRF Se)
26. 16PF B: More Intelligent (16PF B)
27. 16PF I: Tender-minded (16PF I) - reversed
28. 16PF M: Imaginative (16PF M) - reversed
29. 16PF N: Shrewd (16PF N)

# APPENDIX B

## Individual Listing of Correlational Criterion-Related Validity Studies for Biodata Inventories

## CRITERION CATEGORY: TRAINING

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| 116-item Officer Biographical & Attitudinal Survey | Air Force pilot trainees | 593 | Pass/Fail of pilot training | P, CV | Range=.06-.14<br>Median $r$=.10 | Guinn, Vitola, & Leisey, 1976 |
| 8 assorted items | Skilled trade workers | 134 | Training grades & ratings | P | Range=.01-.47<br>Median $r$=.11 | Ronan, 1964 |
| 29 assorted items (Partial MAP-75) | Army enlistees | 278 | Completion of basic training | C | Range=.00-.28<br>Median $r$= .17 | HumRRO, 1976 |
| 45-item Biographical Questionnaire | Marine Corps drill instructors | 114 | Training grades | P, CV | $r$=.38 | Standlee & Abrahams, 1980 |
| 136-item Airman Assessment Inventory | Air Force police trainees | 4,502 | Completion of training | P | $r$=.39 | Guinn, Wilbourn, & Kantor 1977 |
| 5 assorted items | Army Special Forces trainees | 140 | Instructor ratings | P | Range=.11-.47<br>Median $r$=.40 | Berkhouse & Cook, 1961 |
| 8 assorted items | Navy diver school trainees | 296 | Pass/Fail of training | P, CV | Range=.15-.28<br>Median $r$=.16 | Biersner & Ryman, 1974 |
| 8 items: age, education, & family background | Marine Corps recruits | 193 | Final class standing | P, CV | Range=.21-.43<br>Median $r$=.23 | Farr, O'Leary, Pfeiffer, Goldstein, & Bartlett, 1971 |
| 10 assorted item groupings | Navy trainees | 115 | Completion/Non-completion, class rank | P, CV | Range=.55-.59<br>Median $r$=.57 | Helmreich, Bakeman, & Radloff, 1973 |
| 99 assorted items | Naval junior officers | 895 | Class grades | C | Range=.00-.22<br>Median $r$=.07 | Rhea, 1966 |

(Continued)

## CRITERION CATEGORY: TRAINING (CONTINUED)

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| 8 assorted items | British naval candidates | 519 | Training performance marks | P | Range=.11-.33<br>Median r=.20 | Gardner & Williams, 1973 |
| 9 assorted items | Navy enlisted trainees | 7,929 | Disenrollment & final grades | P, CV | Range=.10-.49<br>Median r=.21 | Hoiberg, Booth,& Berry 1977 |
| 87 assorted items | Real estate students | 644 | Attainment of license | P, CV | Range=.36-.46<br>Median r=.41 | Mitchell & Klimoski, 1982 |
| 119 assorted items | Oil refinery workers | 168 | Combination grades & ratings | P, CV | Range=.01-.25<br>Median r=.10 | Matteson, 1978 |
| 100 assorted items | Marine Corps drill instructor candidates | 759 | Average grade | P, CV | r=.38 | Wiskoff, 1980 |
| 67-item Enlisted Profile & 36-item Early Experience Questionnaire | Army enlistees | 2,212 | Peer & instructor ratings of basic training performance | P, CV | Range=.17-.28<br>Median r=.26 | Erwin & Herring, 1977 |
| 52-item Military Applicant Profile-75 | Male Army enlistees | 278 | Completion of basic training | C | Range=.30-.33<br>Median r=.32 | Seeley, Rosen, & Stroad 1978 |
| 11 items related to previous school satisfaction | Navy enlisted personnel | 1,200 | Completion of medical training | P, CV | Range=.00-.23<br>Median r=.13 | Webster, Booth, Graham, & Alf, 1978 |

