1978

A Multidimensional Approach To Personality Inventory Responding

Edward Helmes

A MULTIDIMENSIONAL APPROACH TO PERSONALITY INVENTORY RESPONDING

by

Edward Helmes

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, Ontario
June, 1978

ABSTRACT

The limited number of computer simulations of personality have generally attempted to model a full range of aspects of personality. Determination of the adequacy of such models is difficult, given the disagreement regarding appropriate evaluative criteria among personality theorists. This work takes a more limited approach in dealing only with models of the item response process. This procedure has the advantage of dealing with a quite restricted range of possible behavior, thus making validation of the models easier. The method used here is to compare the response predicted by a computer model to the actual response of an individual.

The requirements of computer models are outlined. As all such models require measures of the items' positions on the dimensions of interest (scale values), a procedure for obtaining multidimensional scale values for dimensions of content (basic structure content scaling) is briefly reviewed. Evidence for the validity of these scale values is then presented before the prediction models themselves are presented.

Several computer models of the item response process are described and evaluated. Predictions were made using the basic structure scale values for two distinct samples of respondents to the Personality Research Form. The first consisted of 92 introductory psychology students and the second of 301 senior high school students. The simplest prediction models are invariant and make the same predictions for all individuals. Analysis of such models is useful in determining which aspects of items are important in the response process. Results for these models showed that social desirability was more effective in accurately predicting responses than were the aspects of content used by the invariant models tested. Jackson's (1968) threshold model of item responding, based on item social desirability, accurately predicted about 60% of responses. This level of prediction was superior to similar threshold models developed to predict responses on the basis of item content. An interesting outcome of these studies was the demonstration that the salience parameter of both Jackson's threshold model and the threshold models for content was very highly correlated with the scale score for the PRF constructs in question. This finding was used to develop cumulative prediction models based upon content using scale scores as the subject parameter. Such a model accurately predicted 64% of item responses in the two samples. A final study examined the predictive accuracy of Cliff's multidimensional spatial model. This model accurately predicted over 70% of the item responses.

The various prediction models were then evaluated in terms of their predictive accu-acy, how well they reproduced the distribution of item p-values, and in terms of the amount of information required to make a prediction. The distributions of predicted p-values for all models could be easily distinguished from the distribution based on real data. However, it was concluded that the best prediction models were predicting the maximum amount of predictable variance in the data. The multidimensional spatial model, although the best predictor, was rejected on the grounds of the excessive amount of information it required to make a prediction about an item, It was concluded that the best prediction model examined was the cumulative model using scale scores.

The three best prediction models were then examined more closely in order to determine which properties of items were the most closely associated with predictive accuracy. The properties most consistently found to be correlated with item predictive accuracy were item content scale value and the degree of scale internal consistency. This was true for both multidimensional spatial and for cumulative scale score models, but not for the threshold model based on desirability. These findings are consistent with expectations based on current practice in modern personality inventory construction.

ACKNOWLEDGEMENTS

I would like to thank those who have variously aided and abetted the completion of this thesis. Foremost among these is my advisor, Dr. Douglas N. Jackson. His encouragement and helpful suggestions and comments throughout my work in this area were invaluable. I thank Drs. R. C. Gardner and W. R. Krane for their detailed comments on an earlier draft of this thesis.

Special thanks are due Dr T. B. Rogers of the University of Calgary, for his generous donation of the data used in Chapter Three and Gail Clarke, for her donation of the data from which the high school sample was taken.

Many people contributed comments and suggestions over the last three years for which I am grateful. Chief among these were D. Chan, R. D. Chrisjohn, P. L. Reed and E. L. Strasburger. In addition, I am deeply grateful to Mary, for her patience, tolerance and continual reminders that it was too late to quit.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Description Page

CHAPTER ONE

INTRODUCTION

The purpose of this research is to examine the processes underlying responses to individual items in personality assessment. Different models of how individuals respond to items are developed in order to predict the actual behavior of responding to personality inventory items. The accuracy of the predicted responses is then used to evaluate the adequacy of the various models.

The personality item, with its accompanying restriction of responses to a binary choice, is perhaps one of the simplest and most basic units of human behavior which can be studied. The item in an inventory or rating scale is the basic unit of measurement for virtually every modern psycnological assessment technique. Therefore, investigations of item properties provide the means for improving future measurement instruments, as well as furthering our understanding of human behavior. The specificity of stimuli and responses in responding to inventory items is advantageous. The advantage lies in the identity of stimulus conditions for all respondents and the presumed generality across respondents of the cognitive processes involved in making the highly stereotyped responses used by most inventories. Thus, differences among individuals in the perception of the situation are reduced, as are the possible responses to be made to that situation. This greatly simplifies the study of item responding.

Various models of item responding have existed for several years (e g. Jackson, 1968; Damarin, 1970). If one views approaches to attitude item scaling such as Guttman's (1941) as being a model of responding, then this time period can be extended further back. The use of models of the response process is useful because such models permit a specification of the parameters in use and the manipulation of the degree of action of variables. Such studies have been performed in this area (e.g. Rogers, 1971) but the use here of Monte Carlo methods is not intended. Instead, the models are used to predict item responses. Those models that predict more accurately than others are assumed to be more valid. At the same time, it must be recognized that accurate prediction is not the sole criterion in the evaluation of the models. A good model should not only predict well, but it should also be of theoretical interest, non-trivial and employ relatively few parameters. The information required by a model in making its predictions is obviously one of the major determinants of the model's utility: the more information required, the less useful the model.

In order to predict item responses, it is necessary to have some numerical estimate of the items' locations on the latent dimensions upon which the predictions are based. In other words, scale values for items on some relevant trait or construct are required. If one is dealing with a large number of items on several constructs, as is the case here, obtaining the required scale values can be a problem. Therefore, a new scaling procedure based on components analysis, basic structure content scaling, was developed for such situations. This procedure is briefly outlined here.

Because of the novelty of the scaling procedure, and the fact that the computer models depend so heavily on the accuracy of the scale values they use, it was felt necessary to demonstrate that the scale values provided by the basic structure procedure accourately reflect the items´ positions on some of the constructs used by the models. This was done by determining the extent to which multidimensional basic structure scale values correspond to scale values derived from traditional unidimensional methods based on judgement data. It is only after a demonstration of the validity of the basic structure scale values that the computer models themselves are discussed.

The analysis of the item response process therefore proceeds through several distinct stages. The first stage discusses the requirements of models of the response process. A method of obtaining multidimensional scale values for item content from item responses, basic structure content scaling, is briefly outlined. This is followed by a brief demonstration of the validity of these scale values. The basic structure scale values for items are then used to investigate the item response process itself. This is done through various computer models of the response process The models are evaluated in terms of how accurately they predict the actual responses of individuals to the items comprising a large modern personality inventory and in terms of how well they reproduce the distribution of item frequency of endorsement (p-values).

