1968

Development And Evaluation Of Four Types Of Structured Personality Scales

John Addison Neill

DEVELOPMENT AND EVALUATION OF FOUR TYPES OF STRUCTURED PERSONALITY SCALES

by

John A. Neill

Department of Psychology

Submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Canada
June, 1968

ABSTRACT

The purpose of the present investigation was to clarify the nature of desirability and faking biases in personality assessment, and with reference to these two kinds of biases, to compare four different types of structured personality scales. Recent advances in personality measurement theory and research, particularly those regarding the subject of construct validity, were utilized in the development and evaluation of scales for measuring Impulsivity, Risk Taking, and Self Esteem. The four types of structured scales developed were nonforced endorsement scales, forced-choice endorsement scales, nonforced desirability-judgment scales, and forced-choice desirability-judgment scales. Scale development was based on item analyses of the data of a large sample of subjects, and the resulting scales were cross-validated against peer behaviour-judgment and self-behaviour-judgment criteria in a new sample of subjects. In order to compare the different types of scales as to their resistance to faking, the scales were administered with standard instructions and with instructions to fake.

The scales were shown to have favourable construct properties in the item-analysis sample; these properties were not attenuated with cross-validation on either the standard-instruction or fake-instruction sample. In both validation samples virtually all of the scales were reliable, relatively uncorrelated with a measure of desirability response style, and able to discriminate among the three constructs.

The four types of scales varied in average validity against behaviour-judgment criteria in different ways in the two validation samples. In the standard-instruction sample, nonforced endorsements were most valid, followed by forced-choice endorsements, nonforced desirability judgments, and forced-choice desirability judgments, in that order. Of the four, only the last had zero validity. In contrast, in the fake-instruction sample all scales had some validity and the differences among the four types of scales in average validity were negligible. This finding represented a decrease in validity with faking for the endorsement scales and an increase in validity with faking for the desirability-judgment scales.

The results are discussed in terms of implications regarding the use of forced-choice and judgmental methods in personality assessment, the nature and control of desirability and faking biases, and the importance of the principles of construct validity.

ACKNOWLEDGMENTS

The author sincerely appreciates the help and encouragement which have come from a number of sources. Dr. Douglas N. Jackson contributed his research talents through many hours of discussion at all stages of the project. The other members of the advisory committee--Dr. D. A. Chambers, Dr. R. C. Gardner, Dr. W. E. Reitz, and Dr. T. F. Siess--provided helpful suggestions regarding the proposed research, as well as constructive criticisms of the manuscript. Miss V. Griffith, Director of the Victoria Hospital School of Nursing, arranged for the participation of student nurses, without whose cooperation the research could not have been completed. Finally, the encouragement and criticism from my patient wife, Betty, have helped to keep the whole project in perspective.

Thanks are due to the Alcoholism and Drug Addiction Research Foundation of Ontario, the Ontario Mental Health Foundation, the University of Western Ontario, and the University of windsor for financial support for the research.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Important advances have been made in the field of personality assessment in the past three decades. Psychologists have learned a great deal about how to recognize and deal with a number of sources of error in the assessment of personality. However, the problems of desirability response style and faking are not yet well understood nor easily dealt with. The purpose of the present investigation was to clarify the nature of desirability and faking biases, and with reference to these two kinds of biases, to compare and evaluate four different types of structured personality scales. Recent developments in personality measurement theory and research were utilized in the development and validation of experimental personality scales for measuring the traits of Impulsivity, Risk Taking, and Self Esteem. The four types of structured scales developed were nonforced endorsement scales, forced-choice endorsement scales, nonforced desirability-judgment scales, and forced-choice desirability-judgment scales.

It should be emphasized that other sources of error besides desirability and faking affect personality assessment; for example, acquiescence response style, the tendency to respond consistently 'true' or 'false' to personality items, has been a troublesome source of error with personality questionnaires. The theory and research regarding acquiescence have been reviewed recently by Damarin and Messick (1965), Jackson (1967a), and Messick (1967). However, like many other problems impinging on personality assessment, acquiescence was not under investigation in the present research, and consequently, was not dealt with beyond the level of attempting to minimize its influence on the experimental personality scales.

Since the examination of more than a few of the many available types of personality scales is beyond the reasonable bounds of a single research project, some major restrictions had to be placed on the research. First, the research was limited to structured personality tests. A structured test may be defined as one in which the subject is given highly specific instructions with regard to the performance of a well-defined task involving a limited range of acceptable responses. In the present research the complete instructions to subjects were in printed form and items required the use of either a 2-category or a 9-category response format. Secondly, only self-descriptive statements were used in composing items. The restriction of the present research to structured personality tests composed of self-descriptive statements excluded such widely used assessment procedures as projective methods, clinical interviews, open-ended questionnaires, and adjective checklists.

Before embarking on a discussion of desirability and faking, the concepts of reliability and validity require clarification. Reliability is used in the present context in the internal-consistency sense. A highly reliable scale is homogeneous with respect to a single dimension of individual consistency. Only the classical definition of validity is required at this juncture; the modern conception of validity is discussed in detail in a later section. For the present, validity is considered to be simply the correlation between personality scale scores and nontest criterion measures of the same trait (American Psychological Association, 1966).

Desirability and Faking Biases

The Problem of Desirability

The personality assessment literature is replete with discussions of desirability concepts. Desirability has been treated both as an individual-difference variable and as an item characteristic. In the present investigation desirability was used principally in the individual-difference sense; the research was in part concerned with the identification and control of desirability response style. However, the discussion of desirability response style presupposes an understanding of desirability as an item characteristic. Therefore, an explanation of desirability as an item characteristic is given before desirability response style is discussed.

Desirability as an item characteristic. It has long been recognized that personality items, self-descriptive statements in the present context, vary greatly with respect to how desirable or undesirable they are considered by people in general. The majority of people, for example, would probably concur in the belief that a person's agreeing with the item 'I get along well with other people' reflects a moderately desirable trait, or that his agreeing with 'I like to hurt people just for the fun of it' reflects an extremely undesirable trait. The item itself is assumed to occupy some position on a latent continuum ranging from extremely desirable to extremely undesirable. The continuum has frequently been referred to as the 'social' desirability continuum because the basic question is about the desirability of the item in people in general, not about its desirability in oneself. In addition, the social desirability of an item is based on the opinion of a group of people, not just on the opinion of one individual (Edwards, 1967a). Only the simpler term 'desirability' is used in the present context.

Following from the notion that items occupy positions on a desirability continuum, items have been scaled for desirability. The usual procedure for determining an item's desirability scale value (DSV) has been to have a large sample of people rate the desirability of the item on a 7-point or 9-point scale, and then to subject the ratings to equal-appearing intervals scaling (e.g., Edwards, 1953) or successive intervals scaling (e.g., Messick & Jackson, 1961). The simple calculation of the arithmetic mean of the ratings yields scale values identical to those of the equal-appearing intervals method and was the method adopted in the present investigation. Whatever the method, the resulting number indicates how desirable or undesirable in people in general a sample of subjects believe the item to be.

The DSVs of items are relevant to personality assessment because it has been shown that the probability that an item will be endorsed, that is, answered 'true,' is positively related to the item's DSV. Using a set of heterogeneous items, Edwards (1953) found that the item popularities or endorsement proportions computed from the responses of one sample of subjects correlated .87 with the same items' DSVs based on another sample. Similar correlations between item popularities and DSVs have been reported by a number of other investigators (Cruse, 1965, 1967; Gordon, 1953; Hanley, 1956; Messick, 1964; Rosen, 1956). Edwards' (1953, 1957) interpretation of the correlation was that an item extreme on the desirability dimension tends to elicit responding to the item's desirability rather than to its content. The connotation of such an interpretation is that most people tend to 'lie' in response to personality items in order to create favourable impressions of themselves. The unfortunate implications regarding the validity of personality scales are obvious. However, an equally plausible interpretation is that some traits are judged desirable because they are more frequently occurring. Interpretations of the finding are still under discussion (Boe & Kogan, 1964, 1966; Cruse, 1965, 1967; Edwards, 1967a, 1967b; Fox, 1967; Jackson & Messick, 1962; Jackson & Singer, 1967; Norman, 1967; Scott, 1963; J. S. Wiggins, 1962, 1966). The research which has ensued has clarified one issue. The correlations reported by Edwards and others were correlations computed across items between two means per item. As such, they were strictly group statistics, and while relevant to the question of validity, they did not constitute evidence for the early interpretations regarding the'sinister' intentions of individuals (Norman, 1967).

Desirability as an individual-difference variable. It has long been apparent that subjects can and do respond to other properties of items besides their psychological content (Cronbach, 1946; Lorge, 1937). It has been evident also that there are individual consistencies in tendencies to respond to items in noncontent ways (Berg, 1967). Desirability has been treated as a response style because people vary considerably and consistently in the tendency to respond to items in terms of their desirability rather than in terms of their psychological content (Jackson & Messick, 1958). For example, a person might respond 'true' to a certain item, not because it was true of him, but because the item described a desirable state of affairs. Obviously, if desirability response style accounted for more than a very minor proportion of the variance in the scores for a given personality scale, it would lower the validity of the scale as a measure of a trait.

There is evidence that desirability response style is not elicited to the same degree by all items. Items extreme on the desirability dimension tend to elicit desirability response style more than do neutral items (Bloxom, 1967; Boe & Kogan, 1964, 1966; Jackson & Messick, 1958, 1961, 1962; Scott, 1963; Stricker, 1963). This relationship between the presence of desirability response style and the desirability of items indicates that during item preparation and item selection for a personality scale steps may be taken to lessen the influence of desirability response style on the scale.

The Problem of Faking

The problem of a faking is regarded as a general problem of self-report measures. It is suspected that people to varying degrees consciously attempt to manage impressions by endorsing items that are not true of them and by denying items that are true. Deliberate faking is assumed to be especially prevalent in settings where practical decisions are based on test scores; for example, faking is assumed to be a serious problem in selection and placement situations (Gordon & Stapleton, 1956; Hedberg & Baxter, 1963; Izard & Rosenberg, 1958; Krug, 1958a, 1958b; Norman, 1963a).

The problem of desirability response style has often been confused with the problem of faking. Clearly, both tend to be elicited more by items with extreme DSVs than by neutral items. In the present context, however, they are treated as conceptually different. While the term faking is used to mean consciously attempting to create a favourable impression by responding in the desirable direction, reference to desirability response style is not meant to imply any awareness of desirability responding on the part of the subject. It remains unclear to what extent faking is best conceived of as an individual-difference variable and to what extent it is best thought of as behaviour elicited by properties of the personality scale, the personality items, or the testing situation. Some methodological and theoretical advances which have a bearing on the foregoing distinctions are presented below. A primary focus of the present investigation is the clarification of some of these issues.

Proposed Solutions

The problems of desirability bias and faking have been outlined in the preceding section. The present section describes several of the more important attempts to assess and control desirability response style and faking. The approaches discussed are empirical keying of items, validity scales, forced-choice methods, and judgmental methods. Each has enjoyed some success but also has had shortcomings. The present research included empirical tests of variations of three of these approaches.

Empirical Keying

The empirical keying or criterion keying of items in the development of structured personality tests has had a long history in personality assessment (see Meehl, 1945). The typically-used procedure begins with the administration of a large heterogeneous pool of self-descriptive statements to preselected criterion groups of people; for example, 500 items of varied content are administered to 100 alcoholic persons and 100 nonalcoholic social drinkers, and those items whose endorsement frequencies differ significantly between the two groups are selected to form an alcoholism scale. Just as the selection of items is purely empirical, so also is the keying of items; continuing the same example, items with significantly higher endorsement frequencies for alcoholics are keyed for alcoholism.

Unlike earlier methods in which item selection and keying were based on the face validity of items, empirical keying methods required no a priori decisions about either the relevance of items or their keying. This was claimed as a major advantage of empirical keying over other methods of scale construction (Meehl, 1945; Seeman, 1952) for it allowed the selection of 'subtle' items. Since subtle items were items lacking in face validity but discriminating between criterion groups, it was argued that they would be difficult to fake (Meehl, 1945; Seeman, 1952).

The above claimed advantage of empirical keying has lost much of its potency in the last two decades, however, because in the interim psychologists involved with personality scale construction have gained a much better understanding both of personality and of the psychology of structured test behaviour (Loevinger, 1957). Because of this increased understanding, many items which would have been considered subtle in the 1940's are now being prepared from considerations of psychological theory, without resort to empirical keying (e.g., Jackson, 1967b; Neill & Jackson, 1967). In addition, there has accumulated a great deal of empirical evidence that the so-called subtle items are actually less valid than the obvious items, especially when cross-validated (Brozek & Erickson, 1948; Duff, 1965; Fricke, 1957; Goldberg & Slovic, 1967; McCall, 1958; Norman, 1963b). In fact, before cross-validation, the validity of all items derived solely from empirical keying should be suspect, since there is always a probability of an invalid item being included in a scoring key by chance (Cureton, 1950; Loevinger, 1957; Travers, 1951b).

Besides the problems with cross-validity, there are other serious disadvantages to empirical keying. Criterion groups are not homogeneous, and consequently, scales based on group differences with regard to a heterogeneous item pool tend to be unreliable (cf. Meehl, 1945) in the internal-consistency sense. Furthermore, scales based on empirical keying have been found lacking in discriminant validity (Edwards, 1957; Jackson & Messick, 1961, 1962), especially when overlapping keys were allowed, as was the case with the Minnesota Multiphasic Personality Inventory (MMPI). A scale is said to have discriminant validity when it correlates more highly with construct-relevant nontest criteria than with measures of irrelevant constructs (Campbell & Fiske, 1959). Contributing to empirically derived scales' inadequate discriminant validity is their evident saturation with response style variance (Jackson, 1960; Jackson & Messick, 1958, 1962; Messick, 1962). If the criterion groups used in constructing a scale differ with respect to desirability response style, for example, then empirical keying of items tends to capitalize on this difference. There is nothing intrinsically wrong with capitalizing on any factors which reliably distinguish between groups when the sole purpose of the scale is gross classification. The problem lies in the confusion that ensues when claims are made with regard to the substantive or construct properties of the scale. Almost nothing can be validly claimed about the psychological content of such a scale (Loevinger, 1957).

From the above discussion it should be clear that empirical keying has not solved the problems of faking and desirability bias. The early promise that subtle items would help solve the problems has not been fulfilled. It is important to recognize, however, that the advent of empirical keying was an important advance in its day; and furthermore, that the resultant instruments, especially the MMPI, have fostered research which has elucidated many methodological problems plaguing personality assessment.

Validity Scales

One of the earliest methods of handling faking and other forms of bias was the inclusion of validity scales in personality questionnaires (see Meehl & Hathaway, 1946). A validity scale is a set of items which, when keyed in a specified manner, detects the presence of responding in inappropriate ways. Examples of validity scales are the Lie scale of the MMPI, which is presumed to detect attempts to fake in the desirable direction (Meehl & Hathaway, 1946); the Infrequency scale of the Personality Research Form (PRF), which was designed to detect nonpurposeful responding (Jackson, 1967b); and the Desirability scale also of the PRF, which was designed to measure desirability response style. The PRF Desirability (PRF-Dy) scale, in addition to serving as a validity scale or detection device, is unlike the other two examples in that it is sensitive along an entire bipolar dimension of individual consistency.

Before discussing the effectiveness of validity scales in overcoming the problems of desirability response style and faking, three scales for measuring individual differences in tendencies to respond desirably are discussed. The three methods are the Edwards Social Desirability (SD) scale (Edwards, 1954), the Marlowe-Crowne Social Desirability (M-C-SD) scale (Crowne and Marlowe, 1960), and the PRF-Dy scale (Jackson, 1967b). These three scales are representative of a much larger set of validity scales (e.g., see Bartlett, 1966; Ford, 1964; Jackson, 1967b; Meehl & Hathaway, 1946; Messick, 1962; Norman, 1963a; Schanberger, 1967).

Edwards' SD scale (1954) was composed of 39 MMPI items of heterogeneous content, all of which had extreme DSVs. The items were keyed in the desirable direction without reference to item content; for example, a subject scored one point for each desirable item answered 'true' and one point for each undesirable item answered 'false,' and his total score was presumed to be a measure of his tendency to respond desirably or undesirably to personality items. Edwards' rationale for selecting items with extreme DSVs for his SD scale was based on the high correlation obtained between item popularity and DSV.

The M-C-SD scale (Crowne & Marlowe, 1960) was different from the Edwards SD scale both in rationale and method of construction. The items in the M-C-SD scale were expressly free from pathological content, thereby avoiding a major criticism of the Edwards SD scale. Crowne and Marlowe selected items which had high DSVs, but very low popularities, or vice versa. A comparison of item-selection methods and a perusal of the items reveals a striking similarity between the M-C-SD scale and the MMPI Lie scale. The M-C-SD scale is probably more suited to the detection of faking than to the measurement of individual differences in desirability response style (Bloxom, 1967).

Jackson's PRF-Dy scale (1967b) was the product of the item analysis of a large pool of heterogeneous items keyed for the desirable response. Responses to items were correlated with the total scores from the Desirability item pool as well as with scores from a number of PRF content scales. After rejecting items which correlated as high with any content scale as with the Desirability item pool, the 40 remaining items with the highest correlations with the Desirability item pool formed two statistically parallel 20-item PRF-Dy scales to be used either together or separately.

A method of scaling items for desirability was outlined earlier. In the construction of the PRF-Dy scales a second index of item desirability was employed, namely, biserial correlation with total scores from a desirability response style scale. The latter method is more pertinent to the assessment of the degree to which an item elicits desirability response style; it was used extensively in the present investigation.

Research using desirability scales of the types discussed has helped to clarify many problems in personality assessment. Partly because of such research, the limitations of existing personality scales are becoming more apparent to psychologists in general. The research has facilitated also the construction of such questionnaires as the PRF, which have fewer response-style-eliciting properties than earlier instruments such as the MMPI.

At the level of individual assessment, however, the use of validity scales has been less fruitful. While such scales have made possible the detection of faking and the measurement of desirability response style, how to assess validly the personality of a person who is faking is a moot question. A common proposal has been to use an individual's validity scale score for statistical correction of his scores on personality scales. However, statistical correction is at best approximate (Jackson, 1967a; Meehl & Hathaway, 1946), and the corrected scores tend to be unreliable. Clearly, the problem of how to assess the personality of someone who intends to fake or is predisposed to responding desirably or undesirably has been only partially solved by validity scales.

Forced-choice Methods

The development of forced-choice methods has brought substantial advances in personality assessment. In an early review of forced-choice methodology, Travers (1951a) credited the basis of the idea of the forced-choice technique in personality assessment to Paul Horst and the development of the technique to R. J. Wherry. The most rapid period of development of forced-choice methodology was in the late 1940's when the United States Army applied the technique to officer performance rating (Zavala, 1965). Other early work with forced-choice methods was conducted by Baier (1951), Campbell and Rundquist (1950), Gordon (1951), Highland and Berkshire (1951), Jurgensen (1944), Kuder (1948), Mais (1951), Richardson (1949a, 1949b), Shipley, Gray, and Newbert (1946), Sisson (1948), and Travers (1951a).

The purpose of the forced-choice technique was to reduce bias in responses to items, whether in personality scales or performance-rating scales. As explained previously, self-descriptive statements are often nonneutral by the criterion of either DSV or item popularity. In fact, there is evidence that proportionately very few items are neutral (Cruse, 1965). The two indices of the favourableness of items are, of course, highly correlated (Edwards, 1953). In essence the forced-choice technique consists of pairing a self-descriptive statement pertaining to a personality trait with a trait-irrelevant filler statement having a very similar index of favourableness. The subject is asked to choose the statement which is more characteristic of himself.

Both DSV and item popularity have been used as the favourableness index used for matching purposes. Edwards (1954, 1957) has preferred matching items on DSV, while Jackson and Payne (1963) preferred matching items on item popularity. The rationale for the former is that subjects forced to choose between items matched on DSV cannot respond in terms of the desirability of items, and therefore, are more likely to respond to the content of items. The result should be a reduction in the influence of desirability response style, increased resistance to faking, and higher scale validity. The rationale for matching on item popularity also involves reducing the influence of response styles. Matching on item popularity has an added advantage. The expected popularity of each forced-choice item is .50, no matter how extreme the popularities of the original statements were. Consequently, the matching procedure should produce an increase in item and scale variance with a subsequent increase in scale reliability (see Magnusson, 1967, pp. 53-77). Because of the relatively high correlation between the two indices, however, the two methods of pairing items probably yield scales with very similar properties.

Research has indicated that forced-choice scales often do have higher reliabilities than their nonforced counterparts (e.g., Jackson & Payne, 1963). On the subject of validity, however, the research indicates that neither nonforced nor forced-choice items have a clear advantage, especially when subjects are instructed to fake (cf. Borislow, 1958; Izard & Rosenberg, 1958; Krug, 1958b; Longstaff & Jurgensen, 1953; Maher, 1959; Mais, 1951; Norman, 1963a; Rusmore, 1956; Waters & Wherry, 1962; Winters, Bartlett, & Leve, 1965). Furthermore, it has become very clear that matching statements on DSV does not prevent people from reliably judging one member of the forced-choice pair to be more desirable than the other (Corah, Feldman, Cohen, Gruen, Meadow, & Ringwall, 1958; Edwards, Wright, & Lunneborg, 1959; Feldman & Corah, 1960; Saltz, Reece, & Ager, 1962). Apparently, placing statements in the forced-choice context accentuates subtle differences in the desirability of items (see Corah et al., 1958; Feldman & Corah, 1960).

In addition to the research just cited, there is increasing recognition of the relevance of individual points of view about desirability to the forced-choice rationale. To the extent to which such individual conceptions of the desirable exist, on logical grounds forced-choice items cannot control for desirability bias

(La Pointe & Auclair, 1961; Messick, 1960; Norman, 1963a; Scott, 1963; N. Wiggins, 1966).

A further issue with forced-choice scales is economy. If each item of relevant content is paired with a construct-irrelevant filler item, then the number of statements an individual must read for a given set of forced-choice personality scales is double the number he would have to read were the items presented in a nonforced format. This is uneconomical. A frequently adopted solution to the problem has been to pair each construct-relevant statement with a statement representing another construct. This procedure is economical, for it does not increase the total number of statements an individual must read, and at the same time it requires half as many responses.

Unfortunately the economy of the latter procedure is accompanied by a statistical problem. The scales are now interdependent or ipsative, and their average intercorrelation must be negative. With ipsative scoring it is impossible for an individual to obtain a high score on every trait. What is measured is the intensity of each trait relative to the other traits--the subject's personality profile. The profile may be useful in individual assessment, where comparisons between traits are important, but most forms of correlational and factor-analytic treatment of ipsative scale scores are methodologically unsound (Broverman, 1962; Clemans, 1964; Horn & Cattell, 1965; Radcliffe, 1963, 1965; Stricker, 1965). Since correlational and factor-analytic treatment of the data were central to the present research, it was decided to sacrifice some economy in order to avoid using ipsative measures.

Judgmental Methods

Although empirical keying, validity scales, and forced-choice methods have each contributed substantially to our understanding of the faking and desirability response style problems, none has been more than a partial solution to the problems. All methods discussed to this point have had one thing in common: the subject was instructed to reveal something about himself by indicating whether or to what degree self-descriptive statements were characteristic of himself. It has been shown, however, that valid information about an individual can be obtained as an indirect consequence of requiring the subject to perform a task which does not involve self-report (Campbell, 1950). Such indirect measures have taken many forms (see e.g., Webb, Campbell, Schwartz, & Sechrest, 1966); but in keeping with the principle stated at the outset, only *structured* indirect measures, in this case judgmental methods, are reviewed. Instead of responding to self-report instructions, that is, endorsement instructions, the subject makes judgments about some quality of items, such as their desirability. Then, on the basis of his judgments, inferences are made about his personality or attitudes.