## CRITERION CATEGORY: JOB PROFICIENCY

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| 15 first-order factors | Salesmen & district managers | 226 | Ratings & sales volume rankings | C | Range=.36-.50<br>Median r=.42 | Baehr & Williams, 1968 |
| 151 assorted items | Senior supervisors & staff | 1,251 | Composite measures of career progression | C, CV | Range=.44-.65<br>Median r=.60 | Laurent, 1970 |
| 22 assorted items | Clerical workers | 493 | Ratings | C, CV | Range=.00-.55<br>Median r=.14 | Black & McKinney, 1963 |
| 38 assorted items | Airline workers | 537 | Ratings | P | Range=.00-.36<br>Median r=.06 | Toole, Gavin, Murdy, & Sells 1972 |
| 300 assorted items | NASA scientists | 300 | Ratings | P, CV | Range=.48-.59<br>Median r=.54 | Taylor & Ellison, 1967 |
| 300 assorted items | NASA scientists | 800 | Organizational level | P, CV | r=.70 | Taylor & Ellison, 1967 |
| 300 assorted items | Scientists & engineers | 355 | Publications & patents | C, CV | Range=.37-.59<br>Median r=.48 | Taylor, 1962 |
| 165 assorted items | Naval company commanders | 809 | Ratings & rankings | P, CV | r=.20 | Manese, Skrobiszewski, & Abrahams, 1976 |
| 13 family background items | Male & female Israeli Army ex-soldiers | 914 | Military rank at discharge | C, CV | r=.36 (males)<br>r=.18 (females) | Nevo, 1976 |
| 104 assorted items | Salesmen from 4 European countries & U.S. | 362 | Pooled multiple overall ratings | C, CV | Range=.24-.56<br>Median r=.38 | Hinrichs, Haanpera, & Sonkin, 1976 |
| Number & type of biodata items not reported | Plant foremen | 173 | Supervisory rankings | C, CV | r=.54 | Cornelius, 1977 |

(Continued)

## CRITERION CATEGORY: JOB PROFICIENCY (CONTINUED)

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| Number & type of biodata items not reported | MBAs with 5 or 10 years experience | 266 | Salary | P, CV | *r*=.16 - 5 years<br>*r*=.22 - 10 years | Harrell, Harrell, McIntyre, & Weinberg, 1977 |
| 15 biodata items | Management consultants | 127 | Performance ratings | P, CV | *r*=.58 | Miner, 1971 |
| Ratings of fathers' occupational level | Male MBAs | 337 | Career success measures | C | Range=.00-.12<br>Median *r*=.06 | Porter, 1965 |
| 220 assorted items | Managers | 382 | Assessment center rating | C, CV | *r*=.40 | Ritchie & Boehm, 1977 |
| 100 assorted items | Electrical engineers | 100 | Possession of patents | C, CV | *r*=.62 | Colson, 1977 |
| 200 items: family, academics, adult life | Scientists & engineers | 203 | Performance ratings & objective measures | C, CV | Range=.41-.60<br>Median *r*=.49 | Ellison, James, & Carron 1970 |
| 160 assorted items | Pharmaceutical scientists | 157 | Various ratings, rank & salary | C, CV | Range=.01-.51<br>Median *r*=.22 | Tucker, Cline, & Schmitt 1967 |
| 119 assorted items | Oil refinery workers | 168 | Performance ratings | P, CV | Range=.00-.22<br>Median *r*=.05 | Matteson, 1978 |
| 100 assorted items | Marine Corps drill instructor candidates | 759 | Supervisory ratings | P, CV | *r*=.16 | Wiskoff, 1980 |
| 185-item Registered Nurse Biographical Inventory | Registered Nurses | 1,018 | Performance composite of rankings & tests | C, CV | Range=.18-.29<br>Median *r*=.21 | Dyer, Cope, Monson, & Drimmelen, 1972 |
| 278-item Aptitude Index Battery | Life Insurance agents | 12,453 | Value of insurance sold | P | Range=.13-.26<br>Median *r*=.20 | Brown, 1981 |
| 9 assorted items | Marine Corps enlisted men of low mental ability | 1,342 | Composite measure of job performance | P | Range=.06-.20<br>Median *r*=.09 | Plag, Goffman, & Phelan, 1971 |
| 17 assorted items | Navy enlisted men of low mental ability | 1,260 | Composite measure of job performance | P | Range=.08-.24<br>Median *r*=.14 | Plag, Goffman, & Phelan 1967 |

(Continued)