In order to further our knowledge of the item response process and to aid in the development of better personality tests, it would be useful to determine what properties of items are associated with the items' being accurately predicted. Therefore, the analysis of the item response process concludes with a brief examination of the extent to which several easily obtained numeric properties of the items used are associated with high predictive accuracy. This is done for the three most accurate prediction models.

Thus the major focus of this thesis, the comparison of different computer models of the item response process, is supplemented by two other matters of interest. The first of these is a demonstration of the validity of the scale values used in the prediction studies. The second is an examination of the item properties associated with the most accurate prediction models.

CHAPTER TWO

MODELLING THE ITEM RESPONSE PROCESS

Responding true or false to a personality inventory or attitude item is a very restricted aspect of behavior. Marking alternatives on an answer sheet does not normally meet one's expectations as an example of the complex behavior associated with an individual's personality. However, this restriction of the possible responses is advantageous if one is interested in predicting behavior or in modelling it. Item responding, because of its specificity, is perhaps the most elementary unit of human behavior for which adequate prediction models can be constructed which meet criteria such as Loehlin's (1968) for the computer simulation of human personalities.

Before describing the computer prediction models to be evaluated here, some previous work relevant to the models will be discussed.

Previous Models

Earlier work on the process by which individuals respond to items, whether ability items or personality items, has been somewhat limited. Most of this work has dealt with ability items, although the field of attitude and personality tests has enjoyed a certain degree of attention. The conventional distinction between these types of items (Horst, 1968) is that ability items have a unique correct answer, whereas this is not true for personality, attitude, value or interest items. However, if one views items as scaling individuals for the degree of attribute present, this distinction need not hold (see

Appendix 1). The disproportionate amount of interest in ability items is undoubtedly due to such items being easier to deal with than personality items, as well as due to the incentive provided by the large number of users of academic and aptitude tests.

Regardless of the type of item considered, virtually all work dealing with the item response process makes the same three basic assumptions. The first of these is that items can be ordered with respect to the degree they reflect the attribute in question, whether arithmetic skills or attitude toward communism. The second assumption is that individuals can be ordered with respect to the degree they have the attribute. Thirdly, one assumes that this level of the attribute is important in determining how the individual will respond to the items. Other assumptions may be made in terms of the form of the function presumed to relate item responses to respondent properties or the form of the distribution of respondent properties, but these three assumptions are basic, and as such, rarely mentioned.

In the majority of psychometric work, emphasis has been placed upon the test rather than the testee. Somewhat surprisingly, for a long period, item properties other than simple frequency of endorsement were ignored as well (Goldberg, 1963). Despite this neglect, the item has assumed a certain degree of importance in more recent work.

In dealing with abilities, the concept of the item characteristic curve is fundamental to both classical test theory (Lord & Novick, 1968) and to latent structure analysis (Lazarsfeld, 1959). The item characteristic curve in aptitude testing relates the probability of a correct response for a given item to the continuum of individuals having different levels of ability. An item characteristic curve involves an implicit model for item responding, but this has rarely been made explicit by specifying the operations required on the part of the subject in making a response. Thus this body of work does not directly deal with the item response process, even though it involves all the basic assumptions required of such models. Nevertheless, there has been some work done using latent traits in modelling the item response process. For example, Damarin (1970) developed a multidimensional latent structure model for responses to the MMPI. Three latent traits were thought sufficient to account for the data: self-descriptive accuracy, favorable bias and a tendency to endorse neutral items.

Models of the response process almost require the use of the concept of a subject or person characteristic curve. Lumsden (1977) has described such a curve for ability tests, one very similar to that described by Jackson (1968) in the area of personality assessment. Weiss (cited in Lumsden, 1977) and Helmstader (1957) have also described similar concepts. In these characterizations, the subject characteristic curve is a function relating the probability of a keyed response to the continuum of items having differing degrees of intensity for the attribute in question. Such a curve is the direct converse of the item characteristic curve and as such is a more direct form of model of the item response process.

There have been more specific attempts at using such models of the response process in the field of personality and attitude measurement than in the area of ability assessment. The earliest such models are also those most clearly incorporating the three assumptions discussed earlier, namely those of Guttman (1941; 1950) and Loevinger (1948). These cumulative homogeneity models state that the individual's score on the unidimensional construct exactly determines how many items will be answered in a given direction. All items below that particular item will be answered in the same direction, whereas all items above the critical item will be answered in the opposite direction. A more modern development of this work is by Rasch (1960) in which the relationship between the probability of a response for an individual and the item position on the construct dimension is specified as an exponential function rather than a step function.

The conceptual simplicity of the cumulative homogeneity models of this type has proven very attractive to many workers in addition to Rasch. For example, Fiske (1966) has stated his preferences for the cumulative homogeneity model when used in an appropriate manner. This includes equalizing the proportion of total test variance among persons, items and residual and incorporating a uniform distribution of frequencies of item endorsement. Goldberg (1963), working from a different point of view, introduced the idea of an area of uncertainty around an individual's location on a content dimension. This idea has since been incorporated into Fiske's conception of the cumulative homogeneity model by Tyler (1968). Other students of Fiske have also continued working with this basic model. This work will be discussed in more detail later.

A model somewhat similar to Fiske's conception of the cumulative homogeneity model has been proposed by Jackson (1968) and explicated by Rogers (1971). The essential difference between these approaches lies in Jackson's model dealing with item social desirability rather than item content as does Fiske's. One important aspect of both these models that differs from the original cumulative homogeneity model (Loevinger, 1948) is that both use continuous functions rather than the original step function and operate in a probabilistic manner rather than in a determinate or error free one. An interesting aspect of Jackson's model is that its two parameters of salience and threshold were recently identified with the response styles of social desirability and acquiesence (Voyce & Jackson, 1977) which in turn correspond to two of the latent traits found by Damarin (1970) in his latent structure model of MMPI item responses.

The approach taken here differs from much of this previous work in that it utilizes computer models to predict item responses. This is quite different from, say, Damarin's (1970) latent structure analysis or Voyce and Jackson's (1977) investigation of the properties of the parameters of models. By evaluating several distinct models of the item response process in terms of how well responses are predicted, one can determine which model best accounts for the observed data.

This procedure is very similar in many ways to previous attempts at computer simulations of personality (Cranton, 1976; Loehlin, 1968; Tomkins & Messick, 1963), but is very much more restricted in scope than such full simulations as Colby´s simulations of neurotic (Colby, 1963) and psychotic (Colby, Weber & Hilf, 1971) personalities. Such models do, however, meet the criteria of Dutton and Briggs (1971) for simulations, despite their restrictions in scope. The models used here examine a defined behavior process, using a theory which describes and explains unambiguously. They also show how that process is affected by the environment (different items) and are formulated such that inferences can be verified by observation. In this case, the latter constitutes the comparison of predictions based on the models with the actual responses. This lack of verification of simulation models has often been used by their critics. For example, it was used by Blackmore (1972) to attack Moser and von Zeppelin´s (1969; 1970) simulation of the Freudian neurotic.