Judgmental methods of attitude assessment. Before discussing the research on judgmental methods in personality assessment, some research in attitude assessment is reviewed briefly because of its close resemblance to the work in personality assessment. The most significant attempt to date to measure attitudes by judgmental methods has been described in a series of studies concerned with attitudes towards racial segregation (Edrich, Selltiz, & Cook, 1966; Selltiz & Cook, 1966; Selltiz, Edrich, & Cook, 1965; Waly & Cook, 1965; Zavalloni & Cook, 1965). The approach used by Cook and his associates was to have subjects judge the plausibility of arguments for racial segregation and for racial integration, and then to compare their judgments with self-report measures of racial attitudes or with group membership implying such attitudes. The plausibility judgments were obtained after subjects were told that their ratings were needed for the construction of psychological tests. The investigators reported positive relationships between plausibility judgments and self-report measures of attitudes. In one study, the mean correlation between plausibility judgments and self-reports was .43 (Selltiz, Edrich, & Cook, 1965); in another the correlations ranged from .54 to .78 (Selltiz & Cook, 1966). The implication that these investigators drew from their findings was that, in principle, arguments for two sides of any issue could be constructed and plausibility judgments used as an indirect measure of attitude towards the issue.

Judgmental methods of personality assessment. There has been in recent years increasing recognition that there are individual consistencies in judgments about personality items. Messick (1960), for example, applied a vector model of multidimensional scaling to desirability judgments of items from the Edwards Personal Preference Schedule (Edwards, 1954) and identified nine dimensions or points of view of desirability. Other investigators have reported (Rosen, 1956; Scott, 1963; N. Wiggins, 1966) or reviewed (Damarin & Messick, 1965) research findings consistent with the multidimensional conception of desirability judgments about personality items.

Parallel to the recognition of individual consistencies in judgments, it has been hypothesized that such judgmental consistencies are predictive of the personalities of the judges. An incidental finding reported by Heineman (1953) lent early support to the hypothesis. Heineman used DSVs for matching statements in the construction of a forced-choice version of the Taylor Manifest Anxiety Scale (Taylor, 1953). When he classified subjects into high and low anxious groups according to their endorsement scores on the original Taylor Scale, he found high anxious subjects had rated the anxiety items as less undesirable than had the low anxious subjects.

Several more recent studies have sought to test the feasibility of using an individual's judgments about items to make valid inferences about his personality. Jackson (1961, 1964) instructed subjects to judge the desirability in other people of 45 personality items, responses to which were previously found to relate to a conformity criterion. The desirability judgments correlated positively with a nonquestionnaire measure of social conformity (r = .29). Loomis and Spilka (1963) failed to replicate Jackson's findings (r = -.15); but Goldberg and Rorer (1966), who, rather than instructing their subjects to judge the desirability of items, instructed them to judge whether their peers would endorse the items, were able to predict the social conformity criterion (r = .30). Stricker, Messick, and Jackson (1966, 1968) found that in a sample of 148 subjects, desirability judgments of personality items correlated with behavioural criteria to almost the same degree as did endorsements of a parallel set of items. However, when the data of subjects suspicious of the deception involved in the study were analyzed separately from the data of nonsuspecting subjects, for the suspicious subjects desirability judgments were more valid than endorsements, and for the nonsuspicious subjects the reverse was true. The authors speculated that the discrepancy between Loomis and Spilka's (1963) finding and the findings of the other two studies may have been due in part to a lack of control for suspicion of deception.

Kusyszyn (1968) contributed to the area by attempting to measure eight personality traits using judgment instructions. Fraternity brothers responded to the items from eight 20-item PRF scales (Jackson, 1967b) under four instructional sets, two of which were endorsements and desirability in others. Kusyszyn correlated the endorsements and desirability judgments with mean peer behaviour ratings of the same traits. Endorsements proved to be the most valid predictors of peer ratings (median r = .40; range = .35 to .71) and judgments of desirability in others were the next most valid (median r = .20; range = .01 to .36).

Some potential advantages of judgmental methods should be noted. Judgmental methods should be much less susceptible to faking since they may be presented plausibly to subjects as performance tasks. Judgmental measures could well be developed as parallel forms to the more direct endorsement measures for purposes of detecting personality or attitude change which may have resulted from treatment. Personality construct measures not involving endorsements would be a valuable contribution to the study of personality structure, since presumably they would share little of the method variance associated with self-report measures. The advantages of increasing the independence of methods are discussed more fully in a subsequent section on convergent and discriminant validity.

Several attempts to relate consistencies in desirability judgments about items to personality traits in the judges have been reviewed. In each study the items to be judged had been previously selected on the basis of the validity, not of judgments, but of endorsements of items. It may be that those items which yield the most valid endorsements are in some aspects different from those items which yield the most valid desirability judgments. If this is borne out, then the validity of desirability judgments for personality assessment, which at present varies widely (cf. Kusyszyn, 1968) could be substantially raised by basing the preliminary selection of items on desirability-judgment data instead of on endorsement data. In the present study, the items comprising the desirability-judgment scales were selected in a way designed to maximize the validity of the scales when administered with desirability-judgment instructions. Novel item-analysis procedures were utilized in selecting items for both endorsement and desirability-judgment scales.

Construct Validity--A Recent Development

The foregoing discussion has assumed the classical definition of validity. Validity in the classical sense is criterion-related; it is the correlation between scores on a personality scale and external measures of the same trait. Any references made above to the effects of faking and desirability response style on validity have pertained to the effects on scale-criterion correlations. The present section shows how the concept of validity in recent years has broadened in response to the rapid growth in psychological and psychometric knowledge.

Of the many advances in test theory in the past two decades, the most important is probably the introduction of the concept of construct validity. The introduction of the term construct validity by the American Psychological Association Committee on Psychological Tests (American Psychological Association, 1954) represented a major shift in emphasis from classical test theory. Classical test theory was primarily a theory of reliability, as exemplified by the fact that what is probably the most notable classical work, Gulliksen's (1950) _Theory of Mental Tests_, devoted only a few isolated sections to validity. Beginning with the American Psychological Association's _Technical Recommendations_ (1954), a number of significant articles and monographs have devoted the bulk of their discussion to validity rather than reliability (e.g., American Psychological Association, 1954, 1966; Campbell, 1960; Campbell & Fiske, 1959; Cronbach & Meehl, 1955; Loevinger, 1957).

Construct validity extends the classical conception of validity to subsume the questions of the extent to which the test measures the construct and the extent to which the construct is related to real traits in people (Cronbach & Meehl, 1955); or as Loevinger (1957) has expressed the questions, the extent to which the test measures a real trait in people and the adequacy of our interpretation of the test scores. In short, the question of construct validity is the question of not just how good is our measurement, narrowly defined, but how good is our psychology. An aspect of the latter is the psychology of structured test behaviour, a topic which includes the whole question of response styles and faking. The balance of this section is a review of some contributions to validity theory by Cronbach and Meehl (1955) and by Loevinger (1957).

The Logic of Construct Validity

Cronbach and Meehl's (1955) explication of construct validity was partly in answer to a characteristic of the psychological research of the 1940's and early 1950's. Many psychologists were involved with developing psychological tests, usually through empirical-keying methods, and validating such tests against a large number of specific criteria. A purely predictive model was in vogue; for any given test, a different validity would be reported for each criterion variable predicted by the test. Cronbach and Meehl argued that emphasis should be shifted from the *prediction* of specific criteria to the *measurement* of psychological variables that are part of psychological theory.

Central to Cronbach and Meehl's argument was their concept of nomological networks. The following is quoted from their definition (Cronbach & Meehl, 1955).

> 1. Scientifically speaking, to "make clear what something *is*" means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a *nomological network*.
>
> 2. The laws in a nomological network may relate (a) observable properties or quantities to each other; or (b) theoretical constructs to observables; or (c) different theoretical constructs to one another. These "laws" may be statistical or deterministic.
>
> 3. A necessary condition for a construct to be scientifically admissible is that it occur in a nomological net, at least *some* of whose laws involve observables [p. 290].

The construct validation of a psychological test, according to Cronbach and Meehl, involves relating test scores to constructs within a nomological network in order to determine the psychological significance of the test scores and, simultaneously, the adequacy of the theory in accounting for the data. If a psychological theory relates a construct to a particular behavioural variable, and measures of the behavioural variable do not correlate appropriately with scores from a test purported to measure the construct, then, quoting from Cronbach and Meehl (1955), the negative finding can be interpreted in the following ways.

> 1. The test does not measure the construct variable.
>
> 2. The theoretical network which generated the hypothesis is incorrect.
>
> 3. The experimental design failed to test the hypothesis properly [p. 295].

Positive results, on the other hand, constitute evidence for both the construct validity of the test and the adequacy of the theory. Such a conclusion from positive results, it should be cautioned, assumes the existence of no evidence from other sources that the test lacks construct validity. For example, the existence of high correlations between scores on the test and measures of desirability response style would suggest the need for redefining the construct or incorporating a desirability response style construct into the nomological network. Probably both would be advisable.

The Components of Construct Validity

Loevinger's (1957) monograph, "Psychological Tests as Instruments of Psychological Theory," was a milestone in the history of psychology and psychometrics. Her explication of construct validity was more radical than that of Cronbach and Meehl's (1955). She argued that the starting point for construct validation is psychological theory. For a test to achieve construct validity, even such aspects as its method of construction should stem from psychological theory. Loevinger argued that construct validity has three components, the substantive component, the structural component, and the external component. As the following discussion shows, Loevinger stressed that if a psychological test is to contribute anything to psychological theory, it is essential that the test achieve all three components of construct validity.

The substantive component. Loevinger (1957) defined the substantive component of construct validity as "the extent to which the content of the items included in (and excluded from?) the test can be accounted for in terms of the trait believed to be measured and the context of measurement [p. 661]." In her definition, she used the term "context" to include psychological theory, especially the psychology of structured test behaviour. She used the term "trait" in the sense of an individual-difference variable which really exists in people, its counterpart in psychological theory being, of course, the construct.

Loevinger argued that the most critical stage in securing the substantive validity of a test is the preparation of the item pool. A construct, or theory of a trait, is spelled out in terms of its content so that a pool of items embodying this content can be prepared. However, Loevinger proposed that items should be drawn not only from the content specified by the construct in question, but also from the content specified by all known alternative theories of the trait. She reasoned that only in this way could the scale derived from the item pool contribute to the comparative evaluation of the alternative constructs which putatively explain the trait in question.

The structural component. The structural component of construct validity "refers to the extent to which structural relations between test items parallel the structural relations of other manifestations of the trait being measured [Loevinger, 1957, p. 661]." Structural validity includes both fidelity of structure and degree of structure. A scale is said to have fidelity of structure when its items are intercorrelated to about the same degree as are nontest manifestations of the trait. If behavioural manifestations of a trait are highly intercorrelated, for example, the items in the scale should also be highly intercorrelated. The degree of structure refers to the degree of interrelatedness of the items. Assuming a scale exhibits fidelity of structure, it is said to have high structural validity, for example, if its items are highly intercorrelated.

Structural validity is secured for a scale through item selection from the substantively-valid item pool. The method of item selection depends on the structural model adopted. For example, the structural model of classical test theory holds that the degree of the trait possessed by an individual is a monotonic increasing function of the number of manifestations of the trait. If the classical model is deemed appropriate, as was the case in the present investigation, then after administering the item pool to a large sample of people, the data of each subject are scored by simply counting the number of keyed responses. Furthermore, if it is correct to assume that a high degree of interitem structure is required for structural fidelity, then item selection based on the magnitudes of correlations between item responses and total item-pool scores produces a scale with high structural validity.

The external component. The third component of construct validity is akin to classical criterion-oriented validity. A scale must be demonstrated to correlate substantially with external criteria that are theoretically related to the trait the scale is purported to measure. Loevinger (1957) stressed that without this empirical confirmation of validity, a scale cannot be expected to demonstrate any practical usefulness. Assuming substantive and structural validity have been secured, if a scale achieves a satisfactory level of external validity, it may not only exhibit practical utility, but may also be used to make valuable contributions to psychological theory. The latter claim may be restated more concretely. A fully construct-validated scale may be used experimentally in testing hypotheses derived from psychological theories which embody the corresponding construct.

In her treatment of the external component of construct validity, Loevinger (1957) insisted that for a scale to have construct validity it must demonstrate, not only substantial correlations with relevant criteria, but also negligible correlations with theoretically irrelevant variables. For example, in Loevinger's view an impulsivity scale should correlate highly with other independent measures of impulsivity, but should not correlate appreciably with known forms of distortion such as desirability response style or with valid measures of theoretically independent constructs. Loevinger's expectations regarding external validity were somewhat overstated; negligible correlations with irrelevant variables are too much to expect from any type of scale. Nevertheless, her treatment of external validity should be recognized as a precursor to Campbell and Fiske's (1959) important exposition of convergent and discriminant validation. The latter is reviewed below.

Discriminant Validity--An Adjunct to Construct Validity

The recognition by those concerned with personality assessment devices that a measurement device should have discriminant validity as well as convergent validity has been slow in coming. Since the first thorough presentation of the subject (Campbell & Fiske, 1959) was antedated by the major expositions of construct validity (i.e., Cronbach & Meehl, 1955; Loevinger, 1957), the latter concentrated on convergent validity. The intent of the present section is, first, to clarify some terminology, and then, to review techniques for attaining and assessing discriminant validity. The review of the methodology is organized according to the substantive, structural, and external components of construct validity (cf. Loevinger, 1957).

Terminology

Convergent validity is the typically-used kind of criterion-oriented validity often referred to as predictive, concurrent, empirical, or external validity. The term 'convergent' validity denotes the degree to which two measures yielded by maximally independent methods of assessing a trait tend to converge on that trait. Convergent validity is distinguished from 'reliability;' reliability is the correlation between two measures yielded by the same method of assessing the trait, rather than by independent methods. Convergent validity is also distinguished from 'discriminant' validity, in that the latter refers to the extent to which a scale is uncorrelated with theoretically irrelevant variables; for example, a novel measure of a trait would be invalidated if it is correlated too highly with established measures of other psychologically unrelated traits. For establishing the construct validity of a scale or for justifying a novel way of measuring a trait, both discriminant and convergent validity are essential.

The Substantive Component and Discriminant Validity

The discriminant validity of scales to be developed for measuring various constructs in a nomological network is affected by even the initial delineation of the constructs. At this early stage prior to scale development, care should be taken to assure a maximum degree of conceptual independence of the constructs (Jackson, 1966a). If a theory embodies two highly related constructs, and scales for measuring them are constructed, the two scales will be probably too highly correlated to be independently useful, but not highly enough correlated to be combined without sacrificing structural validity. In such a case the discriminant validity of scales to be developed is improved by revising the theory either by making the conceptually overlapping constructs more distinct or by replacing them with an intermediate construct.

Preparation of the item pool for each construct may proceed when conceptual independence of the constructs has been established. The relationship between the constructs and the item content has already been discussed from the point of view of establishing convergent validity. The discriminant validity of the proposed scales can be improved by avoiding the assignment of highly similar items to item pools for different constructs. The omission of a careful review of the item pools for substantive independence can result in spuriously high scale intercorrelations. The failure of most MMPI scales in achieving satisfactory levels of discriminant validity, for example, has been at least partially due to the rampant item overlap between scales; item overlap is an extreme case of lack of substantive independence.

The Structural Component and Discriminant Validity

As explained earlier, after administering the item pool to a large sample of people, structural validity is gained through item selection based on the analysis of the item data. The item-selection technique adopted may also contribute to discriminant validity. During the selection of items for a scale, the discriminant validity of the scale is facilitated by excluding, for example, items which correlate too highly with item pools for other constructs or with noncontent variables such as desirability response style. When test-development programs have stressed such techniques for improving discriminant validity, the resulting scales have demonstrated impressive evidence of discriminant and convergent validity (e.g., Jackson, 1967b). Of course, an improvement in construct validity is concomitant with any improvement in discriminant validity.

The External Component and Discriminant Validity

After scales have been developed to measure particular constructs, they must be shown to possess convergent and discriminant validity. Methods for assessing both aspects of the external component of construct validity are presented below. The multitrait-multimethod correlation matrix is the starting point for most methods of assessing discriminant and convergent validity. Therefore, the matrix itself is described before methods are presented for analyzing it.

The multitrait-multimethod matrix. A multitrait-multimethod matrix "presents all of the intercorrelations resulting when each of several traits is measured by each of several methods [Campbell & Fiske, 1959, p. 81]." Each combination of a method of measurement and a trait to be measured is a trait-method unit. The measurement aspects of a scale and the content of the scale can each account for part of the systematic variance of the scale scores. The fact that an individual's response is influenced by both construct-relevant content determinants and construct-irrelevant method determinants draws attention to the importance of including maximally independent methods in the validation procedure. The less similar are two methods of measurement, the lower will be the level of their correlated method variance. Furthermore, the greater the number of independent methods that are brought to bear on the measurement of a trait, the more evidence there will be for or against the construct validity of each trait-method unit. There is an obvious corollary to requiring the inclusion of several methods in the matrix; namely, several traits should also be included. The greater the number of traits that are represented in the matrix, the more evidence there will be for or against the discriminant validity of each trait-method unit. It should be apparent by now that discriminant and convergent validity cannot in any practical way be separated from one another. Both are integral parts of the external component of construct validity.

For purposes of analysis the multitrait-multimethod matrix is usually organized in a special manner. The variables are arranged so that all trait-method units belonging to one type of method are adjacent to each other in the matrix. In this way the correlation matrix is composed of a number of monomethod and heteromethod submatrices, each one containing intercorrelations among all traits under consideration. The traits are arranged in the same order within each submatrix so that all minor-diagonal elements are convergent validity coefficients; hence the minor-diagonal is often termed the 'validity diagonal' and its elements are referred to as monotrait-heteromethod correlations or convergent validities.

Informal procedures. Campbell and Fiske (1959) recommended that three principles be applied in the evaluation of convergent and discriminant validity in multitrait-multimethod matrices. The first principle is concerned with convergent validity and the other two with discriminant validity. First, the validity-diagonal elements should be statistically significant as the minimum requirement for convergent validity. Secondly, each validity-diagonal element should be greater than the off-diagonal correlations in the same row and column within the heteromethod submatrix. Thirdly, each validity-diagonal element should be greater than the correlations involving the same trait in either of the corresponding monomethod submatrices.

Campbell and Fiske's (1959) informal procedure was an interim solution to the problem of assessing the level of discriminant and convergent validity in a matrix as a whole, for their method has presented problems. For example, a simple count of the number of validity-diagonal elements which exceeded all corresponding off-diagonal elements (second principle) can be interpreted only in terms of the size of the matrix. Conversion of the number to a proportion of the total number of validity-diagonal elements is at best a partial answer, because the proportion likely to be attained is affected by the ratio of the number of traits to the number of methods. The most stringent test of discriminant and convergent validity can be made by measuring similar traits by highly dissimilar methods. It follows, therefore, that in using the above informal procedure a set of tests might appear to have convergent and discriminant validity simply because only highly similar measures of highly dissimilar traits were included in the matrix. Furthermore, the use of informal procedures is a very imprecise method for comparing the potency of constructs across methods of measurement or for making a comparative evaluation of the methods used. A more precise analytic procedure is required for answering such questions.

Analytic procedures. The analytic procedures that have been applied to multitrait-multimethod matrices for purposes of separating the effects of traits from the effects of methods have all been based on factor analysis. Ordinary principal components factor analysis has proven unsatisfactory in that it yields, not only trait factors, but also method factors and factors reflecting interactions between traits and methods. Very recently, Jöreskog (1968) and Boruch and Wolins (1968; Boruch, 1968) have presented factor-analytic methods which incorporate restraints aimed at separating method factors from trait factors. However, while their success with trial data appears promising, the methods have not yet been applied to substantive problems. Jackson (1966b) used a rationale similar to that for interbattery factor analysis (Kristof, 1967; Tucker, 1958) in his development of multimethod factor analysis. In Jackson's technique the monomethod submatrices of the multitrait-multimethod matrix are orthogonalized (see Horst, 1965, pp. 566-576) by replacing them with identity matrices, as, for example, is shown in Table 1. Identity matrices have unities for diagonal elements and zeros for off-diagonal elements. The multitrait-multimethod matrix with orthogonalized monomethod submatrices is then subjected to principal components factor analysis. Since the analysis is based on only variance associated with two or more methods, the factor solution is not affected by variance common to only single methods. Multimethod factor analysis has proved to be a viable technique for assessing convergent and discriminant validity in multitrait-multimethod matrices (e.g., Jackson, 1967b, Kusyszyn & Jackson, 1968; Siess & Jackson, 1967). Consequently, it was employed in the present investigation.

Table 1

Orthogonalization of Monomethod Submatrices

in Multimethod Factor Analysis

Method		1			2			3			4		
	Trait	A	B	C	A	B	C	A	B	C	A	B	C
1 True-false	A												
	B		I			R_{12}			R_{13}			R_{14}	
	C												
2 Forced-choice	A												
	B		R_{21}			I			R_{23}			R_{24}	
	C												
3 Self Ratings	A												
	B		R_{31}			R_{32}			I			R_{34}	
	C												
4 Peer Ratings	A												
	B		R_{41}			R_{42}			R_{43}			I	
	C												

Note.-- Identity matrices (I) have been substituted for the diagonal, monomethod submatrices.

The Nature of the Present Investigation

The topic of the chapters which follow is the development and evaluation of a number of experimental scales for measuring the personality traits of Impulsivity, Risk Taking, and Self Esteem. Scales for measuring more than three traits were not developed for reasons of economy; even with only three traits subjects had to contribute either two or four hours of their time. The basic reason for selecting the above three traits, as opposed to other possible traits, was that the three traits differ widely in their apparent desirability; Self Esteem is a very desirable attribute in people, Impulsivity is somewhat undesirable, and Risk Taking probably lies somewhere between the other two. The desirability of the three traits was relevant to the control of desirability response style, an important aspect of the research.

The program of research was not conceived as an attempt to develop a set of ideal scales. Rather, the research involved, first, the development of experimental scales which differed from one another in terms of the methods of measurement involved; and secondly, it involved the comparative evaluation of the different methods under two experimental conditions. One experimental condition employed standard research instructions for responding honestly to all tasks. In the other experimental condition subjects were instructed to fake; more specifically, they were instructed to try to create a favourable impression by the way they answered.

The four types of personality scales developed for comparative evaluation were identified as nonforced endorsement scales, forced-choice endorsement scales, nonforced desirability-judgment scales, and forced-choice desirability-judgment scales. Of the four, only the last has not yet been defined. It was a novel approach which required subjects to select from pairs of statements the statement they regarded as more desirable in other people.

For most of the possible comparisons that are made among methods, traits, and experimental conditions, there was, with one exception, insufficient theoretical or empirical reason for formulating specific hypotheses. The exception was the hypothesis put forward in the review of judgmental methods that such methods are more resistant to faking than are endorsement methods. It was expected, therefore, that desirability-judgment scales would have as high a level of convergent validity with instructions to fake as with standard instructions, but that endorsement scales would not. Of more significance than this specific prediction, however, was the overriding expectation that close attention to considerations of construct validity from the outset would yield important additional insights into the nature and control of desirability response style and faking.

CHAPTER II

DEVELOPMENT OF THE PERSONALITY SCALES

The purpose of the present chapter is to describe the development of experimental scales for measuring Impulsivity, Risk Taking, and Self Esteem. Item pools were prepared and administered to a large sample of subjects, with instructions both to endorse items and to judge the desirability of items. Item-analysis procedures were applied to the endorsement and desirability-judgment data in ways designed to select the subsets of items best representing each construct.