## CRITERION CATEGORY: JOB PROFICIENCY (CONTINUED)

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| 4 items: age, schooling, suspensions, arrests | Navy enlistees in 1st 2 years of service | 3,964 | Composite job effectiveness | P, CV | *r*=.32 | Hoiberg & Pugh, 1978 |
| 8 assorted items | Skilled trade workers | 134 | Ratings & promotions | P | Range=.06-.31<br>Median *r*=.14 | Ronan, 1964 |
| 9 assorted items | Naval recruits | 3,630 | Recommendation for re-enlistment | P, CV | Range=.04-.32<br>Median *r*=.16 | Plag & Goffman, 1966 |
| 25 assorted items | Marine Corps drill instructors | 114 | Job performance ratings | P, CV | *r*=.16 | Standlee & Abrahams, 1980 |
| Standard Oil Biodata Form | Skilled craftsmen & applicants | 328 | Job performance rankings | P | 10 study median *r*=.45 | Standard Oil Company, 1962 |
| Standard Oil Biodata Form | Engineering & technical personnel | 157 | Job performance rankings | P | 2 study median *r*=.29 | Standard Oil Company, 1962 |
| Standard Oil Biodata Form | Office workers | 292 | Job performance rankings | P | Median *r*=.40 | Standard Oil Co., 1962 |
| Standard Oil Biodata Form | Supervisors | 136 | Job performance rankings | P | Median *r*=.41 | Standard Oil Co., 1962 |
| Standard Oil Biodata Form | Potential managers | 64 | Composite managerial success | P | Median *r*=.35 | Standard Oil Co., 1962 |
| 118 assorted items | Research scientists | 132 | Rankings & ratings of job performance | C, CV | Range=.14-.57<br>Median *r*=.38 | Buel, 1965 |
| 33 assorted items | Research scientists | 132 | Rankings & ratings of job performance | C, CV | Range=.29-.52<br>Median *r*=.33 | Buel, 1966 |
| 484 assorted items | Scientists & engineers | 100 | Patents & ratings of performance & creativity | C, CV | *r*=.52 | Smith, Albright, Glennon, & Owens, 1961 |

## CRITERION CATEGORY: JOB INVOLVEMENT

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| 35 assorted items | Naval Academy cadets | 251 | Attrition | P, CV | Range=.11-.22<br>Median r=.17 | Dann & Abrahams, 1970 |
| 70 assorted items | Navy ROTC students | 400 | Tenure | P, CV | Range=.15-.17<br>Median r=.17 | Neumann, Githens, & Abrahams 1967 |
| 16 assorted items | Clerical employees | 160 | Tenure | C, CV | Range=.56-.58<br>Median r=.57 | Cascio, 1976 |
| 13 items: family situation, education & expectations | Credit Union workers | 96 | Tenure | P, CV | r=.74 | Federico, Federico, & Lundquist, 1976 |
| 33 assorted items | Clerical employees | 100 | Tenure | C, CV | Range=.24-.30<br>Median r=.27 | Gebhardt , 1979 |
| 11 items: previous job, education & age | Female clerical workers | 81 | Tenure | P, CV | Range=.29-.56<br>Median r=.43 | Johnson, Newton, & Peek 1979 |
| Unspecified number of application blank items | Clerical workers | 150 | Tenure | C, CV | r=.56 | Lee & Booth, 1974 |
| 4 items: age, job grade, marital status, & company tenure | Clerical workers | 105 | Attrition | P | Range=.11-.30<br>Median r=.25 | Waters, Roach, & Waters 1976 |
| 37 assorted items | Black, male custodial workers | 201 | Absenteeism | C, CV | r=.53 | Schwartz, 1975 |
| 10 assorted items | Life insurance agents | 14,738 | Tenure | P, CV | Range=.25-.42<br>Median r=.33 | Brown, 1978 |
| 68 assorted items | Navy enlisted men (mechanics & electricians) | 6,635 | Re-enlistment | P, CV | Range=.13-.21<br>Median r=.17 | Dann & Abrahams, 1969 |
| 102-item Delinquent Behavior Inventory | Navy enlisted personnel | 2,485 | Basic training attrition | P | r=.25 | Yellen, 1975 |
| 51-item Early Experience Questionnaire | Male Army enlistees | 4,282 | Completion of 180 days service | P, CV | Range=.20-.41<br>Median r=.31 | Frank & Erwin, 1978 |
| 67-item Enlisted Profile & 36-item Early Experience Questionnaire | Army enlistees | 2,182 | Completion of 180 days service | P, CV | Range=.32-.45<br>Median r=.41 | Erwin & Herring, 1977 |