With this earlier work in mind, we shall now turn to a closer examination of the three assumptions necessary for models of item responding.

Requirements of Models

A computer model of any process requires that the process be specified clearly and exactly so that the necessary program can be developed and debugged. In addition, there are the specific structural requirements of the models themselves. It is these that will be discussed.

It is ordinarily assumed that individuals differ with regard to the traits or constructs in question and that these differences can be assessed in some way such that individuals can be ordered with regards to the trait. This assumption underlies all studies of individual differences. If this property were not present, there would be little point in proceeding further.

Similarly, items or situations can be ordered with regard to the construct in question. This is a basic problem in scaling (Torgerson, 1958) and can be done either by obtaining judgements of the degree to which the items reflect the construct or by other methods based upon actual responses to the items. In addition, one must decide whether the construct is unidimensional or multidimensional. This is necessary in order to obtain the correct ordering of the items with regard to the construct. It is only following this that the nature of the procedure used by the individuals in responding to the items can be dealt with. This may involve such processes as the individual´s perceiving the item (stimulus encoding), understanding its meaning (stimulus comprehension), deciding its relevance to himself and comparing the item´s intensity with the degree of self-perceived intensity and thereby determining which response (true or false) is to be made (Rogers, 1974a). To the extent that reality does not match the underlying assumptions, any model using those assumptions will not be valid, nor will it predict responses accurately. We shall therefore review selected literature relevant to some of these assumptions.

The first point of interest is the work by Loehlin (1961; 1967) which indicates that individuals differ in the degree of consistency of perceived meaning which they attribute to self-descriptive words. They attribute consistency of meaning to words to about the same degree that they reliably differentiate among other people. In other words, not all individuals view self-descriptive words as having identical and consistent meaning. This lack of consistency of meaning implies that the use of self-descriptive words alone in assessment (as in adjective checklists) is of dubious reliability or utility. Therefore, the inclusion of items similar to the adjective format (e.g. "I am curious.") in an inventory would prooably lead to the same degree of ambiguity. Such ambiguity means that it will not normally be possible to determine whether different responses to such items are due to differences in the degree of the relevant trait among the individuals or whether the individuals viewed the meaning of the items differently. Only the former reason can be dealt with by the sorts of models discussed here. The inclusion of such items would necessarily lead to lowered validity for the model. At least, this would be the conclusion drawn, even if the underlying cause was the inclusion of a poor item in the inventory. If not all individuals order items identically, any model assuming a universal and consistent ordering of items will predict responses to those items poorly.

It is relevant to note as well that Loehlin (1961) concluded that much of the consistency of subjects with regard to the self-descriptive terms appeared to be associated with the social desirability of the terms. Individuals consistently tended to attribute desirable terms to themselves and not to attribute undesirable terms. It might therefore be expected that social desirability would play a substantial role in determining responses. This was previously pointed out with regard to Damarin's (1970) latent structure model and the success of Jackson's threshold model (Voyce & Jackson, 1977).

A paper by Turner and Fiske (1968) supplements Loehlin's work. Turner and Fiske were searching for indications of the processes actually used by individuals in responding to personality items. They reported seven frequently used strategies, only two of which were thought to be relevant to the test and which would be considered acceptable (or even reasonable) to one interested in constructing models. The relevant processes included the general perception of the self or the use of relevant general experience. These accounted for 57% of the responses. The remaining 43% was accounted for by such processes as the use of a single or specific instance, qualification (placing restrictions on the applicability of the response), reinterpretation of the item, excessive emotional reaction or lack of any experience of the situation invoked by the item. Kuncel (1973; 1977; Kuncel & Fiske, 1974) has replicated these findings.

This is discouraging for the modeller, as it implies that only 60% or so of the responses to an inventory are predictable by models making conventional assumptions. However, there is some encouragement in that Turner and Fiske noted that item homogeneity (as defined by the item-total biserial correlation) tended to be associated with the relevant response categories. This implies that tests with high homogeneity (or more accurately, internal consistency) would elicit more relevant responses and would therefore be more predictable than those used by Turner and Fiske. The items used were taken from the MMPI and the Thurstone Temperament Schedule, neither of which represent the best of the currently available personality inventories.

Other work, however, is more encouraging. Kuncel (1973) also found that the use of inappropriate response processes was associated with increased latency of response and with the proximity of the item to the respondent´s location on the latent dimension. The location of both subject and item had previously been identified by Rasch scaling (Wright & Panchapakesan, 1969). This report supplements previous knowledge that items with moderate levels of endorsement are unstable in test-retest situations (Goldberg, 1963). Kuncel (1977) extended these findings by demonstrating that items with moderate levels of endorsement were particularly prone to both unstable responses over time and to inappropriate response processes. In large part, this is because it is within the range of moderate levels of endorsement that most small subject-item distances occur.

These studies indicate some advantages of operating at the level of the individual subject. From them, one infers that the most stable responses, based on appropriate response processes, arise from items which are located far from the individual subject's location on the latent dimension. In other words, the responses farthest from the individual's location are those that would best be dealt with by a model based on the appropriate response processes and the assumptions outlined earlier. At the same time, Loehlin's (1961) report of social desirability's being associated with those same conditions must be kept in mind. Therefore, the most stable responses should not only be associated with a large subject-item distance, but, at the same time, the role of social desirability, a large factor in producing stable responses, should be minimized in order to obtain uncontaminated measurements.

To satisfy these requirements in actual testing, one must operate at the level of the individual, instead of that of the group. At the present time, this means the use of tailored tests (Lord, 1970). Unfortunately, these are rarely possible in ability assessment and even less so in personality assessment. Tailored testing requires a large pool of calibrated items relevant to the concept to be assessed, as well as a means of rapid initial assessment of the subject's position with regard to the range of possible levels of the concept. Most commonly, this requires interactive computer-assisted testing. Therefore, obtaining large subject-item distances directly will only rarely be practicable. This is due in part to a lack of resources and in part to a lack of calibrated item pools for personality concepts.

However, large subject-item distances may arise as a consequence of what Jackson (1970; 1971) terms content saturation, or the intensity with which an item reflects solely the content of the trait or construct in question, in combination with other item selection strategies aimed at the reduction of irrelevant response styles such as desirability and the promotion of convergent and discriminant validity (Campbell & Fiske, 1959). It was these considerations which led to the use of Jackson's Personality Research Form - Form E (PRF-E) (Jackson, 1974) for the models developed here. Kuncel (1973; 1977) used items taken from an earlier form of the PRF which had not been subject to as extensive item analytic procedures as those used for the E form. At the moment, it remains unknown whether the new form would prove to be superior to the old with regard to the use of appropriate response processes.