Item Pools

It has been emphasized that, if a scale intended for measuring a personality trait is to achieve a high level of construct validity, the item pool from which items are selected must faithfully represent the content assumed to be related to the trait (Loevinger, 1957). The preparation of items should be based on a carefully prepared broad definition of the personality construct. The item pools used in the present investigation were borrowed from two sources. The Impulsivity item pool was that on which was based the development of the PRF Impulsivity scale (Jackson, 1967b). Jackson's Experimental Personality

Inventory, a major personality questionnaire now under construction, provided the Risk Taking and Self Esteem item pools. The three item pools were considered acceptable for present use because they had been constructed according to the principles expressed with regard to representativeness of content. With particular attention to the definitions of the constructs, at least 100 items had been prepared for each pool so as to represent all areas of content stated or implied in the definition of the construct.

One-half of the items in each pool were negative exemplars of the construct, that is, they were phrased to represent the opposite end of the personality dimension in question. In addition to providing more adequate definition for the ends of each continuum, balancing the number of positively-keyed and negatively-keyed items in a pool reduces distortion due to certain kinds of response bias. When items are presented in a true-false format, for example, some people tend consistently to answer 'true,' while others tend consistently to answer 'false' (Jackson, 1967a); or when items are presented in a rating-scale format, where subjects indicate how self-characteristic each item is by rating it on, for example, a 9-point scale, some people tend to give more extreme ratings than do others (Cronbach, 1958). In the past, where the number of positive and negative exemplars have not been balanced, these response tendencies have accounted for a major portion of the variance of total scale scores (Jackson, 1960; Jackson & Messick, 1958, 1961, 1962; Messick, 1967; Peabody, 1966; Stricker, 1963).

The 144-item Impulsivity item pool, the 100-item Risk Taking item pool, and the 124-item Self Esteem item pool were assembled in booklet form in a partially random order with 240 other items. Of the latter items 40 comprised the PRF-Dy scales (Jackson, 1967b) described earlier. The remaining 200 items of heterogeneous content were selected from the item pools of the PRF and were included as filler items. Selection of the 200 filler items was random, except that no Impulsivity items, PRF-Dy scale items, or items with obvious relevance to Risk Taking or Self Esteem were selected. The filler items, which were required in the construction of the forced-choice scales, were included in the questionnaire in order to obtain the item statistics required for matching items.

Subjects

The subjects were 246 female volunteers from introductory psychology classes. Two subjects who attended the first session failed to return for the second, leaving a usable item-analysis sample of 244 subjects. All subjects received research participation credit for their time. No men were used in the first phase of the research because only women were available for the second phase. The validation study required subjects who knew one another well enough to provide reliable behaviour judgments about one another. This requirement was fulfilled with a group of student nurses living in residence.

Procedure

Each subject participated in four hours of testing, two hours one week and two hours a week later. The testing was conducted in large groups and was supervised by a minimum of one proctor to 40 subjects. At the beginning of each session subjects were given printed instruction sheets and the head proctor read the instructions to the group. Subjects recorded their responses on separate answer sheets entitled 'Record of Judgments.'

During the first session subjects were presented with the endorsement instructions. Each subject judged the items in the booklet as to how characteristic or uncharacteristic each item was of herself. Ratings were made on a 9-point scale which ranged from a rating of 1, indicating 'extremely uncharacteristic of self,' to a rating of 9, indicating 'extremely characteristic of self.'

During the second session a week later subjects were presented with the desirability-judgment instructions. This time each subject judged the same set of items as to how desirable or undesirable each item would be in other women if it were true of other women. Again, ratings were made on a 9-point scale. This time a 1 indicated 'extremely undesirable in other women' and a 9, 'extremely desirable in other women.' The complete instructions may be found in Appendix A.

Rationale for sequence adopted. Separate data-gathering sessions were held for endorsements and desirability judgments in order to minimize the probability of carry-over from one task to the other. Scheduling efficiency was the major reason for fixing the interval at one week. The endorsement instructions were consistently presented before the desirability-judgment instructions because of evidence for considerable carry-over from desirability judgments to endorsements (Jackson & Messick, 1967, personal communication; cf. Messick, 1965), but little or no evidence for the reverse. If the study had not been dependent upon minimizing carry-over effects so that comparisons between methods could be made, a procedure allowing the experimental comparison of different orders of presentation would have been adopted. For similar reasons, the same order of presentation was adopted in conjunction with the validation of the experimental scales against behaviour-judgment measures of the traits.

An added control. Whenever subjects are tested in large groups there is some loss of control over the performance of individual group members. Possibly a few individuals will respond carelessly in their attempts to complete the task quickly and leave early. This adds unreliability to the results. In the present context the incentive of leaving early was reduced by adding a filler task. It was announced that upon completion of the questionnaire there would be a second task 'that is not as important as the first task, but is still important.' Subjects were told that no one would finish in less than two hours, and therefore, care should be taken to do the first task well. When a subject had completed rating the items, her materials were collected and she was given a blank piece of paper and a booklet entitled 'General Reasoning Test.' The latter consisted of 23 arithmetic reasoning questions with instructions to answer every question and to show all work. The proctor recorded the starting time at the top of the blank paper to add authenticity to the task. At the end of the two-hour session subjects working on the General Reasoning Test were told they could complete it at the next session. In the second session, following completion of their desirability judgments of items, subjects were told to continue with the General Reasoning Test from where they left off the previous week. Very few subjects managed to complete the filler task by the end of the second session.

Data Reduction

Upon completion of the endorsement and desirability-judgment data gathering, a number of experimental personality scales were constructed by basing item selection on multistage item-analysis strategies. The purpose of the item selection was to prepare for each construct concise scales with both substantive and structural validity. Steps were taken during the construction of each scale both to maximize the variance attributable to the construct and to minimize the irrelevant or unreliable variance.

Before any item analysis could be conducted, a score for each subject had to be computed. The first such score to be computed was the subject's total endorsement score for each item pool. Following the structural model of classical test theory, the score used was the sum of the subject's keyed responses obtained under endorsement instructions. First, the subject's endorsement ratings of positively-keyed items, that is, her ratings of positive exemplars of the construct, were summed to get a part score for that item pool. Next her rating of each negatively-keyed item was subtracted from 10, thereby reversing the direction of the rating. Finally, the reversed ratings of the negatively-keyed items were summed with the part score computed for the positively-keyed items, yielding the subject's total score for the item pool. The scoring procedure was the same for each item pool and for the 40-item PRF-Dy scale. The product of the scoring was an endorsement score for Impulsivity, Risk Taking, Self

Esteem, and desirability response style for each of 244 subjects in the item-analysis sample.

The Pearson product moment correlations between the item-pool endorsement scores and the PRF-Dy scale scores were computed. As expected, the Impulsivity item-pool endorsement scores correlated negatively with the PRF-Dy scale scores ($r = -.36$), and the Self Esteem endorsement scores correlated positively ($r = .55$). Risk Taking was negligibly correlated with the PRF-Dy scale ($r = -.12$). With correlations of moderate size between the PRF-Dy scale and two of the three item pools, scales derived from these item pools also might correlate too highly with the PRF-Dy scale. Therefore, the item-pool endorsement scores were altered so as to reduce the degree to which each reflected desirability variance. The item-pool endorsement scores were corrected for desirability bias by means of the following regression formula (cf. Gulliksen, 1950, p. 425; Peters and VanVoorhis, 1940, pp. 110-112).

$$T_c = T_u - r_{T_u Dy} \frac{S_{T_u}}{S_{Dy}} (Dy - \overline{Dy})$$

where T_c is the corrected endorsement score for an item pool,

T_u is the uncorrected endorsement score for an item pool,

S_{T_u} is the standard deviation of the uncorrected endorsement scores,

Dy is the PRF-Dy scale score,

S_{Dy} is the standard deviation of the PRF-Dy scale scores,

$\overline{Dy}$ is the mean of the PRF-Dy scale scores, and

r_{T_uDy} is the correlation between the uncorrected endorsement scores and the PRF-Dy scale scores.

It can be seen that the magnitude of the correction of an individual's endorsement score for a particular item pool was a function of both the deviation of her PRF-Dy scale score from the mean of the PRF-Dy scale scores and the correlation between the item-pool endorsement scores and the PRF-Dy scale scores. For the item-analysis sample as a whole, the effect of regressing out the PRF-Dy scale scores was to make the corrected endorsement scores for each trait correlate zero with the PRF-Dy scale scores.

As explained in the introduction, the structural model of classical test theory, the model adopted in the present investigation, requires item-selection indices which emphasize considerations of internal consistency. In the present research, both the uncorrected and corrected endorsement scores described above were used in the calculation of the item-total correlations. The endorsement and desirability-judgment ratings of each item in the three item pools were correlated with the uncorrected scores for Impulsivity, Risk Taking, and Self Esteem. In addition the endorsement and desirability ratings of each item comprising a particular item pool were correlated with the corrected endorsement scores for that item pool. Subsequent item selection was based on these item correlations. Since the item-selection strategy for the endorsement scales differed from that for the desirability-judgment scales, the two strategies are described separately in the two sections which follow.

Item Selection for the Endorsement Scales

Jackson and Messick (Jackson, 1967, personal communication; cf. Jackson, 1967b) developed what they called their Differential Reliability Index (DRI). They used it in item selection because it took into account the item's correlation with a desirability response style measure as well as the item's correlation with total scores. In terms of a reduction in desirability bias, the DRI proved quite beneficial in item selection for the PRF (Jackson, 1967b). In the present research, item selection for endorsement was based on the following variation of Jackson and Messick's DRI.

$$DRI_1 = \sqrt{r_{iT_c}^2 - (\sum r_{iT_{ug}}^2 + r_{iDy}^2)}$$

where $r_{iT_c}^2$ is the square of the correlation between the endorsements of an item and the corrected endorsement scores for the corresponding item pool,

$\sum r_{iT_{ug}}^2$ is the sum of the squared correlations between the endorsements of the item and the uncorrected endorsement scores for the other two item pools, and

r_{iDy}^2 is the square of the correlation between the endorsements of the item and the PRF-Dy scale scores.

The foregoing index may be viewed as the square root of the difference between two variance components, the difference being the reliable variance of the item. The advantage of using such an index for item-selection purposes is that it aids the selection of items with the maximum trait saturation relative to their saturation with variance due to other traits, including desirability response style. It differs from Jackson and Messick's DRI in that theirs took into account variance due only to desirability response style.

The DRI_1 served as the major criterion in selecting items for endorsement scales. The positively-keyed and negatively-keyed items of each item pool were ranked separately according to the magnitudes of their DRI_1s. The 10 best items from each subset were retained tentatively for the endorsement scales; for example, of the positively-keyed Impulsivity items, the 10 having DRI_1s of greatest magnitude were selected to form the positively-keyed half of the Impulsivity endorsement scale.

The three 20-item sets of the highest-ranking items were then each subjected to a critical review by two or more editors to assure that the items were still broadly representative of the construct and its item pool,and that item selection had not resulted in any narrowing of content. An example of narrowing of content occurred in item selection for the Risk Taking scale. Four of the 20 initially-selected items dealt with gambling behaviour. Although the four items were unquestionably a part of the content universe of Risk Taking, four gambling items was considered a disproportionate number for a 20-item scale. Therefore, two of them were discarded and each was replaced by the first content-representative item from the appropriate ranked subset of unselected items. Only one to three items per scale required replacement.

Item Selection for the Desirability-judgment Scales

Like the item selection for the endorsement scales, the item selection for the desirability-judgment scales was based on item indices computed from correlations between responses to items and item-pool endorsement scores. The major difference was that the item responses used in computing the correlations were the judgments of the desirability of the items instead of the endorsements of the items. The reason for basing the item indices on correlations between desirability-judgments of items and total scores computed from endorsements of items was to maximize the correlations between the endorsement and desirability-judgment scales resulting from item selection.

The use of a single index for selecting items for the desirability-judgment scales was precluded by the fact that the correlations between desirability-judgments of items and the endorsement totals tended to be moderately low and not as highly discriminating between the appropriate and inappropriate item-pool endorsement scores as were the corresponding endorsements of items. An index exactly parallel to that employed in the development of the endorsement scales probably would have capitalized on chance factors affecting the correlations, thereby reducing reliability. Instead of employing a single index, therefore, item selection was completed in two stages. The first stage was designed to exclude items whose desirability judgments correlated too highly with the PRF-Dy scale scores relative to the correlations with the appropriate item-pool endorsement scores. The second stage was designed to select from the remaining items in each pool those items whose desirability judgments were most highly correlated with the endorsement scores of the construct-relevant item pool, relative to the endorsement scores of the other two item pools.

In order to exclude items whose desirability judgments were too highly associated with scores on the PRF-Dy scale, which is an endorsement scale, the following index was computed for each item.

$$DRI_2 = \sqrt{r_{iT_c}^2 - r_{iDy}^2}$$

where $r_{iT_c}^2$ is the square of the correlation between desirability judgments of an item and the corrected endorsement scores of the corresponding item pool, and

r_{iDy}^2 is the square of the correlation between the desirability judgments of the item and the PRF-Dy scale scores.

As with the endorsement indices the positively-keyed and negatively-keyed items of each item pool were ranked separately according to the magnitude of their DRI_2s. In each set of positively- or negatively-keyed items, the best 20 were retained for further selection. Of the 20 negatively-keyed Risk Taking items, for example, the 20 items having indices of greatest magnitude were selected to form the negatively-keyed half of a provisional Risk Taking desirability-judgment scale.

Each subset of 20 items was reduced to 10 items by ranking the items on a further index, this one designed to select the items whose desirability judgments best discriminated among the three constructs.

$$DRI_3 = \sqrt{r_{iT_u}^2 - \sum r_{iT_{ug}}^2}$$

where $r_{iT_u}^2$ is the square of the correlation between the desirability judgments of an item and the uncorrected endorsement scores for the corresponding item pool, and

$\sum r_{iT_{ug}}^2$ is the sum of the squared correlations between the desirability judgments of the item and the uncorrected endorsement scores for the other two item pools.

As with the construction of the endorsement scales, the resulting 20-item desirability-judgment scales were subjected to a careful editorial review to assure content representativeness. Item replacements were made where necessary.

The foregoing procedure could be applied successfully only to the Impulsivity and Risk Taking items. Item selection for a Self Esteem desirability-judgment scale was hampered by the fact that correlations with the endorsement scores for the Self Esteem item pool were so low that 20 promising items could not be found. In fact, for Self Esteem items the highest correlation between desirability judgments and item-pool endorsement scores was only .20. The reason for the low correlations is probably related to the fact that Self Esteem is a very desirable trait. There is some evidence that desirability judgments of desirable items are less highly correlated with endorsements than are items neutral with respect to desirability (Messick, 1964). It was decided, therefore, to abandon plans for the development and validation of a Self Esteem desirability-judgment scale.

Construction of the Forced-choice Scales

Forced-choice versions of the three 20-item endorsement scales and two 20-item desirability-judgment scales were prepared by pairing each item with an item from among the 200 filler items which had been included in the booklets administered to subjects. Responses to the 200 filler items were correlated with the endorsement scores from the three item pools. Any item correlating as high as .20 with any of the three item pools was considered unacceptable for pairing purposes. Each item in an endorsement scale was paired with a filler item closely resembling it in mean endorsement rating. Similarly, each item in a desirability-judgment scale was matched with a filler item on mean desirability judgment. If two or more filler items were equally good with regard to their mean ratings, the pairing was based on similarity of item variances. If neither mean nor variance could lead to the choice of a particular filler item, the filler item with the lowest average correlation with the three item pools was chosen. In most cases the item means within a pair differed from each other by no more than .01 points on the 9-point scale.

CHAPTER III

PRELIMINARY EVALUATION OF THE PERSONALITY SCALES

The construction of 10 experimental personality scales has been described to this point. Five of the scales utilized a single-statement or nonforced item format, while the other five utilized a two-statement or forced-choice item format. A nonforced endorsement scale and a forced-choice endorsement scale were prepared for each of the three constructs, while for reasons specified in Chapter I neither type of desirability-judgment scale was prepared for measuring Self Esteem. The purpose of developing the scales was to make possible a comparative evaluation of the different kinds of scales in terms of their construct validity and resistance to desirability response style and faking.

In the following section the reanalysis of the item-analysis sample data is described. The reanalysis was done to gain some preliminary information with regard to the effectiveness of the item-selection strategies in securing structural validity, in reducing the influence of desirability response style, and in increasing the independence of the measures. Strictly speaking, the reliabilities, the correlations with the PRF-Dy scale, and the scale intercorrelations reported below must be viewed with caution since they were based on the item-analysis sample (cf. Cureton, 1950). The reported shifts in correlations may have been partially due to capitalization on chance factors. Therefore, although the results of the reanalysis are presented, few conclusions may be drawn before the readministration of the scales to a new sample of subjects. The preliminary results are compared with the results of the cross-validation in Chapter V.

Reanalysis of the Item-analysis Sample Data

Only the five nonforced endorsement scales were considered in the reanalysis, since the forced-choice scales had not been administered to the item-analysis sample. Scoring keys were developed and the endorsement and desirability-judgment data were rescored. Taking into account the direction of keying of items, the sum of the endorsement ratings of the items in each of the three endorsement scales was computed for each subject. Similarly, the sum of the desirability judgments of the items in each of the two desirability-judgment scales was computed for each subject. The uncorrected endorsement scores for the three item pools and the PRF-Dy scale were available from the original analysis. The intercorrelations among the above scores were computed.

In addition to computing the total scores for each subject, each of the foregoing sets of items was divided into odd and even subsets for the purpose of examining scale reliability. Alternate items within each set were assigned to odd and even subsets with the provision that each subset should contain an equal number of positively-keyed and negatively-keyed items. The correlations between corresponding odd and even subsets of items were computed and corrected using the Spearman-Brown prophecy formula.

Results of the Reanalysis

Reliability and content saturation. The corrected split-half reliability coefficients of the item pools and experimental scales are presented as the major diagonal elements of the correlation matrix in Table 2. The reliabilities of the item pools were very high, ranging from .94 to .97. The attainment of such high reliability confirms the conclusion from other research (Jackson, 1966a, 1967b; Neill & Jackson, 1967) that it is possible to prepare from substantive considerations a large internally consistent pool of items to represent a dimension of personality. The reliabilities of the 20-item endorsement scales ranged from .81 to .94, only slightly lower than those of the corresponding item pools. The reliabilities of the two desirability-judgment scales were .80 and .88.

With such a high degree of internal consistency in the item pools, high correlations should be expected between the item pools and the corresponding 20-item scales. As may be seen in Table 2, the correlations with the appropriate item pools were .91, .92, and .94 for the endorsement scales and .49 and .58 for the desirability-judgment scales. The correlations for endorsement scales were expected to be higher than those for the desirability-judgment scales, because the former were part-whole correlations. It may be concluded that the scale development strategies produced scales possessing very satisfactory levels of structural validity.

Table 2

Correlations among the Three Item Pools, the Five Nonforced Experimental Personality Scales, and the PRF-Dy Scale

(N = 244)

Method	Trait		Item Pools Endorsements			20-item Scales Endorsements			20-item Scales Desirability Judgments	
			Im	RT	SE	Im	RT	SE	Im	RT
Item Pools Endorsements	Impulsivity	Im	(94)							
	Risk Taking	RT	69	(95)						
	Self Esteem	SE	17	33	(97)					
20-item Scales Endorsements	Impulsivity	Im	91	55	21	(81)				
	Risk Taking	RT	55	92	28	42	(90)			
	Self Esteem	SE	21	29	94	24	24	(94)		
20-item Scales Desirability Judgments	Impulsivity	Im	49	26	07	47	23	07	(80)	
	Risk Taking	RT	38	58	07	28	59	05	48	(88)
Endorsements	PRF-Dy		-36	-12	57	-25	-05	42	-16	-13

Note.-- The major diagonal contains corrected split-half reliabilities. The .05 and .01 significance levels of r are .14 and .18, respectively. Decimals have been omitted.

Correlation with the PRF-Dy scale. One of the functions of the scale-construction procedure was to reduce the influence of desirability response style. The last row of the matrix in Table 2 contains the correlations with the PRF-Dy scale. A comparison between the item pools and the derived endorsement scales reveals that for Impulsivity the correlation with the PRF-Dy scale was reduced from -.36 to -.25, and for Self Esteem, from .57 to .42. For Risk Taking the correlation remained negligible. The size of the correlation between the Self Esteem and PRF-Dy scales probably reflects something more than desirability bias in the Self Esteem scale; there may be a substantive connection between Self Esteem and desirability response style (cf. Block, 1965).

Discrimination among the personality scales. The initial multitrait-multimethod correlation matrix in Table 2 was fairly satisfactory in terms of the experimental scales' ability to discriminate among the three constructs. In the heteromethod block involving the 20-item endorsement and desirability-judgment scales, the monotrait correlations in the minor diagonal, that is, the validity coefficients, were higher than all off-diagonal correlations in their respective rows and columns. In other words the highest correlations between endorsement and desirability-judgment scales involved single constructs.

The substantial correlation between the Impulsivity and Risk Taking item pools is indicative of a lack of conceptual independence of the personality dimensions. The conceptual overlap may have arisen from the fact that the Impulsivity and Risk Taking item pools had been prepared in the contexts of different scale-development programs. Had their original preparation been for the same questionnaire, a lower degree of overlap could no doubt have been obtained. However, the fact that item selection reduced the correlation between the two from .69 to .42, yet produced scales possessing high internal consistency, indicates that the Impulsivity and Risk Taking constructs are sufficiently distinct from one another to be of independent theoretical and practical importance.

Item overlap between endorsement and desirability-judgment scales. Earlier attempts to measure personality traits through desirability judgments of personality items employed items originally selected for endorsement scales (Jackson, 1961, 1964; Kusyszyn, 1968). If in one item pool the items best suited for an endorsement scale were not the items best suited for a desirability-judgment scale, the practice of using items from endorsement scales in the desirability-judgment response mode could seriously limit the validity of desirability judgments of items for personality assessment. In contrast to previous studies, the present study placed no restrictions on the number of items that the endorsement and desirability-judgment scales could have in common. Instead, items were selected independently for the two types of scales. There was in fact some item overlap. The Impulsivity endorsement and desirability-judgment scales had four items in common, and the two Risk Taking scales shared eight items. Since 60 percent of the Risk Taking items and 80 percent of the Impulsivity items in each scale were unique to that scale, it is reasonable to surmise that items best suited to the different response modes differ from each other in some respects. It was suggested earlier that the items best suited to one response mode may differ in extremity of desirability from those best suited to the other mode, the items best suited for desirability-judgment scales being more neutral with respect to desirability (cf. Messick, 1964).

CHAPTER IV

VALIDATION PROCEDURES FOR THE PERSONALITY SCALES

The topic of the present chapter is the empirical validation procedures for the experimental personality scales whose development and initial evaluation have been described. The scales were validated against the criteria of peer ratings, self-ratings, peer rankings, and self-rankings. The experimental scales and the behaviour-judgment measures are listed in Table 3.

The validity of the scales was assessed under two conditions. Half of the subjects were given the usual instructions to perform the questionnaire tasks honestly, and, therefore, are referred to as the standard-instruction sample. The remaining subjects were instructed to distort their responses in a specified manner and are referred to as the fake-instruction sample. The instructions to fake were introduced in order to determine experimentally the resistance to faking of the different types of personality scales being tested. A direct test of the issue was included because the problem of faking has had an important bearing on the development of forced-choice methods (Norman, 1963a; Zavala, 1965) and judgmental methods (Jackson, 1964; Kusyszyn, 1968; Selltiz, Edrich, & Cook, 1965) of assessment.