(Continued)

## CRITERION CATEGORY: JOB INVOLVEMENT (CONTINUED)

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| 60-item Military Applicant Profile | Army enlistees | 1,457 | Completion of 180 days service | P | Range=.01-.40<br>Median r=.24 | Haymaker & Erwin, 1980 |
| 16 assorted items | Female clerical personnel | 224 | Turnover | C, CV | Range=.33-.49<br>Median r=.41 | Buel, 1964 |
| 31 assorted items | Clerical workers | 561 | Turnover | P, CV | Range=.36-.48<br>Median r=.42 | Shott, Albright & Glennon 1963 |
| 37 assorted items | Female clerical workers | 211 | Tenure | P, CV | r=.57 | Wernimont, 1962 |
| 19 assorted items | Unskilled workers | 150 | Tenure | P, CV | r=.31 | Scott & Johnson, 1967 |
| 86 assorted items | Vocational Rehabilitation Clients | 400 | Tenure | P, CV | r=.53 | Ehrle, 1964 |
| 22 assorted items | Clerical workers | 493 | Turnover | C, CV | Range=.00-.51<br>Median r=.18 | Black & McKinney, 1963 |
| 46 assorted items | Production workers | 232 | Tenure | P, CV | r=.05 | Lefkowitz, 1972 |
| 34-item Naval Background Questionnaire | Naval officers | 1,164 | Tenure | P, CV | Range=.07-.17<br>Median r=.15 | Neumann, Githens, & Abrahams 1967 |
| 18-item Quest 1 Questionnaire | Naval female enlistees | 243 | 18-month attrition | P | r=.24 | Wilcove, Thomas & Blankenship, 1979 |
| 22 assorted items | Black, disadvantaged job placementees | 702 | 3 & 6-month attrition | P | Range=.37-.53<br>Median r=.45 | Richardson, Bellows, Henry, & Co., 1971 |
| 246 assorted items | Project Talent, college students | 20,000+ | School attendance | P, CV | r=.60 | Prediger, 1969 |
| 10 assorted items | Female library clerks | 80 | Tenure | P, CV | r=.10 | Cunningham & DeVitt, 1968* |
| 20 assorted items | Male machinists | 74 | Tenure | P, CV | r=.05 | Oliver, 1969* |
| 25 assorted items | Unskilled male city employees | 100 | Tenure | P, CV | r=.10 | Oliver, 1969* |

(Continued)

*Cross-validated by Schwab & Oliver, 1974.

## CRITERION CATEGORY: JOB INVOLVEMENT (CONTINUED)

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| 33 assorted items | Unskilled female production workers | 166 | Tenure | P, CV | $\underline{r}$=.10 | Packard, 1971* |
| 14 assorted items | Female office workers | 248 | Tenure | P, CV | $\underline{r}$=.61 | Dunnette, Kirchner, Erickson & Banas, 1960 |
| 40 assorted items | Female office workers | 120 | Tenure | P, CV | $\underline{r}$=.57 | Fleishman & Berniger, 1960 |
| 100-item History & Opinion Inventory | Air Force male enlistees | 15,252 | 2-year attrition | P, CV | Range=.19-.27<br>Median $\underline{r}$=.23 | Guinn, Johnson & Kantor 1975 |

*Cross-validated by Schwab & Oliver, 1974.