Once the basic requirements of a simulation model have been developed, one can then incorporate some additional details suggested by experimental work. For example, Rogers (1974a) showed that stimulus encoding, comprehension, and deciding upon a response were all independent components of the response process. He also showed (1974b) that there is at least one further component, the self-referent decision, in selecting a final response. This component involves comparing item content to the concept of self located in memory (Rogers, 1977). This step requires further processing time over a response not referring to the self and therefore leads to increased item response latency. It probably also increases the unreliability of such items. Very few psychological processes are completely reliable, and in general, the addition of a further step in processing will also add a further error component as well. None of the models to be dealt with here reach this level of complexity.

Therefore, we can see that some factors involved in the item response process have been subject to research, but surprisingly little in view of the importance of this process for other work. With the importance of questionnaires and surveys in modern research, the lack of knowledge and empirical data relevant to the processes involved in responding to the items comprising those surveys is appalling.

With this general lack of knowledge in mind, the models to be used here depend only upon the most basic assumptions. One such assumption will be dealt with in more detail in the next section.

Scaling Items

A requirement of virtually every model of the response process is that items be scaled for the intensity of content for the relevant construct or constructs in question. The problem lies in selecting a method to do this.

If one is dealing with relatively few items, on the order of 20 or so, there are many possibilities. In general, the possible methods can be organized in terms of whether the scaling is to be unidimensional or multidimensional and whether the data are to be judgements or responses.

If one is dealing with a unidimensional construct, there are a wide variety of techniques available based upon judgements (Torgerson, 1958). Alternatively, one can deal with item responses and utilize Rasch scaling (Rasch, 1960; Wright & Mead, 1975). This procedure provides independent scale values for both respondents and for items and has been used by Kuncel (1973; 1977) in her research on item responding. Conventional cumulative homogeneous or Guttman scaling (Guttman, 1941; Green, 1956) is another possibility. With unidimensional procedures, there are few restrictions upon the number of items to be scaled. One exception is Rasch scaling, in which an exact solution is not practical with more than 20 or so items. An approximate procedure (Wright & Mead, 1975) is available which gives acceptable results.

However, in many cases it is difficult to justify the assumption of strict unidimensionality for personality items. (See Bejar (1977) as an example and Lumsden (1976) for a review.) This is particularly true if one considers the pervasiveness of evaluation or social desirability in personality assessment. Therefore, it is entirely possible that one would prefer the use of either multidimensional scale values or the scale values from one dimension of a multidimensional space over the use of unidimensional values in a prediction model of the response process. Again, if one is dealing with a small set of items, there are few problems. If using judgements, there is a wide choice of available multidimensional scaling procedures (Shepard, 1972). However, many of these procedures require all possible pairwise judgements of item similarity. With a large number of items to be judged, this can quickly develop into a task impossible for all but the most indefagitable judges. Procedures designed to cut down on the number of pairwise judgements needed are available (Spence & Domoney, 1974), but the amount of reduction still may not be sufficient to make the task feasible for most judges. It is, however, possible to perform multidimensional scaling with responses rather than judgements. Normally, many fewer responses are required than judgements of all possible pairs. One possible method is the Multidimensional Scalogram Analysis series of programs (Lingoes, 1972) for performing multidimensional Guttman scaling upon a set of item responses. Shepard and Carroll (1966) described a procedure they called parametric mapping for the analysis of response data. Unfortunately, this procedure has succumbed to computational difficulties (Carroll. personal communication). The traditional form of multidimensional analysis of response data has been factor analysis (including component analysis) (Shepard, 1972). However, the factor or component loadings of an item factor analysis are not readily interpretable as scale values in the same sense as are the factor or component scores. Therefore, a procedure based upon components analysis has been developed specifically to provide multidimensional scale values for large numbers of items such as are found in modern personality inventories (Jackson & Helmes, 1976). It does of course also have other applications. As this method is distinct from the purposes of this thesis but remains relevant, a detailed description is reproduced here as Appendix 1 and only a brief description of the basic structure content scaling procedure will be presented here.

Basic structure content scaling is a form of components analysis. As such, it deals with dimensions of content as simple linear composites of the original variables (items). The initial step in deriving basic structure content scores is to rescale the data matrix of subjects by responses by removing subject means and altering the subject variances to unity. This is equivalent to calculating a subject by subject correlation matrix and is required to ensure that the items are located appropriately in the multidimensional space. Subjects with similar levels of a particular trait or construct are assumed to respond in a similar manner to a set of items relevant to that construct. Thus similarities among subjects can be used to define a multidimensional space.

The basic structure scale values are defined as the projections of the items upon rotated axes within the multidimensional subject space. They are thus equivalent to component scores and are therefore also referred to as basic structure content scores. The number of axes reflects the number of dimensions of the space and therefore the number of traits or constructs in question. These axes may be selected as either orthogonal or oblique to one another through the use of an appropriate rotation, following the placing of axes in the subject space by conventional computational procedures. The use of orthogonal content scores has the advantage of maintaining the values for items on one dimension (trait or construct) independent of values on other dimensions. This is particularly useful if one wishes to minimize the role of social desirability, for example. If the dimensions are correlated, the oblique content scores may show higher empirical relationships with other variables than the orthogonal scores, as the oblique scores will tend to be higher on relevant dimensions than the orthogonal content scores.

The basic structure content scores or scale values used for the prediction models were calculated for a group of subjects separate from those used in the prediction studies. This sample consisted of 214 North American college students from a random sample of American colleges and universities. More details are presented in Appendix 1.

The basic structure procedure was used in part because of the logistical impossibility of obtaining the 61,776 pairwise judgements of the 352 items of PRF-E required for multidimensional scaling procedures using judgement data, and in part because it was important to identify scale values reflecting response processes. Although judgement methods would probably identify dimensions of relevant content, it is not certain that these would represent precisely the same dimensions used by respondents. In addition, it is possible that some determinants of responses, such as social desirability, are not represented to the same extent when individuals respond to items as when they are asked to judge their similarity to each other or to some concept. The next chapter presents some data in order to evaluate the validity of the scale values provided by the basic structure scaling method. This is necessary because this work represents the first instance of the actual use of basic structure scale values in a substantive problem.

CHAPTER THREE

VALIDATION OF THE BASIC STRUCTURE CONTENT SCORES

Any model of the response process requires the use of at least one parameter for each item. These parameters (e. g. a scale value for content or for desirability) are normally simply assumed to be valid. However, the validity of this assumption should be assessed before the model itself is investigated. That is to say, the parameters used (for example, scale values for items) should have a strong and demonstrable relationship to the 'true' value. In classical test theory, such a relationship is well defined in the theoretical relationship of true score to observed score (Lord & Novick, 1968). However, such theoretical concepts and the necessary mathematical assumptions have not been defined for personality items, particularly in the context required here. Therefore, an empirical approach will be taken to evaluate the validity of the content scale values used in the prediction models.