Table 3

Experimental Personality Scales and Behaviour-judgment Measures Included in the Validation Procedures

Method	Trait
Personality Scales	
1. Nonforced Endorsements	Impulsivity
	Risk Taking
	Self Esteem
2. Forced-choice Endorsements	Impulsivity
	Risk Taking
	Self Esteem
3. Nonforced Desirability Judgments	Impulsivity
	Risk Taking
4. Forced-choice Desirability Judgments	Impulsivity
	Risk Taking
Behaviour Judgments	
5. Self Behaviour Rankings	Impulsivity
	Risk Taking
	Self Esteem
6. Self Behaviour Ratings	Impulsivity
	Risk Taking
	Self Esteem
7. Peer Behaviour Rankings	Impulsivity
	Risk Taking
	Self Esteem
8. Peer Behaviour Ratings	Impulsivity
	Risk Taking
	Self Esteem

Subjects

Subjects were student nurses living in a residence of a large general hospital. Out of a possible 320 subjects, 182 volunteered to take part in the study. Approximately 60 percent were first year students who had lived together for eight months. The remainder were second and third year students who had lived together continuously since the beginning of their first year in training.

Although the women would come to know one another just from living together in the same residence, they could be expected to become especially well acquainted with those women living physically near them in the residence. The list of 320 students in residence, therefore, was divided into subgroups occupying different physical zones of the residence. A zone was defined as any combination of a wing and floor in the residence; for example, Floor 3, Wing B was considered one zone. The list for each zone contained at least 10 but not more than 30 names. Since the rank ordering of more than about 15 names probably would be too difficult and time-consuming for the average subject, any list containing more than 15 names was randomly divided into two sublists. This procedure resulted in a set of 31 lists of from 8 to 15 names of potential subjects. The name of every student in the residence appeared on one or another such list. Each list was the basis for forming a group of subjects who would make behaviour judgments about one another.

A minimum of four volunteer subjects per behaviour-judgment group was required in order to assure stable peer behaviour-judgment measures. Consequently, the data from seven groups containing three or fewer subjects were discarded. In addition, nine subjects were selected randomly from the largest behaviour-judgment groups, and their data were discarded in order to equalize the number of subjects in the standard- and fake-instruction validation samples. After the exclusion of such data, there remained 162 subjects, 81 in each validation sample.

Materials

The materials were organized into two instruction booklets, two corresponding response booklets, and an envelope containing precision data cards. The first instruction booklet contained the PRF-Dy scale (Jackson, 1967b) and two behaviour-judgment tasks for obtaining criterion measures of the three personality constructs. The second instruction booklet contained the experimental personality scales. In the following paragraphs the various tasks are described in the order in which they appeared in the booklets, that is, in the order in which they were administered to subjects. The details of how the materials were administered to subjects are presented below in the Procedure section. The instructions for each task are contained in Appendix B.

PRF-*Dy* scale. The PRF-Dy scale (Jackson, 1967b) was included to aid in the interpretation of the results of the validation procedures. The instructions were to judge each statement as to how characteristic or uncharacteristic it was of self and to record the judgment in the response booklet. Ratings were made on a 9-point scale ranging from 'extremely uncharacteristic of self' to 'extremely characteristic of self.'

Behaviour descriptions. Six behaviour descriptions were prepared for purposes of obtaining behaviour-judgment criterion measures of the traits. Each pole of the Impulsivity, Risk Taking, and Self Esteem dimensions was represented by a one-paragraph description of behaviour typifying a person possessing that trait (see Appendix B). The three pairs of descriptions were carefully edited by four people to assure a degree of conceptual independence, while maintaining a close correspondence to the original definitions of the personality dimensions.

Behaviour-ranking task. Printed on the outside of each envelope were the name list for a potential behaviour-judgment group, and the three behaviour descriptions corresponding to the positive poles of the personality dimensions. The descriptions were labeled 'BEHAVIOUR DESCRIPTION 1,' 'BEHAVIOUR DESCRIPTION 2,' and 'BEHAVIOUR DESCRIPTION 3.' Inside the envelope were three separate decks of precision data cards, one deck of blue-coloured cards, a second of orange cards, and a third of yellow cards. Every card in the blue deck had a name and code number and the label 'BEHAVIOUR DESCRIPTION 1' punched in the card and printed along the top edge of the card. There was one card in the deck for each name appearing in the list on the envelope. The information on the cards of the orange and yellow decks was identical to that on the cards of the blue deck, except that they were labeled, respectively, 'BEHAVIOUR DESCRIPTION 2' and 'BEHAVIOUR DESCRIPTION 3.' A set of such envelopes and contents was prepared from each of the 31 name lists so that each member of the corresponding behaviour-judgment group would have an envelope.

The 14,000 punched data cards required for this task were computer-generated from a master list of the students in residence. The instructions were to use the three decks of cards in the envelope to rank-order the persons with regard to how characteristic each behaviour description was of each of them.

Behaviour-rating task. The behaviour-rating response sheet contained the six behaviour descriptions, a grid for recording ratings, and a list of names and code numbers (see Appendix B). The six descriptions were listed in such a way that the positive and negative descriptions for a construct were not adjacent to each other. Each was labeled 'BEHAVIOUR DESCRIPTION' with a numeral from 1 to 6. The response grid contained seven columns, the first of which had the heading 'degree of acquaintance.' The remaining columns were numbered 1 through 6 with the heading 'BEHAVIOUR DESCRIPTION.' The list of names on the rating form corresponded to the list for the ranking task, and the names were aligned with the rows of the response grid.

The first instruction required the subject to indicate in the first column of the response grid how well she knew each person on her list. Ratings were made on a 9-point scale ranging from 'do not know her at all' to 'know her extremely well.' The next instruction was to study each behaviour description in turn and to rate how characteristic or uncharacteristic it was of each person. Ratings were made on a 9-point scale ranging from 'extremely uncharacteristic' to 'extremely characteristic' of the person.

Nonforced endorsement task. The second instruction booklet contained the experimental personality scales. The nonforced endorsement task was first in the booklet. The 100-item nonforced endorsement task consisted of the 20-item Impulsivity, Risk Taking, and Self Esteem nonforced endorsement scales, whose development was described in Chapter II, and 40 filler items of heterogeneous content. The 100 items were arranged in a partially random order: items from the same scale were separated by at least two other items. The instructions were identical to those for the PRF-Dy scale.

Forced-choice endorsement task. The second task consisted of the three forced-choice endorsement scales. The order of the statements within forced-choice items was varied so that the filler statement preceded the construct-relevant statement in one half of the items of each scale. The 60 items were arranged in a partially random order. Instructions were to indicate in the response booklet which of the two statements from each pair was more characteristic of self.

Nonforced desirability-judgment task. The third task consisted of the 40 items of the Impulsivity and Risk Taking nonforced desirability-judgment scales and 20 filler items of diverse content. Again the items were arranged in a partially random order. The instructions were to judge each item as to its desirability in other women, assuming it to be characteristic of other women. The judgments were made on a 9-point scale, ranging from 'extremely undesirable in other women' to 'extremely desirable in other women.'

Forced-choice desirability-judgment task. The final task was made up of the Impulsivity and Risk Taking forced-choice desirability-judgment scales. As in the forced-choice endorsement task, the filler statement preceded the construct-relevant statement in one half of the items of each scale. The 40 items were arranged in a partially random order: items from the same scale were not adjacent. The instructions paralleled those of the nonforced desirability-judgment task except that subjects now were required to indicate with regard to each pair the statement which they regarded to be more desirable in other women.

Procedure

Two-hour testing sessions were conducted in a large examination hall by a minimum of one proctor to 30 subjects. The instruction booklets containing the PRF-Dy scale and the behaviour-judgment tasks were arranged on desks just inside the entrance to the examination hall. The arrangement of materials was such that a subject could locate easily the set of materials containing the name list which included her own name. Her materials were checked by a proctor before she was directed to a seat to await further instructions. When the group of subjects had assembled in this manner, the head proctor explained to the subjects that when they had completed the tasks before them, a proctor would exchange those materials for a second and final set of materials. Testing proceeded precisely as it had been explained to the subjects. When a subject had completed both sets of tasks, she was allowed to leave. Most subjects completed the tasks within two hours, but a few took longer.

The procedure was identical for all subjects up to the point of finishing the first set of materials. Unknown to subjects, the second set of materials was not identical for all subjects. Some response booklets had face sheets with instructions to 'fake good,' that is, to perform all tasks so as to create a favourable impression of themselves, ignoring, if necessary, what they were really like (see Appendix B). As subjects requested the second set of materials they were given, alternately, a set of materials containing a standard response booklet, or one containing a response booklet with instructions to fake printed on the face sheet. In this manner subjects were assigned alternately to either the standard-instruction sample or the fake-instruction sample.

Data Reduction and Analysis

Before the rather unwieldy raw data could be analyzed statistically, they had to be reduced to a few measures of each trait. These measures were then intercorrelated separately for the standard-instruction and fake-instruction samples. The resulting multitrait-multimethod matrices were examined both informally and analytically for evidence of convergent and discriminant validity. All computational steps in the data reduction and analysis were performed by computer.

Reduction of questionnaire data. Prior to statistical analysis, the questionnaire data of each subject were scored to yield a relatively small number of measures. The nonforced endorsement and desirability-judgment data were scored exactly as the corresponding item-analysis sample data had been scored. Taking into account the direction of keying of items, an individual's ratings of the items in each nonforced scale were summed to form a scale score. The scoring of the forced-choice endorsement and desirability-judgment data was simply a matter of calculating for each subject the number of responses in the keyed direction for each scale: on the forced-choice scales, an individual accumulated one point each time she selected either the construct-relevant statement when it was positively-keyed or the filler statement when the construct-relevant statement was negatively-keyed. In this way the maximum score she could achieve on each 20-item forced-choice scale was 20. The data reduction strategy as described produced 10 personality scale scores for each subject (see Table 3, p. 62). They were three nonforced endorsement scale scores, three forced-choice endorsement scale scores, two nonforced desirability-judgment scale scores, and two forced-choice desirability-judgment scale scores. The PRF-Dy scale data were scored in the manner described previously.

In addition to computing the above total scale scores for each subject, responses to odd- and even-numbered items within each scale were summed separately for purposes of determining the random split-half reliability of each scale.

Reduction of behaviour-judgment data. Reduction of the behaviour-judgment data was more complex than the reduction of the questionnaire data. Since the questionnaire data produced by a given subject pertained to that subject alone, they were reduced to summary scores without reference to the data of the subject's peers. On the other hand, the behaviour-judgment data produced by the same subject pertained not only to herself but also to the others in her behaviour-judgment group, as well as to some of her peers that did not take part in the experiment. Consequently, while some aspects of the data reduction could be accomplished at the level of the individual subject, the interrelatedness of the data necessitated performing a substantial amount of data reduction at the level of the intact behaviour-judgment group.

Each subject had rated from 8 to 15 persons, including herself, on six behaviour descriptions. Her six ratings of each person were reduced to three scores by subtracting her rating of the negative pole of each trait from 10, and then summing the reflected rating with the corresponding rating of the positive pole. The strategy produced for each trait a rating score for each of the subject's peers, as well a self-rating score. The self-rating scores were now in their final form, while the rating scores for the subject's peers had to undergo further reduction.

Each subject produced on the behaviour-ranking task one rank-ordered deck of name-cards per trait. Each person represented by a card received a ranking score for the relevant trait, the score being a function of the ordinal position of her card in the deck. The following formula was used.

$$\text{Ranking score} = \frac{K - (J - 1)}{K}$$

where J is the ordinal position of the card in the deck, and

K is the total number of cards in the deck.

The conversion of raw ranking scores to proportions took into account the variability in the lengths of name lists, thereby making the ranking scores roughly comparable from one behaviour-judgment group to another. Since each card deck rank-ordered by a given subject contained a card bearing the subject's own name, the foregoing strategy produced for each trait a self-ranking score and a ranking score for each of the subject's peers.

The reduction of behaviour-judgment data described to this point yielded for each subject six self-judgment scores, a persons by traits matrix of rating scores, and a similar matrix of ranking scores. Only the self-scores were now in their final form.

Working with the data of one intact behaviour-judgment group at a time, a set of scores based on judgments by peers was produced for each subject. For each subject in turn the rating and ranking score matrices of each of the other subjects in her group were searched for the data pertaining to the subject in question. If a peer had indicated she knew the subject by a degree of acquaintance rating of 2 or greater, her judgments about the subject were considered acceptable and were summed with other peers' corresponding judgments about the subject. After the data of all the other subjects in the behaviour-judgment group were searched, the three rating-score sums and three ranking-score sums for each subject were transformed to means. Use of mean behaviour-judgment scores, rather than sums, standardized the scale of measurement in all behaviour-judgment groups. The strategy just described for one subject was repeated for each of the other subjects within the behaviour-judgment group. The procedure was repeated for each behaviour-judgment group.

For purposes of computing the reliability of the peer behaviour-rating scores, rating scores pertaining to a given subject were alternately used in computing two additional mean peer behaviour-rating scores per trait. Essentially, the new mean scores so formed were random split-half mean rating scores. Therefore, the correlation between them was corrected for double the number of raters by the Spearman-Brown prophecy formula, giving the reliability coefficient of the peer behaviour ratings.

Analysis of summary data. The reduction of a massive set of raw data to a relatively small set of measures for each subject has been described in the preceding paragraphs. Besides the special scores needed for computing reliabilities, the basic set of measures per subject (see Table 3, p. 62) consisted of three nonforced endorsement scale scores, three forced-choice endorsement scale scores, two nonforced desirability-judgment scale scores, two forced-choice desirability-judgment scale scores, three self-behaviour-ranking scores, three self-behaviour-rating scores, three mean peer behaviour-ranking scores, three mean peer behaviour-rating scores, and a PRF-Dy scale score.

The summary data of the standard-instruction sample and the fake-instruction sample were analyzed separately. Pearson product-moment correlations were computed among the scores, yielding a 23 by 23 variable multitrait-multimethod correlation matrix for the standard-instruction sample and a similar matrix for the fake-instruction sample. Each correlation matrix was subjected to multimethod factor analysis, an analytic procedure designed for application to multitrait-multimethod matrices (Jackson, 1966b). In these analyses an identity matrix was substituted for the monomethod submatrix corresponding to each of the eight measurement methods represented, thereby orthogonalizing the traits within each method. The substitution of identity matrices did not affect the Desirability vector, as may be seen in Table A, Appendix C. The multimethod factor matrix of each sample was subjected to varimax rotation of three factors before interpretation.

CHAPTER V

RESULTS OF THE VALIDATION PROCEDURES

The results of the validation procedures described in the preceding chapter are treated in the present chapter. The convergent and discriminant validity of the experimental personality scales are examined in the context of multitrait-multimethod matrices involving behaviour-judgment criterion measures of the three traits, Impulsivity, Risk Taking, and Self Esteem (see Table 3, p. 62). In order to facilitate the comparative evaluation of the four types of personality scales under consideration, the empirical findings of the study are presented in three sections. The first section deals with the experimental personality scales alone, and the second with the behaviour-judgment criterion measures alone. The third and most crucial section evaluates evidence for convergent and discriminant validity in the complete multitrait-multimethod matrices, which involve both the questionnaire methods and the behaviour-judgment methods of measuring the personality traits.

Experimental Personality Scales

In the paragraphs which follow, the experimental personality scales are evaluated in terms of reliability, correlation with the PRF-Dy scale, and convergent and discriminant properties of the four questionnaire methods. The standard-instruction validation sample results are compared with those of the item-analysis sample and with those of the fake-instruction validation sample. Comparisons between the standard-instruction and item-analysis samples are important in that they show the effects of cross-validation. Comparisons with the fake-instruction sample show the effects of faking.

Sample means and standard deviations. The personality scale means and standard deviations are shown in Table 4 separately for the three samples of subjects, the two validation samples and the item-analysis sample. In comparisons of the personality scale means and standard deviations of the standard-instruction validation sample with those of the item-analysis sample, only the nonforced scales may be considered, of course, since no forced-choice scales were administered to the item-analysis sample. As Table 4 indicates, the only significant difference was between the sample means for the Risk Taking nonforced endorsement scale. The student nurses on the average achieved lower Risk Taking scale scores than did the university women. There were no significant differences between the sample standard deviations.

A difference between samples in Risk Taking tendencies might be expected to have an adverse effect on the reliability of

Table 4

Personality Scale Means and Standard Deviations

Based on the Three Samples

Sample Scale	Means			Standard Deviations		
	Standard-instruction Validation Sample (N = 81)	Fake-instruction Validation Sample (N = 81)	Item-analysis Sample (N = 244)	Standard-instruction Validation Sample (N = 81)	Fake-instruction Validation Sample (N = 81)	Item-analysis Sample (N = 244)
Nonforced Endorsements						
Impulsivity	101.5	91.5*	101.7	21.3	25.5	19.0
Risk Taking	94.6	102.7*	104.7*	20.5	22.1	21.9
Self Esteem	107.1	132.9*	109.0	23.4	28.4*	22.1
Forced-choice Endorsements						
Impulsivity	10.5	9.5		3.4	3.5	
Risk Taking	9.2	10.4*		3.0	3.3	
Self Esteem	9.6	12.4*		4.4	4.8	
Nonforced Desirability Judgments						
Impulsivity	100.1	98.7	102.1	13.3	17.9*	13.8
Risk Taking	95.6	97.9	99.6	14.6	18.0*	16.5
Forced-choice Desirability Judgments						
Impulsivity	10.1	9.6		2.3	3.2*	
Risk Taking	9.0	9.1		2.8	3.4*	

Note.-- The values marked with an asterisk (*) are significantly different from the corresponding values based on the standard-instruction validation sample ($p < .05$).

forced-choice measures of Risk Taking. A general drop in the popularity of Risk Taking statements, as reflected in the lowered Risk Taking scale means, most likely would not be accompanied by a comparable drop in the popularity of filler statements, since the latter were of heterogeneous content. Since forced-choice items were composed of statements matched on popularity for the item-analysis sample, the statements would no longer be perfectly matched. Therefore, the forced-choice items could be expected to deviate widely from the expected popularity of .50. The occurrence of such deviations would result in lower item variances and covariances, and consequently, lower scale reliability (Guilford, 1965, pp. 453-464; Magnusson, 1966, pp. 53-58). Results confirming the foregoing argument are presented below.

The standard-instruction validation sample means and standard deviations may be compared with those of the fake-instruction validation sample, also reported in Table 4. The most striking difference between validation samples concerns the standard deviations. Standard deviations for all experimental scales were greater with instructions to fake than with standard instructions, five of the ten differences being statistically significant ($p < .05$). The sample differences in standard deviation of the four desirability-judgment scales were among the five significant differences. By way of contrast, only endorsement scales showed significant differences between validation sample means. The Impulsivity endorsement scale scores tended to be lower with faking, while Risk Taking and Self Esteem endorsement scale scores tended to increase. When subjects attempted to fake on an endorsement scale, however, they were able to effect a mean shift in the desirable direction with in most cases only a minor increase in scale variability. The effects of faking were different, however, with desirability-judgment scales. Subjects apparently had discrepant interpretations of how to fake a desirability-judgment task, thereby increasing the variability of the desirability-judgment scale scores, while not affecting their means. The significance of this finding is discussed more fully in the next chapter.

Finally, the standard deviations reported in Table 4 were in general lower for the desirability-judgment scales than for the corresponding endorsement scales. The finding is in agreement with the widely-held assumption that people who differ greatly with regard to possession of a trait tend to concur on its desirability. It should be noted, however, that this assumption was questioned in Chapter I because of increasing evidence of individual conceptions of the desirable and of their relationship to personality. The issue is raised again in Chapter VI.

Reliability. The corrected split-half reliability coefficients of the experimental personality scales are presented in Table 5. For the nonforced personality scales, the reliabilities based on the standard-instruction sample may be compared with the corresponding reliabilities based on the item-analysis sample. Examination of Table 5 reveals that shifts in reliability with cross-validation were negligible. The finding of generally high reliability

Table 5

Corrected Split-half Reliabilities of the Personality Scales

Based on the Three Samples

Sample Scale	Standard-instruction Validation Sample (N = 81)	Fake-instruction Validation Sample (N = 81)	Item-analysis Sample (N = 244)
Nonforced Endorsements			
Impulsivity	.86	.92	.81
Risk Taking	.88	.82	.90
Self Esteem	.92	.95	.94
Forced-choice Endorsements			
Impulsivity	.65	.67	
Risk Taking	.50	.58	
Self Esteem	.81	.82	
Nonforced Desirability Judgments			
Impulsivity	.75	.82	.80
Risk Taking	.78	.77	.88
Forced-choice Desirability Judgments			
Impulsivity	.32	.63	
Risk Taking	.57	.60	

coefficients for the cross-validated 20-item scales speaks well for the homogeneity of the item pools and for the methods used in item selection. The reliabilities are comparable to those obtained in other studies which emphasized considerations of construct validity from the outset (e.g., Jackson, 1967b; Neill & Jackson, 1967). The findings have an important bearing on the discussion of the empirical validity of the personality scales, since the existence of substantial reliability is a precondition for obtaining high empirical validity (Nunnally, 1959, pp. 98-104).

Figure 1 presents the mean reliability for each type of scale for both the standard-instruction and fake-instruction validation samples. The presentation of mean reliabilities facilitates comparisons between the two validation samples and among the four methods of measurement. Sample differences in mean reliability, although small, were consistently in the direction of higher mean reliability for the fake-instruction sample. This was probably a function of the fact that the personality scales had larger standard deviations in the fake-instruction sample. Examination of Figure 1 reveals that differences between samples in mean reliability were in general smaller than differences between methods within samples. Within either validation sample, endorsement scales were more reliable than corresponding desirability-judgment scales, and nonforced scales were more reliable than forced-choice scales. The finding of lower reliability for forced-choice scales than for nonforced scales is contrary to what is usually expected (cf. Jackson & Payne, 1963).

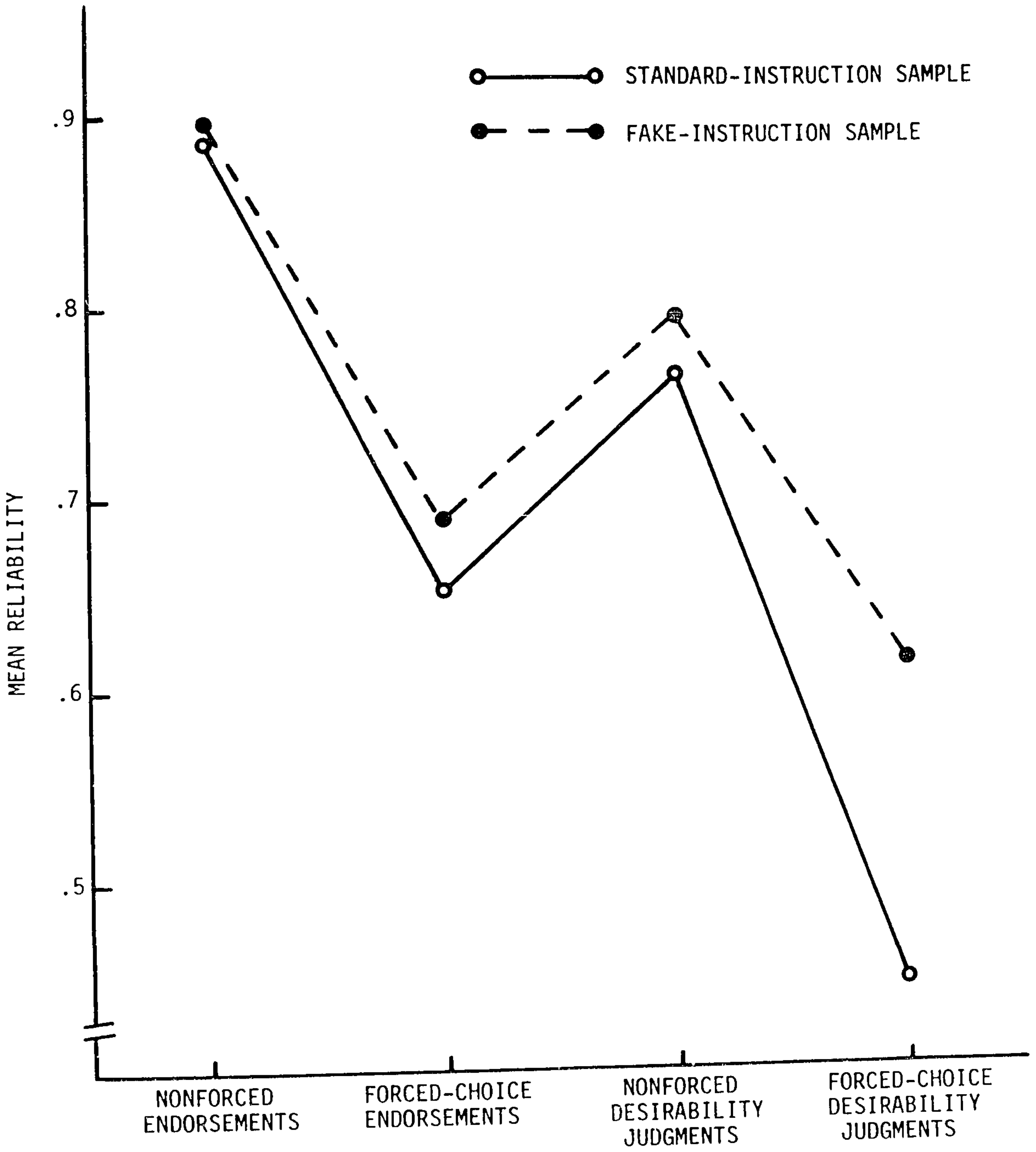

Fig. 1. Mean reliabilities of the four types of personality scales, based on the standard-instruction sample (N = 81) and the fake-instruction sample (N = 81).