## CRITERION CATEGORY: ADJUSTMENT

| Instrument | Sample | N | Criterion | Method | Results | Reference |
|---|---|---|---|---|---|---|
| 5 & 10 assorted items | Grocery & retail store employees | 100 | Employee theft | C, CV | Range=.17-.63<br>Median *r*=.33 | Rosenbaum, 1976 |
| 6 items: previous delinquent behavior & family background | Navy enlistees in basic training | 2,043 | Substance abuse | C | Range=.16-.39<br>Median *r*=.26 | Bucky, Edwards, & Thomas 1974 |
| 8 assorted items | Skilled trade workers | 134 | Emotional problems, grievances, absences | P | Range=.00-.13<br>Median *r*=.07 | Ronan, 1964 |
| 26-item Prediction of Drug Abuse scale | Male Air Force basic trainees | 6,455 | Drug abuse discharge | P, CV | *r*=.36 | Lachar, Sparks, Larsen, & Bisbee, 1974 |
| 18-item Prediction of Emotional Instability scale | Male Air Force basic trainees | 6,185 | Severe adjustment problems to basic training | P, CV | *r*=.20 | Lachar, Sparks, Larsen, & Bisbee, 1974 |
| 100-item History & Opinion Inventory | Male Air Force basic trainees | 15,252 | Discharge for unsuitability & undesirability | P | Range=.07-.31<br>Median *r*=.17 | Guinn, Johnson, & Kantor 1975 |

# APPENDIX C

## Individual Listing of Correlational Criterion-Related Validity Studies*

*Information from these tables is summarized in the body of the text.

CRITERION CATEGORY: JOB INVOLVEMENT

| Inventory | Sample | N | Criterion | Results | Reference |
|---|---|---|---|---|---|
| Kuder Preference Record | Farmers | 50 | Job Satisfaction (Concurrent) | Range = .02 - .28<br>Median *r* = .16 | Brayfield & Marsh, 1957 |
| Kuder Preference Record | Forest Rangers | 125 | Tenure (Predictive) | *r* = .19 | Mayeske, 1964 |
| Minnesota Vocational Interest Inventory | Nurses | 501 | Job Satisfaction (Predictive) | *r* - .08 | Martin, 1968 |
| Navy Vocational Interest Inventory | Navy Enlisted Men | 789 | Re-enlistment (Predictive) | Range = .22 - .40<br>Median *r* = .29 | Lau, Lacey, & Abrahams, 1969 |
| Navy Vocational Interest Inventory | Former Navy Enlisted Men | 127 | Job Satisfaction (Predictive) | Mean *r* = .16 | Lau & Abrahams, 1971 |
| Navy Vocational Interest Inventory | First-term Navy Enlisted Men | 5,485 | Job Satisfaction (Concurrent) | Range = .13 - .46<br>Median *r* = .33 | Dann & Abrahams, 1977 |
| Navy Vocational Interest Inventory | Navy Enlisted Recruits | 3,505 | Job Satisfaction (Predictive) | Range = .08 - .39<br>Median *r* = .23 | Dann & Abrahams, 1977 |
| Strong Vocational Interest Blank | Policemen | 25 | Job Satisfaction (Concurrent) | *r* = .35 | Kates, 1950 |
| Strong Vocational Interest Blank | Former College Students | 663 | Job Satisfaction (Predictive) | Mean *r* = .27 | Strong, 1955 |
| Strong Vocational Interest Blank | Former College Students | 663 | Job Satisfaction (Concurrent) | Mean *r* = .34 | Strong, 1955 |
| Strong Vocational Interest Blank | Life Insurance Agents | 520 | Tenure (Predictive) | Range = .16 - .50<br>Median *r* = .29 | Ferguson, 1958 |

(Continued)

CRITERION CATEGORY: JOB INVOLVEMENT (Continued)

| Inventory | Sample | N | Criterion | Results | Reference |
|---|---|---|---|---|---|
| Strong Vocational Interest Blank (local key) | Computer Programmers | 1,003 | Job Satisfaction (Concurrent) | *r* = .15 | Perry, 1967 |
| Strong Vocational Interest Blank | Insurance Managers | 140 | Job Satisfaction (Concurrent) | *r* = .38 | Dore, 1970 |
| Strong Vocational Interest Blank | Managers | 140 | Job Satisfaction (Concurrent) | Range = .07 - .42<br>Median *r* = .19 | Dore & Meachum, 1973 |
| Strong Vocational Interest Blank | Managers | 101 | Job Satisfaction (Concurrent) | *r* = .38 | Wiener & Klein, 1978 |
| Strong-Campbell Interest Inventory | Former College Students | 130 | Job Satisfaction (Concurrent) | *r* = .31 | Worthington & Dolliver, 1977 |
| Strong-Campbell Interest Inventory | Graduate Students | 43 | Job Satisfaction (Predictive) | Range = .01 - .29<br>Median *r* = .15 | Betz & Taylor, 1982 |
| Vocational Interest Career Examination | Air Force Recruits | 18,207 | Job Satisfaction (Predictive) | Range = .20 - .57<br>Median *r* = .32 | Alley, Wilbourn, & Berberich, 1976 |
| Vocational Interest Career Examination | First-term Air Force Enlisted Men | 3,000 | Job Satisfaction (Concurrent) | Range = .25 - .46<br>Median *r* = .44 | Alley & Matthews, 1982 |
| Local Inventory | Engineers | 109 | Job Satisfaction (Concurrent) | Range = .44 - .62 | Meir & Erez, 1981 |
| Local Inventory | Registered Nurses | 126 | Job Satisfaction (Concurrent) | Range = .08 - .44 | Hener & Meir , 1981 |