In general, this approach is to compare the scale values derived by the basic structure method with item scale values derived from traditional unidimensional scaling methods. It is important to note two points concerning this comparison. The first is that the scaling methods differ in the form of data used. Rating methods are based on judgements of content; the basic structure method uses responses to items. There need not be a correspondence between how items are perceived and processed if one is judging them and how they are perceived and processed if one is responding to them. However, if such a correspondence does not exist, it would be most surprising and would indicate a lack of generality of the concept being dealt with (Stewart, 1974). Empirically, Boyd and Jackson (1966) found a similar structure for attitude statements via both factor analysis and multidimensional scaling, except that the structure obtained from responses (factor analysis) was not as clear as that obtained from multidimensional scaling. Furthermore, responses yielded a large acquiescence factor. The similarity found in this case does give grounds for expecting a fair degree of correspondence between the different methods of scaling, assuming that the content scaling procedure does indeed produce valid scale values for items.

The second point is related to the dimensionality of the scale constructs. The basic structure scaling method is multidimensional; the judgement scaling method used here is unidimensional. A lack of correspondence may arise from differences in dimensionality. Therefore, a comparison of scale values calculated by the two different procedures is advantageous from one point of view in that it allows a determination of the adequacy of the response method's implicit assumption of the unidimensionality of the inventory's scales.

However, the requirement of unidimensionality, or an approximation to it, is normally relevant only to those items which are keyed on a scale. The most prominent aspect of content likely to violate the assumption of unidimensionality is social desirability. The importance of such an evaluative component in personality concepts has been repeatedly demonstrated. It is a confounding variable whose presence must be minimized in any attempt to assess a 'pure' concept. Otherwise, one is forced to deal with concepts on a two-dimensional level, with one level always being social desirability. If the evaluative process differs from judging to responding to items, then there will be a lack of correspondence in the unidimensional scale values. In addition, the relationship of items not keyed on a scale to the keyed items also has a bearing upon the congruence of the results of the two scaling methods. Items on scales not keyed to a particular scale construct may be keyed on scales which are conceptually related to each other, for example, Order and Cognitive Structure, or Defendence and Aggression. Items on one scale may be in some way relevant to another scale via some form of inferential network (Bruner & Taguiri, 1954; Cronbach, 1955; Hays, 1958). The problem lies in whether the relationships perceived by individuals via an inferential network are identical to those actually used by individuals when they respond to items for those scale constructs. To the extent that these networks are not the same, similar scale values will not be obtained. This will be particularly true for items not keyed on a scale, but which do have a conceptual tie to that scale. Judgements of, and responses to, these items will be those most prone to be discrepant.

The presence of relations among scales will also tend to produce lower correlations across methods for basic structure content scores based upon an orthogonal rotation than those scores based upon an oblique rotation. Since some scales are related to one another, an oblique rotation of the basic structure content scores would provide a better fit to the configuration of item points and lead to higher correlations with the judged values. To a certain extent, the use of dimensions based upon scale factors rather than scales themselves would overcome this problem, as the relationships of scales among themselves are contained in the factors.

With these qualifications in mind, the actual comparisons of content scores with judged values are done in three phases. The first of these obtains judgements on items from ten scales of PRF-E. The second uses data obtained by Rogers (1973) and provides information on the generalizability of the basic structure content scores. The third phase examines the relationship of content scores for the PRF-E control scales of Desirability and Infrequency to judgements of those properties for the entire set of PRF-E items.

Comparison of Content Scores with Unidimensional Judged Scale Values

Method

Subjects

A total of 93 introductory psychology students, 48 males and 45 females, completed the entire set of materials to be judged. Approximately half the judges took part in either of two one hour sessions. Judges received credit for research participation required by their course in return for their taking part in the judgement task.

Materials

Each subject received a booklet containing an instruction sheet, target descriptions, item lists and response sheets. The instructions were for the subject to read the target descriptions and then judge the 20 items on the following page on a nine point scale as to how characteristic or uncharacteristic each statement was of the target. The judge was to do this for each of the ten targets.

The targets were brief descriptions of individuals written in such a way as to describe the individual target as an extreme exemplar of a single scale of the PRF. Thus the description for Carl Bates emphasized only Aggression, that for Alfred Anderson emphasized Understanding and no other trait. In writing the target descriptions, use was made both of the description of the scale as given in the PRF manual (Jackson, 1974) and of certain items taken from that scale. However, use of the exact wording of scale items was avoided as much as possible. This procedure was adopted as it was thought to be easier for the judges to conceptualize a person rather than an abstract trait as a target for similarity judgements. A previous study (Rogers, 1973) has shown that judges can reliably perform the latter type of judgement.

The scales for which target descriptions were written were selected at random and included Abasement, Aggression, Change, Dominance, Endurance, Order, Sentience, Social Recognition, Thrill-seeking (the negative pole of Harmavoidance), and Understanding. Each target description was followed by a sheet with 20 items to be judged. Sixteen of these items were from the scale relevant to the target and the other four were taken from the ten scales for which a target was not composed. Each response sheet listed the item number and the possible responses from one to nine with nine always indicating the item was extremely characteristic of the target. Sample instructions, target descriptions, item listings and response sheet are also contained in Appendix 2.

Procedure

Subjects completed the set of targets in one of five orders. These orders were randomly determined with the restriction that the first target was always one of the five targets thought to be easiest for the judge to conceptualize, i.e. one of the targets for Aggression, Dominance, Endurance, Order, or Understanding. This procedure was intended to provide an easy introduction to the task. Item lists for targets were constant across all judges with item ordering within lists being random. Thirty-one subjects (15 males, 16 females) who completed the task in less than 45 minutes were requested to complete a supplementary task consisting in part of an additional target for the negative pole of the Abasement scale. Samples of these materials are contained in Appendix 2. Upon completion of the task or the supplementary task, subjects received an information sheet thanking them for their cooperation and outlining the purpose and significance of the study.

Following collection of all the data, the information was keypunched and verified. Response sheets containing items for which either no response or multiple responses were made were omitted from further analysis. This resulted in from 89 to 92 judges completing any given target.

Mean ratings and successive intervals scale values (Diederich, Messick & Tucker, 1957) were computed for all 20 items for each of the eleven targets (including the target of the supplementary task). The successive intervals scaling procedure provides a set of category boundaries adjusted so as to normalize simultaneously each distribution of judgements on the same baseline. Thus ratings with unequal variances in the raw data can be placed on a scale with equal intervals in terms of the Thurstone model. Such a scale is normally centered at zero with unit intervals. Normally, the successive intervals values correlate very highly with the raw mean ratings (Scott, 1968).

Results and Discussion

The internal consistencies of the judgements on each target, as assessed by two different methods, are given in Table 3-1. Coefficient alpha provides a lower bound estimate of the reliability of a unidimensional scale (Lord & Novick, 1968). Coefficient theta provides an estimate of the reliability of a linear composite (Bentler, 1972). As such, it is a higher lower bound estimate of internal consistency than is coefficient alpha. As it does not assume the scale is unidimensional, theta is probably the more appropriate measure in this case, in which the attributes being assessed frequently assess multiple facets of a fairly broad trait. Nevertheless, the consistency of the judgements under the assumption of unidimensionality (coefficient alpha) is striking.