It is interesting to note that the most unreliable forced-choice endorsement scale was that of Risk Taking (see Table 5). As suggested above, the lower reliability of this scale may have been the consequence of the difference between the item-analysis and validation samples in Risk Taking tendencies.

Correlation with the PRF-Dy scale. One of the aims of the item-selection strategies used in developing personality scales was to produce scales which were maximally saturated with appropriate content variance and minimally saturated with such forms of irrelevant variance as that due to individual differences in the tendency to respond desirably to self-descriptive statements (Jackson, 1966a; Loevinger, 1957; Neill & Jackson, 1967). For this reason the PRF-Dy scale played an important role in scale construction and now has an essential function in evaluating the effectiveness of the item-selection strategies in reducing desirability bias.

By the nature of the item-selection procedures, desirability variance was suppressed at the item level in all scales. The effect of such suppression is likely to be evident at the level of the total scale only when there is an asymmetry in the distribution of item correlations with the PRF-Dy scale, as was the case with the Self Esteem and Impulsivity item pools. By way of contrast, if the total endorsement scores for the item pool were negligibly correlated with the PRF-Dy scale, as was the case with the Risk Taking item pool, desirability suppression at the item level would not be evident at the scale level.

The correlations of the experimental personality scales with the PRF-Dy scale are presented in Table 6. Comparisons between the three nonforced endorsement scale correlations based on the item-analysis sample and those based on the standard-instruction validation sample indicate satisfactory results with cross-validation. The rescoring of the item-analysis sample data showed that correlations with the PRF-Dy scale were lowered through item selection. Had the observed reductions in correlations been spurious, the scales, when administered to a new sample of subjects, probably would have shown increases in correlations with the PRF-Dy scale (see Cureton, 1950). However, while the correlations with the PRF-Dy scale changed somewhat for the three endorsement scales with the new sample of subjects, the mean shift in correlation was .00.

The nonforced and forced-choice endorsement scales correlated with the PRF-Dy scale less highly in the fake-instruction sample than in the standard-instruction sample. In fact, in the fake-instruction sample, only the PRF-Dy correlations with the Self Esteem endorsement scales were statistically significant ($p < .05$). One might expect measures of desirability response style to be predictive of personality scale scores resulting from deliberate attempts to respond desirably, that is, the correlations between the PRF-Dy scale and the experimental personality scales might be expected to be higher in the fake-instruction sample than in the standard-instruction sample. The fact that the findings contradicted this prediction has important implications for understanding the relationship between desirability response style and faking. The

Table 6

Correlations between the Personality Scales and the PRF-Dy Scale Based on the Three Samples

Sample Scale	Standard-instruction Validation Sample (N = 81)	Fake-instruction Validation Sample (N = 81)	Item-analysis Sample (N = 244)
Nonforced Endorsements			
Impulsivity	-.37	-.10	-.25
Risk Taking	-.03	.01	-.05
Self Esteem	.31	.25	.42
Forced-choice Endorsements			
Impulsivity	-.33	-.07	
Risk Taking	-.05	-.00	
Self Esteem	.38	.23	
Nonforced Desirability Judgments			
Impulsivity	-.14	-.14	-.16
Risk Taking	-.12	-.10	-.13
Forced-choice Desirability Judgments			
Impulsivity	-.06	-.10	
Risk Taking	.01	-.05	

Note.-- The .05 and .01 significance levels of r are .22 and .28, respectively, for the validation samples; and for the item-analysis sample they are .14 and .18, respectively.

nature of this relationship is discussed in the next chapter when other pertinent results have been presented.

The striking similarity between the forced-choice and nonforced endorsement scales with regard to the magnitudes of their correlations with the PRF-Dy scale may be observed in Table 6. In either the standard- or fake-instruction sample, the two endorsement scales for each trait had very similar correlations with the PRF-Dy scale. As explained in Chapter I (see pp. 15-16), the rationale for pairing statements on popularity involved an expected reduction in desirability. The fact that no such reduction occurred is at least consistent with an argument against forced-choice methods, also proffered in Chapter I, namely, that pairing statements does not reduce desirability bias because the mere pairing of statements accentuates subtle differences in desirability between the statements.

Finally, it is important to note that the desirability-judgment scale scores were not significantly correlated with the PRF-Dy scale scores in either of the validation samples. The finding supports the argument that judgmental methods of assessing personality are more immune than are endorsement methods to the influence of response styles of the type measured by the PRF-Dy scale. Their susceptibility to other types of response style is discussed below.

Discrimination among the personality scales. One of the aims of the item-selection procedures was to produce maximally independent personality scales. Table 7 presents the inter-

Table 7

Correlations among the Experimental Personality Scales

Method	Trait		Nonforced Endorsements			Forced-choice Endorsements			Nonforced Desirability Judgments		Forced-choice Desirability Judgments	
			Im	RT	SE	Im	RT	SE	Im	RT	Im	RT
Nonforced Endorsements	Impulsivity	Im		51	24	80	40	03	54	22	23	-06
	Risk Taking	RT	52		27	39	62	07	45	58	06	17
	Self Esteem	SE	-30	02		20	03	82	25	12	-09	06
Forced-choice Endorsements	Impulsivity	Im	69	34	-07		30	08	45	18	22	-06
	Risk Taking	RT	43	71	-03	38		-05	34	46	10	20
	Self Esteem	SE	-25	04	67	-21	05		07	04	-11	01
Nonforced Desirability Judgments	Impulsivity	Im	66	50	-02	55	40	01		52	35	07
	Risk Taking	RT	35	64	01	17	57	08	64		15	48
Forced-choice Desirability Judgments	Impulsivity	Im	47	33	01	38	26	-01	68	48		14
	Risk Taking	RT	21	36	05	09	38	12	39	69	41	

Note.-- The Pearson product-moment intercorrelation matrix for the standard-instruction validation sample (N = 81) appears above the major diagonal; the intercorrelations for the fake-instruction sample (N = 81) appear below the major diagonal. In Tables 7 to 10 inclusive, the .05 and .01 significance levels of r are .22 and .28, respectively. Decimals have been omitted from Tables 7 to 11 inclusive.

correlations among the experimental personality scales, above the major diagonal for the standard-instruction sample, and below the major diagonal for the fake-instruction sample. It was shown in Table 2 (p. 57) that the correlation between the Impulsivity and Risk Taking item-pool endorsement scores was .69, and that the correlation was reduced to .42 through item selection. In the standard-instruction validation sample the correlation between Impulsivity and Risk Taking nonforced endorsement scales was .51; it remained below .69 with cross-validation. Comparison of the other nonforced scale intercorrelations for the standard-instruction validation sample with the corresponding intercorrelations for the item-analysis sample (Table 2, p. 57) shows that the reductions in scale intercorrelations through item selection were relatively stable with cross-validation: scale intercorrelations for the most part were not higher for the new sample of subjects.

Between-samples comparisons of the monomethod triangles in Table 7 reveal few differences in the magnitudes of correlations. The endorsement scale intercorrelations were of about the same magnitude for both samples, while the desirability-judgment scale intercorrelations were somewhat higher for the fake-instruction sample. The finding that the scale intercorrelations did not increase appreciably with instructions to fake is contrary to the familiar argument that instructions to fake reduce the dimensionality of responses to items because subjects probably respond to items with reference to a single desirability dimension rather than with reference to several construct dimensions. The implications of the discrepancy between the finding just reported and the popular claim about the effects of faking are discussed in the next chapter.

Table 7 contains two multitrait-multimethod correlation matrices, one for the standard-instruction sample and one for the fake-instruction sample. Since four questionnaire methods of measuring personality traits were involved, the underlined minor-diagonal elements are convergent validities of the four types of scales against one another. For the standard-instruction sample the mean validity was .46, and the validities ranged from .17 to .82. For the fake-instruction sample the mean was .57, and the validities ranged from .36 to .71. The foregoing indicates the presence of good convergent validity among the experimental scales for measuring Impulsivity, Risk Taking, and Self Esteem.

The main criterion of discriminant validity is that the minor diagonal validities should be greater than the off-diagonal correlations in their respective rows and columns (Campbell & Fiske, 1959; Jackson, 1966b). Table 7 shows that without exception in either validation sample the validity-diagonal correlations exceeded all relevant off-diagonal correlations. The finding of good discrimination among scales supports the choice of item-analysis procedures, especially in the light of the fact that two of the traits, Impulsivity and Risk Taking, were conceptually overlapping.

Behaviour-judgment Measures

The reliabilities of the peer behaviour ratings were computed by randomly dividing each behaviour-judgment group into two subgroups, computing the mean peer behaviour rating for each subject from each subgroup, and then within each validation sample and for each trait, computing the correlation between the two mean ratings. Since these correlations were in essence split-half reliabilities, they were corrected with the Spearman-Brown prophecy formula. The corrected reliabilities for Impulsivity, Risk Taking, and Self Esteem peer behaviour ratings were .80, .90, and .82, respectively, for the standard-instruction sample, and .83, .78, and .73, respectively, for the fake-instruction sample. The number of peers contributing to the mean rating of a subject by either subgroup ranged from one to five. In view of the small size of the rating groups, the reliabilities were quite high. They compared favourably with those reported by Kusyszyn (1968), for example, who used larger rating groups and reported a median reliability of .81 for fraternity brothers' ratings of each other on 16 trait-adjectives.

Table 8 contains two multitrait-multimethod matrices, each involving the measurement of the three traits by four behaviour-judgment methods. The underlined minor-diagonal elements are convergent validity coefficients of the methods against one another. The mean validity coefficient in each validation sample was .59. Clearly the four behaviour-judgment methods of measuring the three traits showed adequate convergent validity against one another.

While Table 8 presents evidence of convergent validity for the three traits, it shows a lack of good discriminant validity, especially for the peer behaviour-judgment measures. Many of the

Table 8

Correlations among the Behaviour-judgment Measures

Method	Trait		Self Rankings			Self Ratings			Peer Rankings			Peer Ratings		
			Im	RT	SE	Im	RT	SE	Im	RT	SE	Im	RT	SE
Self Rankings	Impulsivity	Im		59	26	66	53	31	52	56	04	53	48	18
	Risk Taking	RT	41		09	42	77	10	52	53	14	48	47	29
	Self Esteem	SE	19	21		-03	16	74	28	23	45	22	25	45
Self Ratings	Impulsivity	Im	62	34	25		44	11	50	46	32	46	48	10
	Risk Taking	RT	33	77	25	38		22	50	50	25	50	58	34
	Self Esteem	SE	20	22	74	26	37		32	45	44	38	40	44
Peer Rankings	Impulsivity	Im	44	42	31	46	56	38		90	30	90	86	49
	Risk Taking	RT	44	48	41	51	60	35	87		25	86	88	45
	Self Esteem	SE	22	19	37	03	19	37	41	46		21	23	84
Peer Ratings	Impulsivity	Im	47	42	31	52	51	36	92	86	42		85	41
	Risk Taking	RT	41	48	38	48	50	40	84	91	47	88		44
	Self Esteem	SE	34	29	52	49	39	59	53	59	85	53	61	

Note.-- The Pearson product-moment intercorrelation matrix for the standard-instruction validation sample (N = 81) appears above the major diagonal; the intercorrelations for the fake-instruction sample (N = 81) appear below the major diagonal.

validities were exceeded by off-diagonal correlations in the heteromethod blocks, as well as by monomethod correlations. In fact, in both samples, the correlation between the Impulsivity and Risk Taking peer behaviour ratings were as high as or higher than the reliabilities of the two measures. Apparently, the judges possessed very little ability to discriminate among the behaviour descriptions as they applied to other people.

Even though the self-judgment measures could discriminate among traits better than the peer-judgment measures and the latter could discriminate only very poorly, both methods were included in the multitrait-multimethod matrices involving both questionnaire and behaviour-judgment methods. While the peer-judgment measures had the disadvantage of poor discrimination, they had the advantage of being most unlike the questionnaire measures, thereby sharing probably very little error variance with the questionnaire measures. The inclusion of a measurement method which has serious shortcomings by itself, but whose measurement error is uncorrelated with that of other methods, may still make an important contribution to the convergence of the different methods on the personality traits in question (Campbell & Fiske, 1959; Jackson, 1966b). Support for the foregoing argument is provided by the results of the multimethod factor analysis presented below.

Validity of the Experimental Personality Scales

The results which have been presented may be summarized as follows. The personality scales had moderate to high reliability and showed excellent discrimination among one another with respect to the three traits, in spite of the substantial conceptual overlap between Impulsivity and Risk Taking. The behaviour-judgment measures were also reliable, but lacked adequate discrimination with respect to the three traits. The latter finding might be expected to affect the apparent discriminant validity of the personality scales against the criteria of the behaviour-judgment measures.

Table 9 contains the correlations between the experimental personality scales and the behaviour-judgment measures for the standard-instruction validation sample, and Table 10 presents the comparable correlations for the fake-instruction sample. The two are submatrices of the complete 23 by 23 multitrait-multimethod correlation matrices presented in Table A of Appendix C.

The minor-diagonal elements of the 16 heteromethod blocks in Tables 9 and 10 are underlined to indicate that they are convergent validity coefficients of personality scales against behaviour-judgment criteria. For the standard-instruction sample, 27 of the 40 validities were statistically significant ($p < .05$). The validities ranged in magnitude from -.19 to .87, with a mean of .35. For the fake-instruction sample, 31 validities were significant ($p < .05$). These validities ranged from .07 to .53, with a mean of .30. A general trend in both samples was for the personality scales to achieve higher validities against self-judgment measures than against peer-judgment measures. The mean validities against peer behaviour-judgment and self-behaviour-judgment criteria were .26 and .45, respectively, for the standard-instruction sample, and .24 and .37, respectively, for

Table 9

Correlations between the Personality Scales and the Behaviour-judgment Measures

for the Standard-instruction Validation Sample

(N = 81)

Method	Trait		Self Rankings			Self Ratings			Peer Rankings			Peer Ratings		
			Im	RT	SE	Im	RT	SE	Im	RT	SE	Im	RT	SE
Nonforced Endorsements	Impulsivity	Im	65	59	08	73	59	19	55	60	15	53	60	24
	Risk Taking	RT	37	62	16	42	65	22	43	48	12	42	50	28
	Self Esteem	SE	31	12	68	17	25	87	37	35	43	35	41	47
Forced-choice Endorsements	Impulsivity	Im	62	48	16	63	45	20	47	50	10	43	51	19
	Risk Taking	RT	18	34	03	37	40	00	16	17	05	13	18	05
	Self Esteem	SE	16	02	69	-06	17	68	22	20	37	17	26	36
Nonforced Desirability Judgments	Impulsivity	Im	38	31	06	48	32	18	11	15	05	15	19	10
	Risk Taking	RT	15	22	10	23	25	09	11	14	09	08	17	20
Forced-choice Desirability Judgments	Impulsivity	Im	-01	-02	-12	25	09	-17	-10	-11	-06	-19	-06	-06
	Risk Taking	RT	05	-11	03	09	-01	05	-01	01	-06	-09	-05	04

Table 10

Correlations between the Personality Scales and the Behaviour-judgment Measures

for the Fake-instruction Validation Sample

(N = 81)

Method	Trait		Self Rankings			Self Ratings			Peer Rankings			Peer Ratings		
			Im	RT	SE	Im	RT	SE	Im	RT	SE	Im	RT	SE
Nonforced Endorsements	Impulsivity	Im	31	22	-03	44	23	15	29	34	15	30	27	29
	Risk Taking	RT	31	43	04	33	53	13	20	22	04	22	16	21
	Self Esteem	SE	10	17	51	18	19	43	13	19	21	14	18	31
Forced-choice Endorsements	Impulsivity	Im	36	22	08	51	24	16	30	33	11	32	26	32
	Risk Taking	RT	15	37	08	30	50	11	15	22	-16	19	16	02
	Self Esteem	SE	18	12	47	16	20	47	22	19	20	24	19	27
Nonforced Desirability Judgments	Impulsivity	Im	36	25	04	39	23	24	22	39	17	29	36	32
	Risk Taking	RT	24	27	03	24	35	14	07	19	-05	12	19	12
Forced-choice Desirability Judgments	Impulsivity	Im	22	16	-09	25	22	13	19	34	22	25	29	26
	Risk Taking	RT	21	07	06	24	24	07	10	22	04	14	19	13

the fake-instruction sample. The questionnaire measures, which were based on the subject's own responses, probably shared more method variance with self-judgment measures than they did with peer-judgment measures.

The results may be examined to determine the effects of instructions to fake on the relative validity of the four types of personality scales. In the standard-instruction sample, the different methods yielded validities of clearly different magnitudes. Nonforced endorsement scales were the most valid, with a mean validity of .60. Forced-choice endorsement scales were next with a mean validity of .44. Then came nonforced desirability-judgment scales with a mean of .24. The forced-choice desirability-judgment scales were least valid, with a mean validity of -.03. By way of contrast, differences between the validities of the four methods were much less pronounced in the fake-instruction sample. The mean validities for the nonforced endorsement scales, forced-choice endorsement scales, nonforced desirability-judgment scales, and forced-choice desirability-judgment scales were, respectively, .35, .35, .28, and .20. The foregoing indicates that instructions to fake lowered the validities of the endorsement scales and increased the validities of the desirability-judgment scales. Figure 2 further illustrates the interaction between type of scale and instructional set by showing the mean validities of the four types of scales under the two instructional sets, this time using only the peer behaviour-judgment criteria. Implications of these findings are presented in the next chapter.

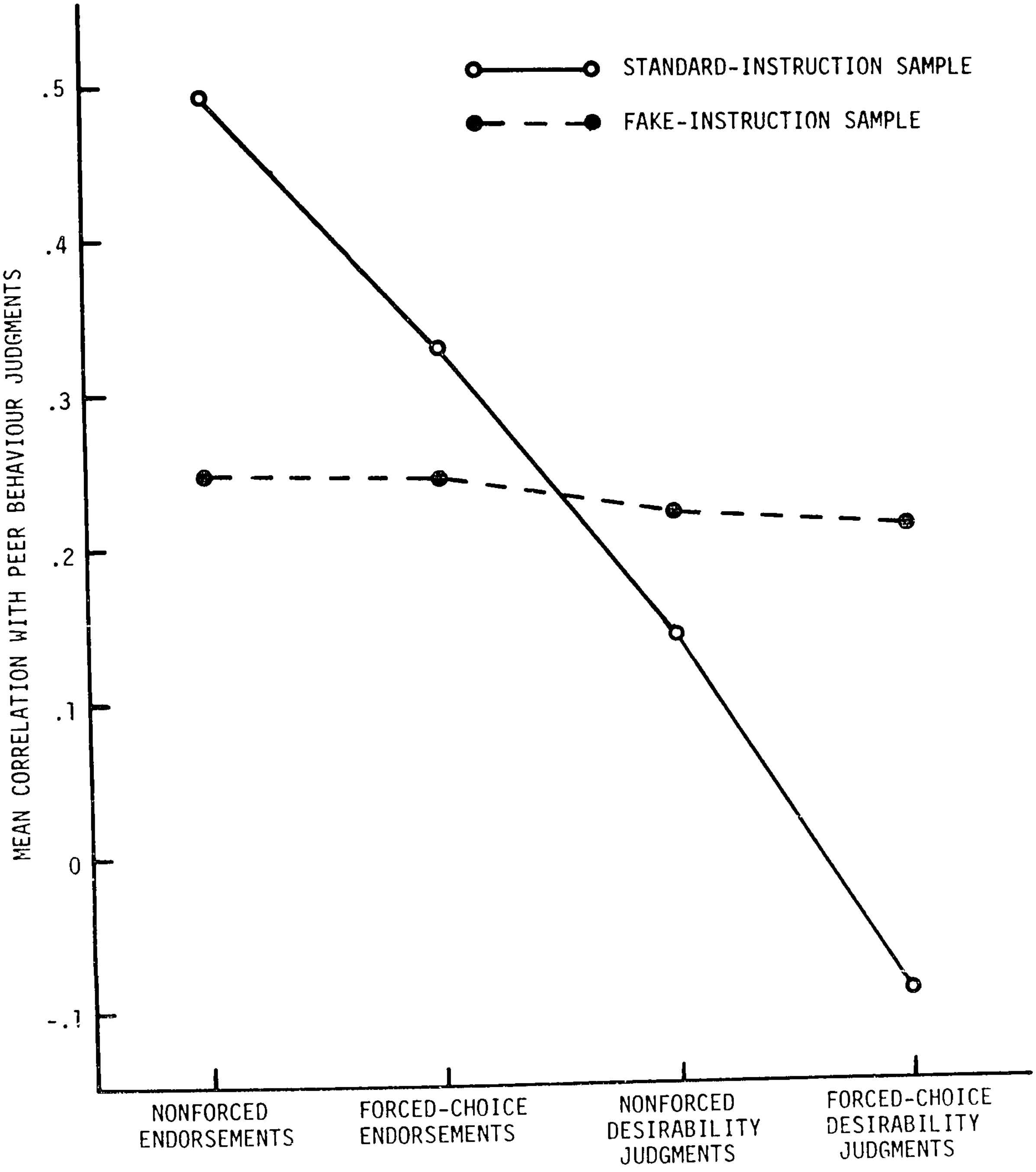

Fig. 2. Mean correlations of the four types of personality scales with peer behaviour judgments, based on the standard-instruction sample (N = 81) and the fake-instruction sample (N = 81).

Any discussion of validity must deal with, not only convergent validity, but also discriminant validity. When an attempt is made to use informal procedures for determining the level of discriminant validity in the intercorrelation submatrices in Tables 9 and 10, however, the inadequacy of such procedures is evident. The matrices represent few traits relative to the number of methods. Therefore, when the criterion measures of two traits are almost totally overlapping, as was the case with Impulsivity and Risk Taking, potentially two-thirds of the personality scales, which showed good discrimination among themselves, could appear to be lacking in discriminant validity, simply because of the criterion overlap. In this kind of situation, an analytic procedure, such as multimethod factor analysis, could be expected to yield a more meaningful solution (Jackson, 1966b).

Results of the multimethod factor analysis. As explained in the introduction, Jackson's multimethod factor analysis is an analytic technique designed specifically for appraising evidence of convergent and discriminant validity in a multitrait-multimethod matrix as a whole. In utilizing the technique, identity matrices were substituted for each of the eight monomethod submatrices as shown in Table A of Appendix C. In this way the substantial method variance components, as evidenced by the magnitudes of the monomethod correlations, cannot influence the factor structure. This is not to say, of course, that there remains no method variance in the matrices. The variance associated with two or more methods is not removed, but is reflected in the heteromethod correlations; the variance uniquely associated with a particular method is removed. The multimethod factor matrices, which have been subjected to varimax rotation, are presented in Table 11 for both validation samples. Except for the PRF-Dy scale factor loadings, loadings below .40 have been omitted. Both the unrotated and rotated multimethod factor matrices are presented in an unabbreviated form in Tables B, C, and D of Appendix C.