CRITERION CATEGORY: JOB PROFICIENCY

| Inventory | Sample | N | Criterion | Results | Reference |
|---|---|---|---|---|---|
| Kuder Preference Record | Farmers | 50 | Performance Ratings (Concurrent) | Range = .05 - .40<br>Median $r$ = .08 | Brayfield & Marsh, 1957 |
| Kuder Preference Record | Forest Rangers | 464 | Performance Ratings (Concurrent) | Range = .01 - .15<br>Median $r$ = .07 | Miner, 1960 |
| Navy Vocational Interest Inventory | Navy Enlisted Recruits | 783 | Performance Ratings (Predictive) | Range = .15 - .38<br>Median $r$ = .25 | Lau & Abrahams, 1970 |
| Navy Vocational Interest Inventory | Navy Enlisted Men | 1,133 | Performance Ratings (Predictive) | Range = .05 - .19<br>Median $r$ = .15 | Dann & Abrahams, 1977 |
| Navy Vocational Interest Inventory | Former Navy Enlisted Men | 127 | Self-reported Job Performance (Predictive) | Mean $r$ = .20 | Lau & Abrahams, 1971 |
| Strong Vocational Interest Blank | Managers | 116 | Administrative Level (Predictive) | $r$ = .24 | Williams & Horrell, 1964 |
| Strong Vocational Interest Blank | Supervisors | 59 | Performance Ratings (Predictive) | Range = .01 - .34<br>Median $r$ = .20 | Strong, 1943 |
| Strong Vocational Interest Blank | Foremen & Assistant Foremen | 30 | Effectiveness Ratings (Concurrent) | $r$ = .38 | Schultz & Barnabas, 1945 |
| Strong Vocational Interest Blank | Bakery Shop Manager Trainees | 37 | Costs/Sales Ratio (Predictive) | $r$ = .53 | Knauft, 1951 |
| Strong Vocational Interest Blank | Stanford MBA Alumni | 195 | Measures of Administrative Level (Predictive) | Range = .10 - .48<br>Median $r$ = .33 | Porter, 1962 |

(Continued)

CRITERION CATEGORY: JOB PROFICIENCY (Continued)

| Inventory | Sample | N | Criterion | Results | Reference |
|---|---|---|---|---|---|
| Strong Vocational Interest Blank (local key) | Managers | 468 | Performance Rankings (Concurrent) | $\underline{r}$ = .33 | Nash, 1966 |
| Strong Vocational Interest Blank (revised key) | Same as in Nash 1966 Study | 468 | Performance Rankings (Concurrent) | $\underline{r}$ = .30 | Johnson & Dunnette, 1968 |
| Strong Vocational Interest Blank (local key) | Naval Academy Cadets | 2,400 | Military Bearing Ratings (Predictive) | Median $\underline{r}$ = .31 | Abrahams & Neumann, 1973 |
| Strong Vocational Interest Blank (local key) | Navy Recruiters | 385 | Performance Ratings (Predictive) | Range = .02 - .26<br>Median $\underline{r}$ = .19 | Borman, Toquam, & Rosse, 1979 |

CRITERION CATEGORY: TRAINING (Continued)

| Inventory | Sample | N | Criterion | Results | Reference |
|---|---|---|---|---|---|
| Vocational Preference Inventory | Nursing Students | 53 | School Grades (Predictive) | Range = .03 - .25<br>Median r = .14 | Johnson, 1973 |
| Locally Developed Scales | Air Force Enlisted Personnel | 1,000 | Final School Grade (Predictive) | Range = .03 - .30<br>Median r = .14 | Pickrel, 1954 |