It should be noted that all the judgements are quite reliable, and that the judgements for the reversed Abasement target, which involved one third the number of judges as the other targets, are as reliable as those with more judges. In part this may be due to the nature of the reversal of the Abasement target, which produced a target which was somewhat easier to judge. It is difficult to conceptualize the positive pole of Abasement and still maintain a moderate level of social desirability.

The degree of correspondence between the judged scale values and the orthogonal and oblique content scores are given in Table 3-2. These content scores are based upon the responses of 214 U. S. college students, as described in Appendix 1. Correlations are given separately

Table 3-1. Internal Consistency of Content Judgements

Target	Scale	Coefficient Alpha	Coefficient Theta
JT*	Abasement	85	98
GF	Abasement	73	92
CB	Aggression	75	91
TH	Change	74	90
FH	Dominance	81	94
HM	Endurance	77	93
TB	Thrill-seeking	77	93
MW	Order	78	92
GB	Sentience	76	92
RJ	Social Recognition	76	92
AA	Understanding	70	92
	Mean	765	926

* Target based upon negative pole of Abasement.

Note: Decimals omitted. Means calculated upon unrounded figures.

Table 3-2. Correspondence between Basic Structure Content Scores and Successive Intervals Scale Values

	Orthogonal Content Scores				Oblique Content Scores			
Scale	All 20 Items		16 Scale Items		All 20 Items		16 Scale Items	
	Raw	Corrected	Raw	Corrected	Raw	Corrected	Raw	Corrected
Abasement (Positive)	82	89	83	90	88	95	88	95
Abasement (Negative)	-85	-87	-86	-88	-91	-93	-92	-94
Aggression	84	92	93	99	94	99	96	99
Change	80	89	85	94	84	93	89	98
Dominance	94	99	97	99	93	99	97	99
Endurance	82	89	86	93	92	99	94	99
Thrill-seeking	-95	-99	-96	-99	-96	-99	-97	-99
Order	87	94	89	97	89	96	90	98
Sentience	63	68	68	74	82	90	90	98
Social Recognition	95	99	96	99	95	99	96	99
Understanding	82	89	84	90	86	93	87	94
Mean	845	905	875	933	900	962	924	979

Note: Decimals omitted. Means calculated upon unrounded figures.

for the 16 items keyed on the target scales. These were expected to be higher, on the grounds of a lack of involvement of an inferential network, as mentioned previously. In addition, it may be that judges find the keyed items more salient and therefore easier to judge accurately. Table 3-2 also gives the correlation between content scores and judged scale values as corrected for the unreliability of the judgements by applying the standard correction for attenuation. In several cases, this results in correlations in excess of 1.0, which have been reduced to 0.99.

These data are encouraging for several reasons. First, the fact that judges can provide a consistent set of scale values for all PRF-E scales used strongly supports the idea that personality scale items can be scaled (Rogers, 1973). This work extends Rogers´ findings in showing that this can be done by using a target person rather than a trait description. In addition, these scale values do in fact provide an accurate ordering of the items. This is indicated in part by the agreement of the judgements with the content scores, and also in part by the reversal in sign of the correlation of the judgements for the reversed Abasement target when compared to the original Abasement target. Second, the substantial correlations in Table 3-2 indicate that the scaling method based upon responses has a high degree of validity. The scale values obtained from responses are very similar to those obtained from judgements. Third, an assumption of unidimensionality for each scale construct appears justified. This is supported by the correspondence between unidimensional scale values from the successive intervals procedure and the multidimensional scale values for a single scale from the basic structure method. As the latter bases each set of scale values for a construct upon a single rotated component, each set of content scale values from the basic structure method must be unidimensional. The correspondence between the two sets of values is evidence not only for the validity of response scaling and the basic structure method, but also provides support for the methods of construction of the PRF-E which emphasized the relationship of the item to its own and other scales (Jackson, 1974). However, it is apparent that the scale constructs are not strictly unidimensional, as indicated by the substantial increment in internal consistency of coefficient theta over coefficient alpha (Table 3-1). This is partially due to theta being a better lower bound estimate of reliability than alpha (Bentler, 1972), and partly due to the nature of the scales as a composite of correlated items which define more than one aspect of the scale trait.

Thus we can tentatively conclude that the content scores do in fact reflect the intensity of item content. This was demonstrated using content scores based upon a North American college sample (Helmes & Jackson, 1977) and judges who were predominantly freshman students at this University (Table 3-2). This implies a certain degree of generality of the content scores across subject groups. The generality of the content scores can further be evaluated in this context by using item scale values determined elsewhere, using judges from a different university.

Comparisons of Content Scores with Values from the Rogers Study

To determine the degree to which the content scores from the basic structure scaling method are valid for other groups, data were obtained from a previous study by Rogers (1973), in which judgements were made as to the degree to which selected items from Form A of the PRF reflected the traits Autonomy and Impulsivity. Such information is useful if prediction models using the basic structure content scores will be used for different samples of respondents than those upon which they were calculated.

Comparisons were made between the scale values obtained in the Rogers study and those determined from the basic structure method. The data kindly provided by Dr. Rogers consist of the mean ratings of the Impulsivity and Autonomy content of 40 PRF-A items by 108 University of Calgary undergraduates. Separate groups of 54 students judged the items on either Autonomy or Impulsivity, using descriptions of these traits as provided in the PRF manual. Judgements were made on a seven-point scale.

Of the 40 PRF-A items used by Rogers, 15 remain unaltered on PRF-E. Of these, nine are from the Autonomy scale and six are from the Impulsivity scale. Correlations between content scores and the judged values obtained by Rogers are given in Table 3-3. Content scores based upon both orthogonal and oblique rotations are reported. The correlations for all items for both scales are substantial for both rotations, and are even higher if only the items keyed on a scale are

Table 3-3. Correlations of Judged Scale Values and Content Scores

Judged Scale Values	Content Scores					
	All Items		Autonomy Items Only		Impulsivity Items Only	
Scale	Orthogonal	Oblique	Orthogonal	Oblique	Orthogonal	Oblique
Autonomy	.85	.88	.96	.98	-.14	.35
Impulsivity	.65	.73	.15	.54	.94	.95

Note: Judged scale values are based upon Rogers (1973). There are nine Autonomy items and six Impulsivity items.

considered. As expected, the correlations for oblique content scores are somewhat higher in all cases. The sharp rise in the correlations of scores for Autonomy items on Impulsivity and vice versa with oblique content scores is somewhat surprising, given the fact that these two scales generally load different factors and are only moderately correlated (about .20) (Jackson, 1974; Skinner, Jackson & Rampton, 1976). However, the original judgements of Autonomy and Impulsivity in Rogers' study correlated 0.63 with each other, and this correlation from the judgement data may be the cause of that reported here. Such a correlation is also indicative of the inadequacy of judgement techniques in such circumstances.