The standard-instruction sample yielded three clearly-interpretable factors. The variables with the highest loadings on Factors 1, 2, and 3 were, respectively, the Impulsivity endorsement scales, the Self Esteem endorsement scales and self-behaviour-judgment measures, and the Risk Taking endorsement and desirability-judgment scales. The peer behaviour-judgment measures of Impulsivity also had high loadings on Factor 1 and those of Self Esteem had high loadings on Factor 2. As expected from its correlations with the Impulsivity and Self Esteem scales, respectively, the PRF-Dy scale had a moderate negative loading on Factor 1 and a moderate positive loading on Factor 2.

The only loadings which did not support the one-trait-per-factor interpretation given above were those of the Risk Taking behaviour judgments. The Risk Taking behaviour judgments had higher loadings on Factor 1, the Impulsivity factor, than on Factor 3, which was defined by Risk Taking scales. This anomaly can be explained easily by the high correlation between the Impulsivity and Risk Taking behaviour judgments. The correlations were high enough to warrant the interpretation that subjects did not

Table 11

Rotated Multimethod Factor Matrices

Standard-instruction and Fake-instruction Samples

	Standard-instruction Sample			Fake-instruction Sample		
Method Factor	I	II	III	I	II	III
Measures of Impulsivity						
Nonforced Endorsements	88			71		
Forced-choice Endorsements	81			72		
Nonforced Desirability Judgments	44			68		
Forced-choice Desirability Judgments				55		
Self Rankings	76			63		
Self Ratings	73			74		
Peer Rankings	72			63		
Peer Ratings	73			63		
Measures of Risk Taking						
Nonforced Endorsements	44		73			81
Forced-choice Endorsements			74			84
Nonforced Desirability Judgments			78			83
Forced-choice Desirability Judgments			55			63
Self Rankings	64					46
Self Ratings	65					59
Peer Rankings	75			63		
Peer Ratings	72			60		
Measures of Self Esteem						
Nonforced Endorsements		88			69	
Forced-choice Endorsements		84			71	
Self Rankings		83			84	
Self Ratings		86			79	
Peer Rankings		62		43	47	
Peer Ratings		63		52	56	
PRF-Dy	-39	48	-02	-24	58	-02

Note.-- Except for the PRF-Dy loadings, factor loadings below .40 have been omitted.

distinguish between the Risk Taking and Impulsivity behaviour descriptions as they applied to their peers. Although the multimethod factor analysis excluded the actual monomethod correlations, it did not alter the fact that subjects used the Impulsivity and Risk Taking behaviour descriptions interchangeably. Examination of each heteromethod submatrix involving peer behaviour judgments in Table 9 indicates that the correlation of the Impulsivity scale with the Impulsivity peer behaviour judgments and the correlation of the Risk Taking scale with the Risk Taking peer behaviour judgments were both exceeded by the correlation between the Impulsivity scale and the Risk Taking peer behaviour judgments. In simpler terms, the Risk Taking behaviour description was regarded by subjects as the description most appropriate to the Impulsivity dimension.

The remarkable thing is that in spite of the above problem with the peer behaviour-judgment criteria, the good discrimination among the personality scales and the scales' high reliability were sufficient to determine three trait factors. The preceding is good evidence of the convergent and discriminant validity of the personality scales when administered under standard instructions.

When subjects were given instructions to fake by responding in the desirable direction to all questionnaire items, the factor pattern changed very little. Each factor was still defined by the questionnaire measures of one trait. The only important difference was that with instructions to fake the desirability-judgment scales achieved higher factor loadings, a finding which should follow from their higher convergent validities. The pattern of loadings for the peer behaviour-judgment measures shifted slightly, but was still consistent with the nonindependence of the behaviour-judgment measures, as explained in the discussion of the factor matrix for the standard-instruction sample. The PRF-Dy scale again had a negative loading on Factor 1 and a positive loading on Factor 2.

The multimethod factor analysis, in summary, demonstrated good convergent and discriminant validity, and therefore, construct validity, for the experimental personality scales. The validity held up even when subjects were instructed to fake. The technique made possible the appraisal of convergent and discriminant validity in spite of the problem of nonindependent criterion measures of two of the traits.

CHAPTER VI

DISCUSSION

In Chapter I the problems of desirability response style and faking were discussed as they related to questionnaire measures of personality traits. Among the several proposed remedies discussed were forced-choice methods and judgmental methods, two approaches which underwent further testing in the present investigation. Also introduced in Chapter I was the topic of construct validity. It was shown how construct validity subsumes the problems of desirability and faking biases. Construct validity is in part grounded in the psychology of structured test behaviour, which includes, of course, desirability responding and faking. It seemed that probably the most fruitful approach to further elucidation of the two problems would be to construct new scales with construct validity considerations in the forefront from the beginning. Therefore, while three of the four types of experimental personality scales were not new in format, they did provide a better opportunity for understanding desirability response style and faking, and for possibly doing something about the two problems.

In Chapters III and V a number of detailed comparisons among the experimental personality scales were made and briefly discussed. Because of the complexity of the data, however, it has been necessary to limit most of the discussion to individual findings. The purpose of the present chapter is to integrate the various findings in order to present a more coherent interpretation. Central to the ensuing discussion is the theme of construct validity. Findings concerning desirability response style and faking are integrated with this theme.

Nonforced Endorsement Scales

Nonforced endorsement scales are by far the most common type of personality scale. More research on the problems of desirability response style and faking has been conducted on this type of scale than on any other. In addition, the most ambitious efforts at achieving construct validity in personality scales have been with nonforced endorsement scales (e.g., Jackson, 1967b). Consequently, it was considered advantageous to include in the present research the development of nonforced endorsement scales and to use the results with such scales as the focal point in the comparative evaluation of the different types of scales.

The study has again demonstrated the feasibility of building construct valid endorsement scales for measuring personality traits, by beginning with substantive considerations (cf. Jackson, 1967b; Neill & Jackson, 1967). High internal consistency was achieved by item pools in which item preparation was based on constructs from psychological theory. Item analysis and subsequent item selection was designed to achieve structural validity and at the same time to reduce desirability variance and scale intercorrelations. The resultant 20-item endorsement scales were highly correlated with and almost as reliable as the item-pool total endorsement scores. Reductions in correlations with the PRF-Dy scale and with other scales held up fairly well under cross-validation.

The fact that reductions in desirability variance were achieved in constructing short endorsement scales from substantively valid item pools is consistent with the results of other studies in which similar explicit attempts to reduce desirability variance were made (Jackson, 1967b; Neill & Jackson, 1967). It should be noted that the criterion used for evaluating reduction in desirability variance involved correlations between total scale scores, while the actual reduction in desirability was effected at the item level by incorporating item correlations with the PRF-Dy scale into the item-selection indices. Therefore, while items correlating highly with the PRF-Dy scale were excluded from all scales, the effects of this exclusion on total scale scores could only be apparent if the distribution of item correlations with the PRF-Dy scale was not symmetrical about zero, as, for example, was the case with the Self Esteem item pool. Perhaps the substantial validity of the endorsement scales was in part due to a reduction in desirability bias that was not even apparent at the level of the total scale.

The scale-development program sought to develop instruments possessing discriminant validity. The basic strategy of attempting to secure construct validity included steps aimed at enhancing convergent validity and reducing scale intercorrelations through item selection. As was shown in the preceding chapter, the scales did achieve satisfactory levels of convergent and discriminant validity. In spite of the fact that Impulsivity and Risk Taking were conceptually overlapping and the descriptions depicting the two traits were almost indistinguishable to subjects, both the Impulsivity and Risk Taking scales were sufficiently different from each other and had sufficient internal consistency to each define a factor in the multimethod factor analysis for each validation sample. The Self Esteem scales also defined a factor for each sample. In short, the scale-development program was able to produce scales that, while substantially correlated, demonstrated the ability to measure distinct constructs.

Forced-choice Endorsement Scales

One of the two types of endorsement scales under examination involved the forced-choice item format. Subjects were required to select from each pair of statements the more self-characteristic statement. The statements were matched on popularity so as to maximize item variance and improve scale reliability. Because of the high correlation reported between item popularity and desirability scale value, it has been reasoned that matching items on popularity should also curtail the influence of desirability response style and make the items more resistant to faking, thereby enhancing convergent and discriminant validity. Contrary to expectation, however, the forced-choice items deviated from the expected popularity of .50, as evidenced by the fact that the scale means deviated from the expected mean of 10; the forced-choice endorsement scales were less reliable than their nonforced counterparts; they had lower convergent validities under standard instructions than the corresponding nonforced scales; their convergent validities were lower for the fake-instruction sample than for the standard-instruction sample; and they correlated with the PRF-Dy scale to about the same degree as did the nonforced scales.

The foregoing evidence may be summarized by stating that the forced-choice scales developed in the present investigation were not superior to or even equal to the nonforced scales by any of the criteria employed. This fact, coupled with forced-choice scales' uneconomical feature of requiring subjects to read two statements for every response, might easily be construed to indicate that the forced-choice technique is simply not a viable technique for assessing personality traits. Such a conclusion is probably not justified, however, because of other considerations. For one thing, the forced-choice endorsement scales did possess some empirical validity under standard instructions, even though not as much as the nonforced scales; and under fake instructions the two had very similar convergent validities. The finding of nonzero validity under fake instructions is contrary to Norman's (1963a) finding of zero validity for forced-choice scales when subjects were instructed to fake. The most plausible explanation for the discrepancy in findings is a basic difference in approaches to scale construction. A construct approach was adopted in the present investigation while Norman employed an empirical-keying approach. This difference suggests a less contentious conclusion about the forced-choice method. A construct approach led to valid nonforced and forced-choice scales, but the nonforced scales demonstrated better construct properties and seem to merit more consideration for research and practical use than forced-choice scales.

Desirability-judgment Scales

Two types of personality scales under investigation were nonforced and forced-choice desirability-judgment scales. For the former, subjects were required to judge single statements as to their desirability in other women, and for the latter, subjects had to select from paired statements the one they regarded as more desirable in other women. The items selected for the desirability-judgment scales were those whose desirability judgments were most highly correlated with the total endorsement scores for the corresponding item pools. The filler statements in the forced-choice scales were matched on DSV with construct-relevant statements. It was reasoned that by selecting items in the manner described the validity of the desirability-judgment scales against behaviour-judgment criteria would be greater than had previously been reported. In previous studies no attempts were made to enhance the validity of desirability judgments through item selection; existing endorsement scales were administered without alteration in the desirability-judgment response mode.

The item-selection strategies employed in scale construction accomplished the aim of producing desirability-judgment scales which correlated substantially with the corresponding endorsement scales. The correlations between the two types of scales were higher than the highest of the comparable correlations reported by Kusyszyn (1968). Contrary to expectation, the increases in correlations of desirability-judgment scales with endorsement scales of the same constructs did not result in any improvement in the external validity of the desirability-judgment scales; in the standard-instruction sample the nonforced desirability-judgment scales had slightly lower validities against peer-judgment criteria than the mean validity obtained by Kusyszyn (1968), and the forced-choice desirability-judgment scales had zero validity.

The fact that desirability-judgment scales correlated substantially with the endorsement scales but not with the peer-judgment criteria suggests the presence of suppressor effects. Individual differences in judgments of desirability apparently account for a part of the reliable variance of the endorsements that is not related to the variance of the criterion measures. If the effects of individual differences in desirability judgments were regressed out of the endorsement scale scores, an increase in the external validity of the endorsement scales is a distinct possibility (cf. Kusyszyn, 1968).

There is evidence to suggest, however, that such suppressor effects may be operative only in scales composed of items whose endorsements are relatively neutral with respect to desirability.

First, for the construct most closely associated with the desirability response style construct, Self Esteem, the correlations between the desirability judgments of items and the Self Esteem item-pool total endorsement scores were very low. Secondly, in Kusyszyn's (1968) data, the rank-order correlation between the endorsement scales' correlations with the PRF-Dy scale and their correlations with the corresponding desirability-judgment scales was -.41, indicating that the scales whose endorsements share the most variance with desirability judgments tend to be the scales least associated with desirability response style. Thirdly, Messick (1964) found that correlations between endorsements and desirability judgments were highest for items having neutral DSVs and lowest for items having extreme DSVs.

In order to integrate these three findings, the issue of response styles must be raised in relation to desirability-judgment scales. The desirability-judgment scales were in general not correlated with the PRF-Dy scale, unlike their endorsement-scale counterparts. This was probably due to the fact that the PRF-Dy scale was designed to measure a response style associated with endorsements of items. There are also response styles associated with desirability judgments; for example, there are consistent individual differences in the tendency to use extreme or neutral categories while judging the desirability of items. There is some evidence that items not neutral with respect to desirability tend to elicit this 'extremity response style' from subjects who are judging the desirability of items (Jackson, 1968, personal communication). The operation of extremity response style is probably fundamental to any explanation of the findings cited in the preceding paragraph; for example, the reason why Self Esteem desirability-judgment scales could not be developed may well have been that extremity response style was accounting for too large a proportion of the variance of the desirability judgments of Self Esteem items.

Discussion of the desirability-judgment scales to this point has been concerned primarily with their characteristics when administered with standard research instructions. It was predicted in Chapter I (p. 39) that against peer-judgment criteria they would be no less valid for the fake-instruction sample than for the standard-instruction sample. The finding of higher validity with faking went beyond expectation. The results suggest that the validity of judgmental tasks may be enhanced by presenting them as maximum-performance tasks, where impression management is expected. Obviously, further research on both judgmental methods and impression management must precede the firm acceptance of such a proposition.

The Problem of Faking--An Overview

The present investigation concerned in part the effects of faking on different types of personality scales. Results have been presented and discussed in relation to each type of scale individually. The purpose of the present section is to attempt to yield further insight into the nature of faking by summarizing and integrating the various findings regarding faking.

Results presented above showed that the convergent validities of the four types of personality scales were differentially affected by deliberate faking. In the standard-instruction sample, the four methods were ordered neatly with regard to validity against peer-judgment criteria: nonforced endorsements were most valid, then forced-choice endorsements, then nonforced desirability judgments, and finally, with virtually no validity at all, forced-choice desirability judgments. In the fake-instruction sample, in contrast, the four types of scales showed only minor variations in validity (see Figure 2). The shifts in validity with faking represented decreases in validity for the endorsement scales and increases in validity for the desirability-judgment scales. An interpretation of the latter was presented in the preceding section. However, one important aspect of the combined findings should be stressed; although some of the convergent validities were low with faking, in no instance did instructions to fake reduce a scale's validity to zero, contrary to what is commonly assumed to occur (cf. Norman, 1963a).

The effects of faking were evidenced not only by differences between samples in the convergent validity of the experimental personality scales, but also by sample differences in the reliability, variability, and discriminant validity of the scales. Before reviewing these results, however, it would be appropriate to summarize the popular conception of the nature and effects of faking. Underlying the popular conception are two related assumptions. First, people can relate personality items to a single desirability dimension, and secondly, people tend to agree on what constitutes a desirable response to an item. Several implications can be derived from these assumptions. First, if people can respond to items from personality scales of varying content with reference to a single desirability dimension, instructions to fake by responding desirably should increase the intercorrelations among the scales and reduce the scales' discriminant validity as construct scales. Secondly, on the same assumption, when subjects are instructed to fake, scores computed from combining scales should be reliable in the internal-consistency sense. Thirdly, if people tend to agree on what constitutes a desirable response to an item, their scale scores should exhibit less variability with instructions to fake than with standard instructions.

The results reported in the preceding chapter were clearly conflicting with the popular conception of faking just outlined. The experimental scales were generally not more highly intercorrelated in the fake-instruction sample than in the standard-instruction sample. This, in combination with high scale reliability with faking, allowed the scales to demonstrate good discriminant validity as construct scales. The reliabilities of combined scales were not computed; but in view of the fact that with faking the degree of discrimination among the scales and the magnitudes of the reliabilities of individual scales were sufficient for the scales to define three construct factors in the multimethod factor analysis, reliability coefficients of combined scales would necessarily have been low. Finally, the variability of the experimental scales was consistently greater with faking than with standard instructions, exactly opposite to what the popular conception suggested should have happened.

The discrepancy between the present findings and the popular view of the effects of faking indicates that the concept of faking needs rethinking. For one thing, faking has often been assumed to be akin to desirability response style; that is, the _ability_ to respond desirably intentionally has been assumed to be related to the _tendency_ to respond desirably as a response style. Coupled with the findings just reviewed, the fact that the experimental scales had lower correlations with the PRF-Dy scale in the fake-instruction sample than in the standard-instruction sample indicates that no such relationship exists. Open to reinterpretation are the results of many published studies where increased control of desirability response style has been assumed to have reduced the problem of faking, or vice versa.

Construct Validity

While the disparities between past thinking and present findings have been brought into focus in the preceding section, an overall explanation for the disparities has yet to be offered. The principles of construct validity have been required to account for a number of specific findings. The aim of the present section is to show that the concept of construct validity must be central to any overall explanation of the findings.

Construct validity considerations were fundamental to the scale development program and yielded dividends in the empirical investigation of the resultant scales. Impulsivity, Risk Taking, and

Self Esteem were conceived as three theoretically distinct constructs. In keeping with the requirements for the substantive component of construct validity, a separate item pool was prepared to represent the content specified by each construct. Item-analysis and item-selection procedures were designed to produce personality scales which possessed high structural validity, discriminated among the three constructs, and elicited minimal amounts of desirability response style. The resultant scales were uniquely associated with the relevant constructs in the item-analysis sample; this was a necessary consequence of the scale-construction procedures. However, the scales remained just as reliable and discriminating in the standard-instruction validation sample; the usual finding with scales not based on a construct approach is a loss in reliability and discrimination. Furthermore, when the scales were cross-validated on a sample of subjects who were instructed to fake, the scales continued to show about the same levels of reliability and discrimination among the constructs. In short, in contrast to other types of scales, the construct scales performed in satisfactory ways independently of sample or instructional set.

The major implication of the fact that each construct produced its own array of individual response consistency regardless of instructional set is that the psychology of personality and the psychology of structured test behaviour are even more closely related than has been realized previously. The study offers unequivocal support for Loevinger's (1957) thesis that psychological tests should be considered instruments of psychological theory.

The findings reported and the interpretation with regard to construct validity add a new complexion to the theory and research in personality and personality assessment.

Implications for Further Research

The research described was concerned with many aspects of scale construction and scale validation. In the context of presenting specific findings the need for experimental explication was often suggested. This final section of the discussion is not concerned with ways of extending specific findings, but is concerned with implications for continuing and broadening the research program of which this has been the beginning.

The fact that each construct determined its own array of individual response consistency under two instructional sets suggests that the research should be broadened to include empirical examination of the construct properties of the scales under yet other instructional sets. The scales, or better, scales representing a wider range of traits, could be administered, for example, with instructions to fake so as to create an unfavourable impression, or with instructions to create the correct impression for applying for particular kinds of jobs. If the construct properties of the scales were maintained under these additional conditions, the conclusions of the preceding section would be further authenticated.

Since attention to construct validity in the construction of scales has produced both endorsement and desirability-judgment scales with favourable construct properties, it seems a distinct possibility that a detailed examination of the characteristics of items shared (and not shared) by endorsement and desirability-judgment scales could lead to even greater degrees of construct validity in both types of scales. In other words, it is suggested that the construct properties of personality scales could be improved by selecting items which are simultaneously good endorsement and desirability-judgment items. Considerable research into the measure of the differences between ideal and less than ideal items is required. Such research could provide much-needed feedback to the task of preparing item pools.

CHAPTER VII

SUMMARY AND CONCLUSIONS

The research concerned the development and validation of a number of experimental scales for measuring the personality traits of Impulsivity, Risk Taking, and Self Esteem. Using the data of a large sample of subjects, four types of experimental scales were developed through item analysis and selection from substantively valid item pools. Nonforced and forced-choice endorsement scales were developed for measuring all three traits. Nonforced scales were composed of self-descriptive statements. Subjects rated each statement as to how characteristic or uncharacteristic it was of self. The forced-choice scales were composed of the same statements as the nonforced scales. In the forced-choice scales, however, each statement was paired with a construct-irrelevant filler statement. Subjects selected the more self-characteristic statement in each pair. Nonforced and forced-choice desirability-judgment scales were prepared for the Impulsivity and Risk Taking constructs. Nonforced scales were composed of specially selected self-descriptive statements. Subjects rated each statement as to how desirable or undesirable it was in other people. The forced-choice desirability-judgment scales contained the same statements as the nonforced scales, but in these scales the statements were paired with construct-irrelevant filler statements. Subjects selected from each pair the statement they judged to be more desirable in others.

The experimental scales were validated against peer behaviour-judgment and self-behaviour-judgment criteria in a new sample of subjects. The four types of scales were comparatively evaluated in terms of reliability, freedom from desirability response style influence, and convergent and discriminant validity. The scales were administered with standard instructions and with instructions to fake in order to compare the different types of scales as to their resistance to faking.

The scales, which were prepared on the basis of the principles of construct validity, were shown to have favourable construct properties in the sample used for item-analysis. The favourable construct properties were not attenuated with cross-validation on either the standard-instruction or fake-instruction validation sample. In both validation samples most of the experimental scales were reliable and were able to discriminate among the three constructs.

The four types of scales varied in average validity against behaviour judgments in different ways in the two validation samples. In the standard-instruction sample nonforced endorsements were most valid, followed by forced-choice endorsements, nonforced desirability judgments, and forced-choice desirability judgments, in that order.

Of the four, only the last had zero validity. In contrast, in the fake-instruction sample all scales had some validity and the differences among the four types of scales in average validity were rather minor. This finding represented a decrease in validity with faking to a nonzero level for the endorsement scales and an increase in validity with faking for the desirability-judgment scales.

The following conclusions may be drawn from the results of the study.

1. Highly reliable item pools for such constructs as Impulsivity, Risk Taking, and Self Esteem can be prepared from considerations of psychological theory, without resort to such methods as empirical keying.

2. The attainment of convergent and discriminant properties of personality scales can be facilitated by item-selection strategies which emphasize structural validity.

3. Substantial correlations between endorsement scales and desirability-judgment scales for such constructs as Impulsivity and Risk Taking can be attained through item analysis and item selection. The corresponding correlations are much lower for constructs which are more extreme on the desirability dimension, such as Self Esteem.

4. Nonforced endorsement and desirability-judgment scales exhibit better construct properties than their derivative forced-choice scales.

5. Endorsement scales suffer some loss in convergent validity when administered with instructions to fake, but, at least for

constructs such as Impulsivity, Risk Taking, and Self Esteem, other construct properties of the scales are maintained with instructions to fake.

6. Desirability-judgment scales of the type examined exhibit higher convergent validity when administered with instructions to fake than when administered with standard instructions. This suggests that desirability-judgment scales may be more valid when presented as maximum-performance tasks rather than as tasks with no right or wrong answers.

7. Personality scale-development strategies of the type adopted produce scales with construct properties that are relatively stable with cross-validation.

8. Personality scale-development strategies of the type adopted produce scales with construct properties that are stable across two instructional sets, instructions to respond honestly, and instructions to respond desirably. This finding suggests that the construct properties of such scales may be stable across yet other instructional sets.

9. Scales developed through a construct approach may be used to make substantial increases in our understanding of both the psychology of personality and the psychology of structured test behaviour. In the present investigation, for example, the relationship between desirability and faking biases was clarified in a way that has not been possible with personality scales lacking in construct validity.

REFERENCES

American Psychological Association. _Standards for educational and psychological tests and manuals_. Washington: American Psychological Association, 1966.