Nevertheless, these data show a good correspondence between mean ratings of content and content scores for a group of judges separated in time and space from the subjects used in deriving the content scores. This is encouraging for the application of the content scores to different groups. It also parallels claims of independence of scores for subjects and items made for unidimensional Rasch scaling (Rasch, 1960; Wright & Mead, 1975).

Comparison of Desirability and Infrequency Content Scores to Judged Desirability and Frequency of Endorsement

An analysis of the two control scales of PRF-E is of interest from several points of view. First, there is no homogeneous content dimension associated with the items on these scales (Jackson, 1974) (unless one wishes to interpret desirability in this way). Therefore, any scalability for these scales will be due more to the conceptual strength of the rationale underlying the inclusion of these scales and to the power of the scaling method than to the strict scalability of the items. This is not to deny that desirability and frequency of endorsement are meaningful or useful in their own right. It is simply that desirability and frequency of endorsement are attributes of items logically distinct from item content itself. Second, the prevalence of response styles such as social desirability (Berg, 1967; Edwards, 1970) indicates that desirability scale values in particular provide useful information about items. If there is a strong correspondence betweeen content scores for Desirability and those found by direct judgements of item desirability, then there is a potential saving of experimental effort in obtaining desirability scale values through the use of the basic structure scaling method. Third, several of the computer models to be discussed later use desirability as a basis for prediction. It would therefore be useful to know the degree of correspondence between scale values based on judgements and the values based upon responses from the basic structure scaling procedure.

One might ask what results could be expected in a comparison of derived content scores and actual unidimensional judgements for the control scales. As in the case of the content scales, the major consideration is that the most salient items are those keyed on the two scales in question. Most items on PRF-E are fairly neutral with regard to desirability and frequency of endorsement. This was a requirement for inclusion of items in the inventory (Jackson, 1974). The only exceptions to this are the items on the two control scales. It is thus to be expected that a correlation between scores will be higher for those items keyed on a scale than for all items. This is because the relatively neutral non-keyed items are subject to much more error of estimation by both procedures. This increase in error will act to lower the correlation over all items, as was the case with the content scales. In addition, considerations of content are more important when responding to items than when one is asked to judge the desirability of items. Thus, in this case the responses are multidimensional, while the judgements are more nearly unidimensional. This will also act to lower the degree of correspondence between the two sets of scale values.

The subjects used and method of obtaining judgements of desirability and frequency of endorsement have been described in detail elsewhere (Helmes, Reed & Jackson, 1977). Judges consisted of 237 introductory psychology students. Instructions were standard for this type of task.

The values obtained in the Helmes et al. (1977) study were correlated with both orthogonal and oblique content scores, both over the entire set of 352 items and over the 16 items keyed on a given control scale. This was repeated for both the original mean ratings and for the successive intervals scale values (Diederich et al., 1957) reported by Helmes et al. (1977).

The results are summarized in Table 3-4 and generally confirm expectations. Correlations of judged values are uniformly higher with oblique basic structure content scores than with orthogonal content scores. In part, this may be a reflection of the correlation perceived by judges between desirability and frequency of endorsement (Jackson & Messick, 1969). This correlation was previously determined to be 0.45 on the PRF (Helmes, et al., 1977) and, on the MMPI, as 0.87 by Jackson and Messick (1969). In addition, correlations over the entire set of items are lower than those over only scale items. Items keyed on the scales are more extreme with regard to frequency of endorsement and desirability and thus would be both easier to judge and more liable to yield a higher correlation than more neutral items.

The magnitude of the correlations themselves generally indicates good agreement between the two sets of scale values. Correlations for frequency judgements have been reflected, as the basic structure content scores are based upon infrequency of occurrence. The lowest correlations are for orthogonal content scores with judged desirability when all 352 items are considered. These are of a reasonable size, although low in comparison to those for frequency of endorsement and the

Table 3-4. Correlations between Content Scores and Judged Scale Values for PRF-E Validity Scales

		Desirability Judgements	
Content Scores		Mean Rating	Successive Intervals Values
352 Items	Orthogonal	.38	.36
	Oblique	.60	.58
16 Items	Orthogonal	.79	.78
	Oblique	.89	.87

		Frequency Judgements	
Content Scores		Mean Rating	Successive Intervals Values
352 Items	Orthogonal	.71	.67
	Oblique	.73	.69
16 Items	Orthogonal	.99	.83
	Oblique	.99	.95

content scales. There are several possible reasons for these correlations being lower than the others. This situation may be a case in which the cognitive processes used in judging items and responding to them are different. This in turn may involve the use of the Differential Reliability Index (DRI, Jackson, 1970) in selecting PRF-E items. The use of this index acts to minimize the role of desirability responding by eliminating from the test any item which correlates more highly with desirability than with the item´s own scale. Therefore, it is possible that desirability is in fact lower in PRF-E items from a response viewpoint, but not from a judge´s viewpoint when he is asked to consider only item social desirability.

In general, the results of this comparison are quite encouraging, particularly for items keyed on a scale. The ranking of items keyed on scales as determined by judges is nearly identical to that obtained by the content scaling method. However, the magnitude of the correlation between desirability content scores and judged desirability scale values over all items implies that somewhat different processes may underlie desirability response scale values and judged values.

Summary

The findings reported in this section provide strong encouragement for the use of the PRF-E basic structure content scores in models of the item response process. The idea that personality items can be scaled in terms of their content (Rogers, 1973) was supported. Judges in a conventional unidimensional scaling task were highly consistent in their judgements of item content and of item properties such as desirability and frequency of endorsement. The judgements agreed highly with the item content scores for ten PRF-E content scales and one control scale. This strongly indicates that the content scores have a high degree of validity in that the content scale values accurately mirror the degree to which the intensity of trait content is represented in the items. The exception to this may be the scale values for desirability. The content scores also correlated highly with scale values obtained in a previous study (Rogers, 1973), which indicates a certain degree of generalizability for the content scores.

Taken together, these results are sufficiently clear to allow the use of the content scores in different models of the item response process.

CHAPTER FOUR

MODELS OF THE RESPONSE PROCESS

In this section, several models of the item response process will be described and then evaluated in terms of how accurately each model can predict the actual responses to the same set of items. This approach provides direct data on the utility of the models in predicting behavior and also supplies an interesting contrast to previous studies of the response process which have used different methodologies. For example, a computer simulation of responses to the MMPI was performed by Rogers (1971) but was analyzed in terms of the resulting factor structure rather than by comparison of the predicted responses to actual responses. Other studies of personality item responding (e.g. Tyler, 1968; Voyce, 1973; Voyce & Jackson, 1977) have investigated the properties of the model parameters used. A very few studies (Cliff, Bradley & Girard, 1973; Cliff, 1977) have compared actual responses with those predicted by the models used.