American Psychological Association. _Technical recommendations for psychological tests and diagnostic techniques_. Washington: American Psychological Association, 1954.

Baier, D. E. Reply to Travers' "A critical review of the validity and rationale of the forced-choice technique." _Psychological Bulletin_, 1951, _48_, 421-434.

Bartlett, C. J. Forced-choice response style measurement and instructional sets. Paper presented at the Symposium on Interest Measurement, University of Minnesota, Minneapolis, Minnesota, 1966.

Berg, I. A. (Ed.) _Response set in personality assessment_. Chicago: Aldine, 1967.

Block, J. _The challenge of response sets_. New York: Appleton-Century-Crofts, 1965.

Bloxom, B. Social desirability as a predictor of behavior on neutral social desirability items. _Educational Testing Service, Research Bulletin_, 1967, No. 26.

Boe, E. E. & Kogan, W. S. An analysis of various methods for deriving the social desirability score. Psychological Reports, 1964, 14, 23-29.

Boe, E. E. & Kogan, W. S. Social desirability in individual performance on thirteen MMPI scales. British Journal of Psychology, 1966, 57, 161-170.

Borislow, B. The Edwards Personal Preference Schedule (EPPS) and fakability. Journal of Applied Psychology, 1958, 42, 22-27.

Boruch, R. F. Factor analysis with constraints (or analysis of variance with fewer constraints) and multitrait-multimethod data. Unpublished manuscript, 1968.

Boruch, R. F. & Wolins, L. A procedure for estimating the amount of trait, method, and error variance attributable to a measure. Paper presented at the meeting of the Psychometric Society, Chapel Hill, North Carolina, 1968.

Broverman, D. M. Normative and ipsative measurement in psychology. Psychological Review, 1962, 69, 295-305.

Brozek, J. & Erickson, N. K. Item analysis of the psychoneurotic scales of the Minnesota Multiphasic Personality Inventory in experimental semistarvation. Journal of Consulting Psychology, 1948, 12, 403-411.

Campbell, D. T. The indirect assessment of social attitudes. Psychological Bulletin, 1950, 47, 15-38.

Campbell, D. T. Recommendations for APA test standards regarding construct, trait, or discriminant validity. American Psychologist, 1960, 15, 546-553.

Campbell, D. T. & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Campbell, J. T. & Rundquist, E. A. Scale items for inclusion in forced-choice rating forms. *American Psychologist*, 1950, *5*, 280. (Abstract)

Clemans, W. V. An analytical and empirical examination of some properties of ipsative measures. *Psychometric Monographs*, 1964, No. 14.

Corah, N. L., Feldman, M. J., Cohen, I. S., Gruen, W., Meadow, A., & Ringwall, E. A. Social desirability as a variable in the Edwards Personal Preference Schedule. *Journal of Consulting Psychology*, 1958, *22*, 70-72.

Cronbach, L. J. Response sets and test validity. *Educational and Psychological Measurement*, 1946, *6*, 475-494.

Cronbach, L. J. Proposals leading to analytic treatment of social perception scores. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior*. Stanford, California: Stanford University Press, 1958. Pp. 353-379.

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

Crowne, D. P. & Marlowe, D. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 1960, *24*, 349-354.

Cruse, D. B. Social desirability scale values of personal concepts. *Journal of Applied Psychology*, 1965, *49*, 342-344.

Cruse, D. B. Social desirability scale values of questions and answers. *Journal of General Psychology*, 1967, *77*, 17-30.

Cureton, E. E. Validity, reliability, and baloney. *Educational and Psychological Measurement*, 1950, *10*, 94-96.

Damarin, F. & Messick, S. Response styles as personality variables: a theoretical integration of multivariate research. *Educational Testing Service, Research Bulletin*, 1965, No. 10.

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

Edrich, H., Selltiz, C., & Cook, S. W. The effects of context and of raters' attitudes on judgments of favorableness of statements about a social group. *Journal of Social Psychology*, 1966, *70*, 11-22.

Edwards, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *Journal of Applied Psychology*, 1953, *37*, 90-93.

Edwards, A. L. *Manual for the Edwards Personal Preference Schedule*. New York: Psychological Corporation, 1954.

Edwards, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957.

Edwards, A. L. The social desirability variable: a broad statement. In I. A. Berg (Ed.), *Response set in personality assessment*. Chicago: Aldine, 1967. Pp. 32-47. (a)

Edwards, A. L. The social desirability variable: a review of the evidence. In I. A. Berg (Ed.), *Response set in personality assessment*. Chicago: Aldine, 1967. Pp. 48-70. (b)

Edwards, A. L., Wright, C. E., & Lunneborg, C. E. A note on "Social desirability as a variable in the Edwards Personal Preference Schedule." Journal of Consulting Psychology, 1959, 23, 558.

Feldman, M. J. & Corah, N. L. Social desirability and the forced choice method. Journal of Consulting Psychology, 1960, 24, 480-482.

Ford, L. H., Jr. A forced-choice, acquiescence-free, social desirability (defensiveness) scale. Journal of Consulting Psychology, 1964, 28, 475.

Fox, J. Social desirability, prediction equation, regression equations, and intrinsic response bias. Psychological Bulletin, 1967, 67, 391-400.

Fricke, B. G. Subtle and obvious test items and response set. Journal of Consulting Psychology, 1957, 21, 250-252.

Goldberg, L. R. & Rorer, L. G. Use of two different response modes and repeated testings to predict social conformity. Journal of Personality and Social Psychology, 1966, 3, 28-37.

Goldberg, L. R. & Slovic, P. The importance of test item content: an analysis of a corollary of the Deviation Hypothesis. Journal of Counseling Psychology, 1967, 14, 462-472.

Gordon, L. V. Validities of the forced-choice and questionnaire methods of personality measurement. Journal of Applied Psychology, 1951, 35, 407-412.

Gordon, L. V. Some interrelationships among personality item characteristics. Educational and Psychological Measurement, 1953, 13, 264-272.

Gordon, L. V. & Stapleton, E. S. Fakability of a forced-choice personality test under realistic high school employment conditions. *Journal of Applied Psychology*, 1956, *40*, 258-262,

Guilford, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1965.

Gulliksen, H. *Theory of mental tests*. New York: Wiley, 1950.

Hanley, C. Social desirability and responses to items from three MMPI scales: D, Sc, and K. *Journal of Applied Psychology*, 1956, *40*, 324-328.

Hedberg, R. & Baxter, B. Favorableness ratings of forced-choice statements: applicants vs. non-applicants. *Personnel Psychology*, 1963, *16*, 23-27.

Heineman, C. E. A forced-choice form of the Taylor Anxiety Scale. *Journal of Consulting Psychology*, 1953, *17*, 447-454.

Highland, R. W. & Berkshire, J. R. A methodological study of forced-choice performance rating. *Air Training Command Human Resources Research Center*, *Research Bulletin*, 1951, No. 9.

Horn, J. L. & Cattell, R. B. Vehicles, ipsatization, and the multiple-method measurement of motivation. *Canadian Journal of Psychology*, 1965, *19*, 265-279.

Horst, P. *Factor analysis of data matrices*. New York: Holt, Rinehart, and Winston, 1965.

Izard, C. E. & Rosenberg, N. Effectiveness of a forced-choice leadership test under varied experimental conditions. *Educational and Psychological Measurement*, 1958, *18*, 57-62.

Jackson, D. N. Stylistic response determinants in the California Psychological Inventory. _Educational and Psychological Measurement_, 1960, _20_, 339-346.

Jackson, D. N. Assessing conformity with desirability judgments. _American Psychologist_, 1961, _16_, 446. (Abstract)

Jackson, D. N. Desirability judgments as a method of personality assessment. _Educational and Psychological Measurement_, 1964, _24_, 223-238.

Jackson, D. N. A modern strategy for personality assessment: the Personality Research Form. _University of Western Ontario, Research Bulletin_, 1966, No. 30. (a)

Jackson, D. N. Multimethod factor analysis in the appraisal of convergent and discriminant validity. Paper presented at the meeting of the Society for Multivariate Experimental Psychology, Atlanta, Georgia, 1966. (b)

Jackson, D. N. Acquiescence response styles: problems of identification and control. In I. A. Berg (Ed.), _Response set in personality assessment_. Chicago: Aldine, 1967. Pp. 71-114. (a)

Jackson, D. N. _Personality research form manual_. Goshen, New York: Research Psychologists Press, 1967. (b)

Jackson, D. N. & Messick, S. Content and style in personality assessment. _Psychological Bulletin_, 1958, _55_, 243-252.

Jackson, D. N. & Messick, S. Acquiescence and desirability as response determinants on the MMPI. _Educational and Psychological Measurement_, 1961, _21_, 771-790.

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

Jackson, D. N. & Payne, I. R. Personality scale for shallow affect. *Psychological Reports*, 1963, *13*, 687-698.

Jackson, D. N. & Singer, J. E. Judgments, items, and personality. *Journal of Experimental Research in Personality*, 1967, *2*, 70-79.

Jöreskog, K. G. Confirmatory factor analysis. Paper presented at the meeting of the Psychometric Society, Chapel Hill, North Carolina, 1968.

Jurgensen, C. E. Report on the "Classification Inventory," a personality test for industrial use. *Journal of Applied Psychology*, 1944, *28*, 445-460.

Kristof, W. Orthogonal inter-battery factor analysis. *Psychometrika*, 1967, *32*, 199-227.

Krug, R. E. A selection set preference index. *Journal of Applied Psychology*, 1958, *42*, 168-170. (a)

Krug, R. E. The effect of specific selection sets on a forced-choice self-description inventory. *Journal of Applied Psychology*, 1958, *42*, 89-92. (b)

Kuder, G. F. *Kuder preference record--personal*. Chicago: Science Research Associates, 1948.

Kusyszyn, I. Comparison of judgmental methods with endorsements in the assessment of personality traits. *Journal of Applied Psychology*, 1968, *52*, in press.

Kusyszyn, I. & Jackson, D. N. A multimethod factor analytic appraisal of endorsement and judgment methods in personality assessment. *Educational and Psychological Measurement*, 1968, *28*, in press.

La Pointe, R. E. & Auclair, G. A. The use of social desirability in forced-choice methodology. *American Psychologist*, 1961, *16*, 446. (Abstract)

Loevinger, J. Objective tests as instruments of psychological theory. *Psychological Reports*, 1957, *3*, 635-694.

Longstaff, H. P. & Jurgensen, C. E. Fakability of the Jurgensen Classification Inventory. *Journal of Applied Psychology*, 1953, *37*, 86-89.

Loomis, R. & Spilka, B. Social desirability and conformity in a test situation. *American Psychologist*, 1963, *18*. 358. (Abstract)

Lorge, I. Gen-like: Halo or reality? *Psychological Bulletin*, 1937, *34*, 545-546. (Abstract)

Magnusson, D. *Test theory*. Reading, Mass.: Addison-Wesley, 1967.

Maher, H. Studies of transparency in forced-choice scales: I. Evidence of transparency. *Journal of Applied Psychology*, 1959, *43*, 275-278.

Mais, R. D. Fakability of the Classification Inventory scored for self confidence. *Journal of Applied Psychology*, 1951, *35*, 172-174.

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

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

Meehl, P. E. & Hathaway, S. R. The K factor as a suppressor variable in the Minnesota Multiphasic Personality Inventory. Journal of Applied Psychology, 1946, 30, 525-564.

Messick, S. Dimensions of social desirability. Journal of Consulting Psychology, 1960, 24, 279-287.

Messick, S. Response style and content measures from personality inventories. Educational and Psychological Measurement, 1962, 22, 41-56.

Messick, S. Desirability judgments and inventory responses in the assessment of personality. Educational Testing Service, Research Memorandum, 1964, No. 13.

Messick, S. Psychology and methodology of response styles. Paper presented at the meeting of the Western Psychological Association, Honolulu, 1965.

Messick, S. The psychology of acquiescence: an interpretation of research evidence. In I. A. Berg (Ed.), Response set in personality assessment. Chicago: Aldine, 1967. Pp. 115-145.

Messick, S. & Jackson, D. N. Desirability scale values and dispersions for MMPI items. Psychological Reports, 1961, 8, 409-414.

Neill, J. A. & Jackson, D. N. An empirical evaluation of item-selection techniques in personality assessment. University of Western Ontario, Research Bulletin, 1967, No. 59.

Norman, W. T. Personality measurement, faking, and detection: an assessment method for use in personnel selection. Journal of Applied Psychology, 1963, 47, 225-241. (a)

Norman, W. T. Relative importance of test item content. _Journal of Consulting Psychology_, 1963, _27_, 166-174. (b)

Norman, W. T. On estimating psychological relationships: social desirability and self-report. _Psychological Bulletin_, 1967, _67_, 273-293.

Nunnally, J. C. _Tests and measurements, assessment and prediction_. New York: McGraw-Hill, 1959.

Peabody, D. Authoritarianism scales and response bias. _Psychological Bulletin_, 1966, _65_, 11-23.

Peters, C. C. & Van Voorhis, W. R. _Statistical procedures and their mathematical bases_. New York: McGraw-Hill, 1940.

Radcliffe, J. A. Some properties of ipsative score matrices and their relevance for some current interest tests. _Australian Journal of Psychology_, 1963, _15_, 1-11.

Radcliffe, J. A. Review of Edwards Personal Preference Schedule. In O. K. Buros (Ed.), _The Sixth Mental Measurements Yearbook_. Highland Park, New Jersey: Gryphon Press, 1965. Pp. 195-200.

Richardson, M. W. An empirical study of the forced-choice performance report. _American Psychologist_, 1949, _4_, 278-279 (Abstract). (a)

Richardson, M. W. Forced-choice performance reports: a modern merit-rating method. _Personnel_, 1949, _26_, 205-212. (b)

Rosen, E. Self-appraisal, personal desirability, and perceived social desirability of personality traits. _Journal of Abnormal and Social Psychology_, 1956, _52_, 151-158.

Rusmore, J. T. Fakability of the Gordon Personal Profile. _Journal of Applied Psychology_, 1956, _40_, 175-177.

Saltz, E., Reece, M., & Ager, J. Studies of forced-choice methodology: individual differences in social desirability. Educational and Psychological Measurement, 1962, 22, 365-370.

Schanberger, W. J. A motivational distortion scale for the Sixteen Personality Factor Questionnaire, Form A. Paper presented at the meeting of the Western Psychological Association, San Francisco, California, 1967.

Scott, W. A. Social desirability and individual conceptions of the desirable. Journal of Abnormal and Social Psychology, 1963, 67, 574-585.

Seeman, W. "Subtlety" in structured personality tests. Journal of Consulting Psychology, 1952, 16, 278-283.

Selltiz, C. & Cook, S. W. Racial attitude as a determinant of judgments of plausibility. Journal of Social Psychology, 1966, 70, 139-147.

Selltiz, C., Edrich, H., & Cook, S. W. Ratings of favorableness of statements about a social group as an indicator of attitude toward the group. Journal of Personality and Social Psychology, 1965, 2, 408-415.

Shipley, W. C., Gray, F. E., & Newbert, N. The Personal Inventory--its derivation and validation. Journal of Clinical Psychology, 1946, 2, 318-322.

Siess, T. F. & Jackson, D. N. A personalogical approach to the interpretation of vocational interests. In Proceedings of the 75th Annual Convention of the American Psychological Association. Washington: American Psychological Association, 1967. Pp. 353-354.

Sisson, D. Forced choice--the new army rating. Personnel Psychology, 1948, 1, 365-381.

Stricker, L. J. Acquiescence and social desirability response styles, item characteristics, and conformity. Psychological Reports, 1963, 12, 319-341.

Stricker, L. J. Review of Edwards Personal Preference Schedule. In O. K. Buros (Ed.), The Sixth Mental Measurements Yearbook. Highland Park, New Jersey: Gryphon Press, 1965. Pp. 200-207.

Stricker, L. J., Messick, S., & Jackson, D. N. Judgmental consistencies as predictors of social behavior. Educational Testing Service, Research Memorandum, 1966, No. 17.

Stricker, L. J., Messick, S., & Jackson, D. N. Desirability judgments and self-reports as predictors of social behavior. University of Western Ontario, Research Bulletin, 1968, No. 79.

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

Travers, R. M. W. A critical review of the validity and rationale of the forced-choice technique. Psychological Bulletin, 1951, 48, 62-70. (a)

Travers, R. M. W. Rational hypotheses in the construction of tests. Educational and Psychological Measurement, 1951, 11, 128-137. (b)

Tucker, L. R. An inter-battery method of factor analysis. Psychometrika, 1958, 23, 111-136.

Waly, P. & Cook, S. W. Effect of attitude on judgments of plausibility. Journal of Personality and Social Psychology, 1965, 2, 745-749.

Waters, L. K. & Wherry, R. J., Jr. The effect of intent to bias on forced-choice indices. Personnel Psychology, 1962, 15, 207-214.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. Unobtrusive measures, nonreactive research in the social sciences. Chicago: Rand McNally, 1966.

Wiggins, J. S. Strategic, method,and stylistic variance in the MMPI. Psychological Bulletin, 1962, 59, 224-242.

Wiggins, J. S. Social desirability estimation and "faking good" well. Educational and Psychological Measurement, 1966, 26, 329-341.

Wiggins, N. Individual viewpoints of social desirability. Psychological Bulletin, 1966, 66, 68-77.

Winters, S., Bartlett, C. J., & Leve, R. Instructional and response style factors with forced-choice response. Paper presented at the meeting of the American Psychological Association, Chicago, Illinois, 1965.

Zavala, A. Development of the forced-choice rating scale technique. Psychological Bulletin, 1965, 63, 117-124.

Zavalloni, M. & Cook, S. W. Influence of judges' attitudes on ratings of favorableness of statements about a social group. Journal of Personality and Social Psychology, 1965, 1, 43-54.

APPENDIX A

Instructions to the Item-analysis Sample

Appendix A

Endorsement Instructions

The booklet given you contains a number of statements that a person might make in describing himself or in expressing an opinion.

You are to judge to what degree each statement applies to you. That is, you are to judge how characteristic or uncharacteristic each statement is of you.

Look at the following statement. "I believe in mercy killing." You might judge this item to be extremely characteristic of you, slightly characteristic, or perhaps moderately uncharacteristic of you.

Indicate your judgment about how characteristic or uncharacteristic each item is of you by writing one of the numbers 1 to 9 to the right of each item number on the "Record of Judgments" sheet. The numbers 1 to 9 will tell to what degree each statement is characteristic or uncharacteristic of you, as indicated below.

9 Extremely characteristic

8 Very characteristic

7 Moderately characteristic

6 Slightly characteristic

5 Neutral

4 Slightly uncharacteristic

3 Moderately uncharacteristic

2 Very uncharacteristic

1 Extremely uncharacteristic

For example, someone has indicated how characteristic or uncharacteristic each of three statement is of himself:

Booklet Statement

A. I have the use of only one leg.
B. I enjoy stories about the sea.
C. Social contact between friends is healthy.

Record of Judgments

A.__1__ B.__5__ C.__9__

The person who made these judgments felt that the statement "I have the use of only one leg" was extremely uncharacteristic of him; that the statement "I enjoy stories about the sea" was neutral when applied to him; and that the statement "Social contact between friends is healthy" was extremely characteristic of him. Your judgments might be different.

Indicate in the same manner on the Record of Judgments sheet the degree to which every statement is characteristic or uncharacteristic of you.

Be sure that the number next to each statement in the booklet is the same as the number in the Record of Judgments sheet.

Be sure to make a judgment about every statement.

Keep these instructions in front of you when making your judgments to help you remember what the numbers mean.

Appendix A (continued)

Desirability-judgment Instructions

The booklet given you contains a number of statements to which a person might respond "true" or "false" in a personality questionnaire. Each response to a statement reflects certain tendencies, preferences, or traits of the person making it.

You are to judge the degree to which a true response to each statement would reflect a desirable or undesirable characteristic. You should judge how desirable or undesirable a true response to these statements would be in other people, not how desirable a true response would be for you.

Look at the following statement "I believe in mercy killing." You might judge it to be extremely desirable for other people to answer true to this item. Or you might judge it to be neutral or extremely undesirable for people to answer true.

Indicate your judgment of every item by writing one of the numbers 1 to 9 to the right of each item number on the "Record of Judgments" sheet. The numbers 1 to 9 represent different degrees of desirability of a true response by other people, as indicated below.

9. Extremely desirable

8. Very desirable

7. Moderately desirable

6. Slightly desirable

5. Neutral

4. Slightly undesirable

3. Moderately undesirable

2. Very undesirable

1. Extremely undesirable.

For example, someone has indicated below his estimate of the desirability of undesirability or a true response to each of three statements.

Booklet Statement	Record of Judgments
A. I often feel like punishing my enemies.	A. 2 B. 5 C. 7
B. I like to read psychological novels.	
C. Social contact between friends is healthy.	

The person who made these judgments believes that a true response to "I often feel like punishing my enemies" is very undesirable; A true response to "I like to read psychological novels" is neither particularly desirable nor undesirable in other people; and a true response to "Social contact between friends is healthy" is moderately desirable. You might feel otherwise and might wish to answer differently.

Indicate in the same manner on the Record of Judgments sheet your own judgments of the desirability or the undesirability of a true response to every statement which appears in the booklet. In making your judgments try to use all nine gradations of desirability.

Be sure the number next to each statement is the same as the number on the Record of Judgment sheet.

Be sure to make a judgment about every statement.

Remember, you are to judge the characteristic implied by a true response to each statement in terms of whether you consider it desirable or undesirable in others.

Keep these instructions in front of you while making your judgments to help you remember what the numbers mean

APPENDIX B

Selected Materials for the Validation Samples[1]

[1] The 20-item scales are not reproduced here, but are available for bona-fide research purposes. Inquiries should be directed to the Department of Psychology, University of Western Ontario, London, Canada.

Appendix B

Behaviour Descriptions

Positive Impulsivity

Tends to do things on the "spur of the moment" and without deliberation; speaks without hesitation; gives vent readily to feelings and wishes; will follow an impulse, even to do something silly; quick-thinking; spontaneous.

Negative Impulsivity

Tends to speak slowly and deliberately; likes to plan something in detail before doing it; rarely buys something on an impulse; not impetuous or rash; not easily excited; controlled; reserved; patient.

Positive Risk Taking

Enjoys taking risks, especially if the possible gains are high; would probably like gambling; would accept an insecure job for the sake of potentially higher rewards; would invest borrowed money on the chance of making a profit; not bothered by danger; carefree.

Negative Risk Taking

Tends to avoid situations involving personal risk, even when the rewards could be great; cautious about situations with uncertain outcomes; not likely to place a bet; more likely to save money than to invest it; doesn't take chances; security-minded; conservative.

Positive Self Esteem

Tends to be comfortable in most social situations; is not made uneasy by being the centre of attention; makes a good first impression; composed and self-possessed in her surroundings; has high self-regard; poised; self-confident.

Negative Self Esteem

Tends to feel awkward when with a group of people, especially if strangers are present; feels ill at ease socially; prefers to remain unnoticed at social events; has a low opinion of herself as a group member; easily flustered; self-conscious.

Instructions for the PRF-Dy Scale

FORM A

The following 2 pages contain a number of statements that a person might make in describing himself or in expressing an opinion.

You are to judge to what degree each statement applies to you. That is, you are to judge how characteristic or uncharacteristic each statement is of you.

Look at the following statement. "I believe in mercy killing." You might judge this item to be extremely characteristic of you, slightly characteristic, or perhaps moderately uncharacteristic of you.

Indicate your judgment about how characteristic or uncharacteristic each item is of you by writing one of the numbers 1 to 9 to the right of each item number under "Form A" on the Answer Sheet. The numbers 1 to 9 will tell to what degree each statement is characteristic or uncharacteristic of you, as indicated below.

9 Extremely characteristic

8 Very characteristic

7 Moderately characteristic

6 Slightly characteristic

5 Neutral

4 Slightly uncharacteristic

3 Moderately uncharacteristic

2 Very uncharacteristic

1 Extremely uncharacteristic

For example, someone has indicated how characteristic or uncharacteristic each of three statements is of himself:

Statement

X. I have the use of only one leg.
Y. I enjoy stories about the sea.
Z. Social contact between friends is healthy.

Answer Sheet -- FORM A

X. 1 Y. 5 Z. 9

The person who made these judgments felt that the statement "I have the use of only one leg" was extremely uncharacteristic of him; that the statement "I enjoy stories about the sea" was neutral when applied to him; and that the statement "Social contact between friends is healthy" was extremely characteristic of him. Your judgments might be different.

Indicate in the same manner on the Answer Sheet the degree to which every statement is characteristic or uncharacteristic of you.

Be sure that the number next to each statement is the same as the number on the Answer Sheet.

Appendix B (continued)

Instructions for the Behaviour-ranking Task

FORM B

Look at the envelope marked Form B. A list of names of some women including yourself, appears on the outside. Find your name and circle it.

Beside the list of names are three behavior descriptions marked:

BEHAVIOUR DESCRIPTION 1
BEHAVIOUR DESCRIPTION 2
BEHAVIOUR DESCRIPTION 3

You are to rank the women on your list according to how characteristic each description is of them.

Inside the envelope you will find three decks of IBM cards: a blue deck marked "BEHAVIOUR DESCRIPTION 1", an orange deck marked "BEHAVIOUR DESCRIPTION 2", and a yellow deck marked "BEHAVIOUR DESCRIPTION 3". Each deck contains one card for each person on your list.

You are to rank the persons on your list by rearranging the cards in each deck as follows. First, study BEHAVIOUR DESCRIPTION 1 until you have a clear idea of it. Then, using the blue deck, find the card for the person on your list of whom BEHAVIOUR DESCRIPTION 1 is most characteristic. Write the word "most" on that card and put it on top of the rest of the deck. Next find the card for the person of whom BEHAVIOUR DESCRIPTION 1 is least characteristic. Write the word "least" on that card and put it on the bottom of the deck. Then, put the card for the person of whom the description is the second most characteristic under the first card. Continue until all the individuals on your list are in order, starting with the person of whom BEHAVIOUR DESCRIPTION 1 is most characteristic and ending with the person of whom it is least characteristic. Then replace the elastic and put the completed deck into the envelope.

Follow the same procedure for BEHAVIOUR DESCRIPTION 2, using the orange deck. Finally, do the same with BEHAVIOUR DESCRIPTION 3, using the yellow deck.

If you are not sure which of two persons a behaviour description fits best, make your best guess. DO NOT leave anyone out of the ranking.

When you have finished ranking all three decks, write "complete" in the box on the envelope and go on to FORM C.

NOTE: All information you give in this experiment is entirely confidential and will be seen only by the experimenter. Only code numbers and not actual names will be used in the analysis.

Appendix B (continued)

Instructions for the Behaviour-rating Task

FORM C

The same list of names as used in FORM B appears on the answer sheet for FORM C. Find your name on the list and circle it.

Because it is very improbable that you will know each woman equally well, you are asked to indicate how well you know each person by using one of the numbers 1 to 9 : a 9 would mean that you know her extremely well, and a 1 would mean that you don't know her at all.

extremely well								not at all
9	8	7	6	5	4	3	2	1

For each person you can indicate your degree of acquaintance in the column entitled "degree of acquaintance", immediately to the right of the names. Please complete the "degree of acquaintance" section NOW, before going on.

You will find a list of six "BEHAVIOUR DESCRIPTIONS" on the answer sheet. You are to judge the degree to which each behaviour description is characteristic or uncharacteristic of the women whose names are on your list.

You are to use one of the numbers 1 to 9 to indicate how characteristic or uncharacteristic each behaviour description is of each person. The numbers 1 to 9 will tell the degree which you judge each behaviour description to be characteristic or uncharacteristic of a person, as indicated below.

9 Extremely characteristic
8 Very characteristic
7 Moderately characteristic
6 Slightly characteristic
5 Neutral
4 Slightly uncharacteristic
3 Moderately uncharacteristic
2 Very uncharacteristic
1 Extremely uncharacteristic

For example, if you felt that BEHAVIOUR DESCRIPTION 3 was very characteristic of Jane Doe, you would write the number 8 in column 3 opposite her name on the Answer Sheet. If you judged the same behaviour description to be extremely uncharacteristic of Jill Doe, you would write a 1 in column 3 opposite Jill Doe's name. When making your judgments, try to use all nine gradations of the scale.

You should proceed with your judgments as follows. First, study BEHAVIOUR DESCRIPTION 1 until you have a clear idea of it. Then, in column 1, write your judgment about how characteristic BEHAVIOUR DESCRIPTION 1 is of the first person on your list. Then go on and judge how characteristic the description is of the second person on your list. When you have indicated for BEHAVIOUR DESCRIPTION 1 your judgment about every person, including yourself, study BEHAVIOUR DESCRIPTION 2 and proceed in the same manner.

Continue to make judgments about all the individuals for each behaviour description in turn until you have made a judgment about every person for all six behaviour descriptions.

NOTE: As soon as the experiment is completed, the name lists will be cut from the answer sheets and only the code number will be used in the analysis.

Response Sheet for the Behaviour-rating Task

FORM C

BEHAVIOUR DESCRIPTION 1

Tends to avoid situations involving personal risk, even when the rewards could be great; cautious about situations with uncertain outcomes; not likely to place a bet; more likely to save money than to invest it; doesn't take chances; security-minded; conservative.

BEHAVIOUR DESCRIPTION 2

Tends to be comfortable in most social situations; is not made uneasy by being the centre of attention; makes a good first impression; composed and self-possessed in her surroundings; has high self-regard; poised; self-confident.

BEHAVIOUR DESCRIPTION 3

Tends to speak slowly and deliberately; likes to plan something in detail before doing it; rarely buys something on an impulse; not impetuous or rash; not easily excited; controlled; reserved; patient.

BEHAVIOUR DESCRIPTION 4

Enjoys taking risks, especially if the possible gains are high; would probably like gambling; would accept an insecure job for the sake of potentially higher rewards; would invest borrowed money on the chance of making a profit; not bothered by danger; carefree.

BEHAVIOUR DESCRIPTION 5

Tends to feel awkward when with a group of people, especially if strangers are present; feels ill at ease socially; prefers to remain unnoticed at social events; has a low opinion of herself as a group member; easily flustered; self-conscious.

BEHAVIOUR DESCRIPTION 6

Tends to do things on the "spur of the moment" and without deliberation; speaks without hesitation; gives vent readily to feelings and wishes; will follow an impulse, even to do something silly; quick-thinking; spontaneous.

		degree of acquaintance	BEHAVIOUR DESCRIPTION					
			1	2	3	4	5	6
MISS [redacted]	2601							
MISS [redacted]	2603							
MISS [redacted]	2605							
MISS [redacted]	2607							
MISS [redacted]	2609							
MISS [redacted]	2611							
MISS [redacted]	2613							
MISS [redacted]	2615							
MISS [redacted]	2617							
MISS [redacted]	2619							

9	Extremely characteristic
8	Very characteristic
7	Moderately characteristic
6	Slightly characteristic
5	Neutral
4	Slightly uncharacteristic
3	Moderately uncharacteristic
2	Very uncharacteristic
1	Extremely uncharacteristic

Appendix B (continued)

Instructions for the Nonforced Endorsement Task

FORM D

The following 4 pages contain a number of statements that a person might make in describing himself or in expressing an opinion.

You are to judge to what degree each statement applies to you. That is, you are to judge how characteristic or uncharacteristic each statement is of you.

Look at the following statement. "I believe in mercy killing." You might judge this item to be extremely characteristic of you, slightly characteristic, or perhaps moderately uncharacteristic of you.

Indicate your judgment about how characteristic or uncharacteristic each item is of you by writing one of the numbers 1 to 9 to the right of each item number under "Form D" on the Answer Sheet. The numbers 1 to 9 will tell to what degree each statement is characteristic or uncharacteristic of you, as indicated below.

9 Extremely characteristic

8 Very characteristic

7 Moderately characteristic

6 Slightly characteristic

5 Neutral

4 Slightly uncharacteristic

3 Moderately uncharacteristic

2 Very uncharacteristic

1 Extremely uncharacteristic

For example, someone has indicated how characteristic or uncharacteristic each of three statements is of himself:

Statement	Answer Sheet -- FORM D
	X. 1 Y. 5 Z. 9

X. I have the use of only one leg.
Y. I enjoy stories about the sea.
Z. Social contact between friends is healthy.

The person who made these judgments felt that the statement "I have the use of only one leg" was extremely uncharacteristic of him; that the statement "I enjoy stories about the sea" was neutral when applied to him; and that the statement "Social contact between friends is healthy" was extremely characteristic of him. Your judgments might be different.

Indicate in the same manner on the Answer Sheet the degree to which every statement is characteristic or uncharacteristic of you.

Be sure that the number next to each statement is the same as the number on the Answer Sheet.

Appendix B (continued)

Instructions for the Forced-choice Endorsement Task

FORM E

The following 4 pages contain a number of pairs of statements. You are to judge which of the two statements in each pair is more characteristic of you.

Indicate your judgment about which statement is more characteristic of you by writing either the letter A or the letter B to the right of each item number under "Form E" on the answer sheet. The letters "A" and "B" refer to the two statements in each pair, as indicated by the following examples.

<u>Pair of Statements</u> <u>Answer Sheet-Form E</u>

X. A. I like to read psychological novels. X. A Y. B
   B. Social contact between friends is healthy.

Y. A. Going barefoot in the cool grass is great fun.
   B. I don't tire easily.

The person who made these judgments felt that the statement "I like to read psychological novels" was more characteristic of her than the statement social contact between friends is healthy"; and the statement "I don't tire easily" was more characteristic of her than the statement "Going barefoot in the cool grass is great fun." Your judgments might be different.

Indicate in the same manner on the Answer Sheet which statement in each pair is more characteristic of you. If both statements describe you, select the one which is <u>more nearly</u> characteristic of you, or is <u>more often</u> correct. If neither statement describes you well, select the one which is <u>more nearly</u> characteristic of you, or is <u>more often</u> correct.

Be sure that the number next to each pair of statements is the same as the number on the Answer Sheet.

Instructions for the Nonforced Desirability-judgment Task

FORM F

The following 3 pages contain a number of statements to which a person might respond "true" or "false" in a personality questionnaire. Each response to a statement reflects certain tendencies, preferences, or traits of the person making it.

You are to judge the degree to which a true response to each statement would reflect a desirable or undesirable characteristic. You should judge how desirable or undesirable a true response to these statements would be in other women, not how desirable a true response would be for you.

Look at the following statement. "I believe in mercy killing." You might judge it to be extremely desirable for other women to answer true to this item. Or you might judge it to be neutral or extremely undesirable for women to answer true.

Indicate your judgment of every item by writing one of the numbers 1 to 9 to the right of each item number under "Form F" on the Answer Sheet. The numbers 1 to 9 represent different degrees of desirability of a true response by other women, as indicated below.

9	Extremely desirable
8	Very desirable
7	Moderately desirable
6	Slightly desirable
5	Neutral
4	Slightly undesirable
3	Moderately undesirable
2	Very undesirable
1	Extremely undesirable

For example, someone has indicated below her estimate of the desirability or undesirability of a true response to each of three statements.

Statement	Answer Sheet - Form F
X. I often feel like punishing my enemies.	X. 2 Y. 5 Z. 7
Y. I like to read psychological novels.	
Z. Social contact between friends is healthy.	

The person who made these judgments believes that a true response to "I often feel like punishing my enemies" is very undesirable; A true response to "I like to read psychological novels" is neither particularly desirable nor undesirable in other women; and a true response to "Social contact between friends is healthy" is moderately desirable. You might feel otherwise and might wish to answer differently.

Indicate in the same manner on the Answer Sheet your own judgments of the desirability or the undesirability of a true response to every statement which appears on the next 3 pages. In making your judgments try to use all nine gradations of desirability.

Be sure the number next to each statement is the same as the number on the Answer Sheet.

Remember, you are to judge the characteristic implied by a true response to each statement in terms of whether you consider it desirable or undesirable in other women.

Appendix B (continued)

Instructions for the Forced-choice Desirability-judgment Task

FORM G

The following 3 pages contain a number of pairs of statements. You are to judge which statement in each pair of statements is more desirable. You should judge which statement would be more desirable in <u>other women</u>, not which would be more desirable for you.

Indicate your judgment about which statement is more desirable in other women by writing either the letter A or the letter B to the right of each item number under "Form G" on the answer sheet. The letters "A" and "B" refer to the two statements in each pair, as indicated by the following example.

<u>Pair of Statements</u>

X. A. I like to read psychological novels.
B. Social contact between friends is healthy.

Y. A. Going barefoot in the cool grass is great fun.
B. I don't tire easily.

<u>Answer Sheet-Form G</u>

X. B Y. A

The person who made these judgments felt that the statement "Social contact between friends is healthy", was more desirable in women than the statement "I like to read psychological novels"; and that the statement "Going barefoot in the cool grass is great fun", was more desirable than the statement "I don't tire easily". Your judgments might be different.

Indicate in the same manner on the Answer Sheet your own judgment about which statement in each pair is more desirable in other women.

Be sure that the number next to each pair of statements is the same as the number on the Answer Sheet.

Appendix B (continued)

Instructions to Fake

NOTE: Read the following special instructions before beginning FORM D. The special instructions apply to all four remaining tasks (FORM D, FORM E, FORM F, AND FORM G).

A person tends to convey a certain impression of herself by the way she makes judgments about herself, about other people, and about personality statements. In the balance of the experiment, your task is not the usual one of trying to give an accurate impression of yourself. Instead, you are to attempt to convey a very favourable impression of yourself by the judgments you make. In other words you should try to make yourself look good. In order to create a very favourable impression, you may occasionally have to disregard what you really believe or what you are really like.

Since half the women participating in this experiment do not receive these special instructions, it is important that you do not discuss this aspect of the experiment with anyone until the experiment is completely over. If, for your own information, you later wish to try FORMS D, E, F, and G (or a different personality questionnaire) in the normal manner, you will be given the opportunity to do so.

NOTE: You should follow the standard instructions for FORMS D, E, F, and G in every way except that you are to try to make yourself look good by the answers you give. Now go ahead with FORM D.

APPENDIX C

Tables from the Multimethod Factor Analysis

Appendix C, Table A

Intercorrelations among the Personality Scales and the Behaviour-judgment Measures

		Nonforced Endorsements			Forced-choice Endorsements			Nonforced Judgments		Forced-choice Judgments		Self Rank			Self Rating			Peer Ranks			Peer Ratings			
		Im	RT	SE	Im	RT	SE	Im	RT	Im	RT	Im	RT	SE	Im	RT	SE	Im	RT	SE	Im	RT	SE	Des
Nonforced Endorsements																								
Impulsivity	Im				80	40	03	54	22	23	-06	65	59	08	73	59	19	55	60	15	53	60	24	-37
Risk Taking	RT				39	62	07	45	58	06	17	37	62	16	42	65	22	43	48	12	42	50	28	-03
Self Esteem	SE				20	03	82	25	12	-09	06	31	12	68	17	25	87	37	35	43	35	41	47	31
Forced-choice Endorsements																								
Impulsivity	Im	69	34	-07				45	18	22	-06	62	48	16	63	45	20	47	50	10	43	51	19	-33
Risk Taking	RT	43	71	-03				34	46	10	20	18	34	03	37	40	00	16	17	05	13	18	05	-05
Self Esteem	SE	-25	04	67				07	04	-11	01	16	02	69	-06	17	68	22	20	37	17	26	36	38
Nonforced Judgments																								
Impulsivity	Im	66	50	-02	55	40	01			35	07	38	31	06	48	32	18	11	15	05	15	19	10	-14
Risk Taking	RT	35	64	01	17	57	08			15	48	15	22	10	23	25	09	11	14	09	08	17	20	-12
Forced-choice Judgments																								
Impulsivity	Im	47	33	01	38	26	-01	68	48			-01	-02	-12	25	09	-17	-10	-11	-06	-19	-06	-06	-06
Risk Taking	RT	21	36	05	09	38	12	39	69			05	-11	03	09	-01	05	-01	01	-06	-09	-05	04	01
Self Rank																								
Impulsivity	Im	31	31	10	36	15	18	36	24	22	21				66	53	31	52	56	04	53	48	18	-30
Risk Taking	RT	22	43	17	22	37	12	25	27	16	07				42	77	10	52	53	14	48	47	29	-22
Self Esteem	SE	-03	04	51	08	08	47	04	03	-09	06				-03	16	74	28	23	45	22	25	45	26
Self Rating																								
Impulsivity	Im	44	33	18	51	30	16	39	24	25	24	62	34	25				50	46	32	46	48	10	-26
Risk Taking	RT	23	53	19	24	50	20	23	35	22	24	33	77	25				50	50	25	50	58	34	-12
Self Esteem	SE	15	13	43	16	11	47	24	14	13	07	20	22	74				32	45	44	38	40	44	32
Peer Ranks																								
Impulsivity	Im	29	20	13	30	15	22	22	07	19	10	44	42	31	46	56	38				90	86	49	-07
Risk Taking	RT	34	22	19	33	22	19	39	19	34	22	44	48	41	51	60	35				86	88	45	-12
Self Esteem	SE	15	04	21	11	-16	20	17	-05	22	04	22	19	37	03	19	37				21	23	84	12
Peer Ratings																								
Impulsivity	Im	30	22	14	32	19	24	29	12	25	14	47	42	31	52	51	36	92	86	42				-12
Risk Taking	RT	27	16	18	26	16	19	36	19	29	19	41	48	38	48	50	40	84	91	47				-14
Self Esteem	SE	29	21	31	32	02	27	32	12	26	13	34	29	52	49	39	59	53	59	85				06
Desirability	Des	-10	01	25	-07	-00	23	-14	-10	-10	-05	-12	-11	49	-17	-02	55	-04	03	08	-06	-04	20	

Note.-- The correlations for the standard-instruction sample (N = 81) appear above the major diagonal and those for the fake-instruction sample (N = 81) appear below the diagonal. Monomethod correlations have been omitted.

Appendix C, Table B

Unrotated Multimethod Factor Matrix for the Standard-instruction Sample

Method	Factor					
	I	II	III	IV	V	VI
Measures of Impulsivity						
Nonforced Endorsements	.76	-.39	-.23	.23	.19	.08
Forced-choice Endorsements	.71	-.30	.25	.26	.12	.08
Nonforced Desirability Judgments	.47	-.23	.15	.50	.17	-.38
Forced-choice Desirability Judgments	.04	-.31	.22	.55	.31	-.31
Self Rankings	.69	-.20	-.24	.28	-.26	.21
Self Ratings	.64	-.42	-.07	.39	-.13	.11
Peer Rankings	.75	.03	-.21	-.23	-.14	-.01
Peer Ratings	.70	-.03	-.31	-.28	-.25	.06
Measures of Risk Taking						
Nonforced Endorsements	.65	-.17	.54	-.27	-.11	-.18
Forced-choice Endorsements	.39	-.25	.62	-.03	-.07	-.19
Nonforced Desirability Judgments	.34	-.11	.71	.01	-.19	.31
Forced-choice Desirability Judgments	.04	-.01	.57	.19	-.43	.37
Self Rankings	.67	-.26	.10	-.44	.30	-.20
Self Ratings	.74	-.15	.14	-.40	.20	-.20
Peer Rankings	.77	-.06	-.18	-.18	-.14	.08
Peer Ratings	.76	-.03	.14	-.10	-.14	-.14
Measures of Self Esteem						
Nonforced Endorsements	.54	.73	-.04	.22	-.13	-.09
Forced-choice Endorsements	.36	.76	.01	.13	-.09	-.18
Self Rankings	.40	.73	.06	.10	-.00	-.04
Self Ratings	.51	.72	-.07	.18	-.19	-.05
Peer Rankings	.33	.52	.14	.02	.59	.32
Peer Ratings	.48	.47	.18	-.04	.54	.40
PRF-Dy	-.15	.58	.16	-.03	-.13	-.42
Eigenvalue	7.30	3.78	2.06	1.66	1.46	1.20

Appendix C, Table C

Unrotated Multimethod Factor Matrix for the Fake-instruction Sample

	Factor					
Method	I	II	III	IV	V	VI
Measures of Impulsivity						
Nonforced Endorsements	.56	-.39	-.33	.32	-.10	-.22
Forced-choice Endorsements	.55	-.25	-.41	.23	-.02	-.12
Nonforced Desirability Judgments	.62	-.35	-.22	.37	-.13	-.14
Forced-choice Desirability Judgments	.51	-.34	-.15	.34	-.22	-.04
Self Rankings	.59	-.10	-.24	-.04	.48	.04
Self Ratings	.68	-.10	-.31	-.00	.44	.11
Peer Rankings	.66	.11	-.21	-.49	.15	-.04
Peer Ratings	.69	.07	-.17	-.43	.17	-.16
Measures of Risk Taking						
Nonforced Endorsements	.55	-.38	.50	.09	-.02	.16
Forced-choice Endorsements	.47	-.44	.55	.06	-.07	-.13
Nonforced Desirability Judgments	.45	-.41	.57	.21	.21	.08
Forced-choice Desirability Judgments	.38	-.27	.43	.21	.41	.04
Self Rankings	.59	-.10	.19	-.42	-.44	.21
Self Ratings	.67	-.09	.32	-.41	-.33	.23
Peer Rankings	.76	.09	-.07	-.14	-.16	-.34
Peer Ratings	.69	.11	-.12	-.32	-.14	-.18
Measures of Self Esteem						
Nonforced Endorsements	.29	.59	.22	.03	.19	.01
Forced-choice Endorsements	.31	.58	.28	-.06	.33	.02
Self Rankings	.43	.71	.16	.12	-.00	-.18
Self Ratings	.52	.63	.11	.29	-.14	-.25
Peer Rankings	.43	.42	-.27	.24	-.13	.62
Peer Ratings	.63	.40	-.15	.37	-.11	.48
PRF-Dy	.03	.56	.28	.32	-.19	-.33
Eigenvalue	6.94	3.35	2.15	1.79	1.35	1.23

Appendix C (continued), Table D

Rotated Multimethod Factor Matrices

Standard-instruction and Fake-instruction Samples

	Standard-instruction Sample			Fake-instruction Sample		
Method Factor	I	II	III	I	II	III
Measures of Impulsivity						
Nonforced Endorsements	88	-07	08	71	22	18
Forced-choice Endorsements	81	-01	03	72	-12	05
Nonforced Desirability Judgments	44	-02	32	68	-12	28
Forced-choice Desirability Judgments	06	-25	28	55	-14	27
Self Rankings	76	07	01	63	07	13
Self Ratings	73	-14	20	74	09	12
Peer Rankings	72	30	01	63	30	10
Peer Ratings	73	23	-09	63	28	16
Measures of Risk Taking						
Nonforced Endorsements	44	14	73	21	01	81
Forced-choice Endorsements	22	-02	74	13	-06	84
Nonforced Desirability Judgments	10	09	78	10	-03	83
Forced-choice Desirability Judgments	-15	06	55	10	03	63
Self Rankings	64	03	33	38	19	46
Self Ratings	65	16	37	36	26	59
Peer Rankings	75	24	06	63	36	25
Peer Ratings	72	26	08	60	33	19
Measures of Self Esteem						
Nonforced Endorsements	23	88	-04	-00	69	05
Forced-choice Endorsements	05	84	-04	-02	71	11
Self Rankings	08	83	02	13	84	01
Self Ratings	22	86	-07	24	79	05
Peer Rankings	06	62	12	43	47	-19
Peer Ratings	20	63	22	52	56	00
PRF-Dy	-39	48	-02	-24	58	-02