The studies reported here include some of the models used in these earlier studies. Therefore, there will be opportunities to compare directly the relative effectiveness of the models proposed here and those used previously by other workers in modelling the item response process.

Subjects

Two separate groups of respondents to PRF-E (Jackson, 1974) were used in the prediction studies. The first of these consisted of 92 introductory psychology students (43 males and 49 females) at the University of Western Ontario (UWO). These students received research participation credit and a copy of their PRF-E profile in return for their cooperation. The second sample consisted of 301 Grade 12 and 13 students (162 females and 139 males) selected from a larger sample from a large suburban Ontario high school. These latter subjects all had person reliability coefficients (Jackson, 1976) in excess of 0.5. This coefficient consists of the correlation between two sets of scale scores, each scale consisting of one-half of the full length scale. This criterion was applied to eliminate as many unreliable and erratic respondents as possible from a sample in which the respondents were not supervised in completing the inventory.

In both samples, no respondent had Infrequency scale scores in excess of 3 or more than 3 omitted items. Both these conditions indicate careless or random responding on the PRF (Jackson, 1974). Any omitted item was scored as a false response.

Invariant Models

Perhaps the most elementary approach to the prediction of item responses is to predict the same response for all individuals on the basis of some property of the items being predicted. Models of this type ignore individual differences and are therefore called invariant models.

Because invariant models make the same predictions for all respondents, they cannot predict all respondents equally well, except in the trivial case in which all respondents do in fact respond in the same way. Such models are thus of less interest than models which at least have the potential to predict equally well across all respondents. Invariant models are also of rather low theoretical interest. Behavior is known to be complex and very simple models can be expected to add little to our knowledge of individual differences and their interaction with the stimulus elements of an item in producing a response characteristic of the respondent.

However, if one selects a collection of invariant models which are based upon different factors which in turn are presumed to contribute to the selection of an item response, then one can at least assess the relative importance of these contributory factors. For example, if an invariant model based upon desirability predicted responses substantially more accurately than a model based upon item content, one might then conclude that item desirability was a more important determinant of the response than was the scale content of the item.

Five different invariant models were used, each being based upon a different aspect of personality items. These models are described in detail below.

Random

This model is intended primarily to confirm one's expectations as to the level of accuracy to be expected from truly random predictions, that is, 50% accuracy with binary responses. Accordingly, a random number generating algorithm (for the CDC Cyber 73) was used to produce 352 random binary digits which were then used to predict the item responses for both samples of respondents. The same random vector was used to predict the responses of all individuals.

Scoring Key

The direction of keying for an item is one of the more salient aspects of an item and represents one fairly obvious attribute of item content. An individual who is sensitive only to this aspect of content would receive either the maximum or minimum scale score, depending upon whether the individual responds in either the keyed or non-keyed direction. Only the first case is of interest here. Therefore, a true-keyed item was predicted as being a true response and false-keyed item as being a false response.

Disjunctive Content

Another aspect of items related to their content is the scale value for the scale construct on which they are keyed. For example, an item keyed on an Aggression scale should be assignable an unique number which reflects the degree to which the item content expresses Aggression. Normally, the highest scale value for any given item will be on the scale on which the item is keyed. Therefore, one can use the direction of the scale value for the scale on which the item is keyed to predict the direction of the response to that item. Therefore, an item with a positive scale value on the scale on which it is keyed was predicted as being a true response, and vice versa for an item with a negative scale value. A content key was made up for the 352 items of PRF-E in this way, with one additional modification. If the scale value for any particular item for social desirability was higher in the opposite direction than the scale value for the scale on which the item is keyed, then the direction for desirability was used in the predictions. Both content and desirability scale values used in this case were the basic structure content scores. The term disjunctive is use to describe this model, following Coombs´ (1964) terminology for models of this general type.

Social Desirability

The importance of social desirability in personality inventory responding has been demonstrated repeatedly (Berg, 1967; Edwards, 1970; Rogers, 1971). Therefore, the item scale values for desirability reported by Helmes et al. (1977) and used in the previous chapter were transformed to binary form suitable for use as predictors. An item with a positive successive intervals scale value was predicted as a true response. An item with a negative scale value was predicted as a false response.

P-values

One of the most basic pieces of information concerning an item is its level of endorsement or p-value. As the average response to that item for some reference group, the p-value must by definition be a good predictor of responses. As such, the p-value provides an indication of the upper bound to accurate prediction as the maximum-likelihood estimator of the true response. For the prediction model, the binary prediction key was based upon the p-values for the U. S. college sample reported by Helmes et al. (1977). This sample was also used to derive the basic structure content scores. Items with p-values of 0.5 or higher were predicted as true responses. Other items were predicted as false responses.

Computational Procedures

For each of the above models, a binary true-false key was composed as described. This key was then used as the basic set of predictions for all respondents in both samples.

To evaluate the above models, as well as those yet to be described, the following basic matrix of outcomes was used.

		Predicted Response	
		True	False
Actual	True	TT	TF
Response	False	FT	FF

This is basically a 2 x 2 contingency table. If a model predicts all responses correctly, then the two elements of the major diagonal (TT and FF, indicating agreement) will be positive and the two off-diagonal elements (TF and FT) will be zero. A model less accurate than this will have a greater proportion of entries in the diagonal than in the off-diagonal elements. A model making random predictions will have entries determined entirely by the marginals. A conventional chi-square test of independence will detect such a case of accurate prediction, but is equally sensitive to the obverse case in which the off-diagonal elements are greater than the diagonal elements. In addition, variables may show perfect association (the opposite of independence) in that one variable may be entirely predictable from knowledge of the other and yet show no agreement (Bishop, Fienberg & Holland, 1975, p.394). Thus, a test of independence of the rows and columns of the table is not appropriate in this case. The best measure available for this case is Cohen's (1960) kappa, originally designed as a measure of inter-rater agreement. As such, it determines the degree to which a given frequency of entries in the major diagonal exceeds the frequency predicted by the marginal totals. That is,

$$\text{Kappa} = \frac{p_o - p_c}{1 - p_c}$$

where p_o is the observed proportion of agreements and p_c is the proportion of agreements predicted by chance. With all entries in the major diagonal, kappa will equal +1. In the converse case, kappa will be -1. If there is no agreement above the chance level, kappa will equal zero. With a large sample, kappa is approximately normally distributed with mean 0 and variance $p_o(1-p_o)/ N(1-p_c)^2$, where N is the number of observations. Thus tests of the null hypothesis of chance agreement are easily made, as are pairwise comparisons of two sets of data. In all cases in which pairwise comparisons are made in the following text, Cohen's (1960) test for two values of kappa has been used, unless specifically stated otherwise. The single case in which kappa is less than successful is the one in which all entries in the outcome table are in a single cell. This happens in the case of items of the Infrequency scale, to which respondents do make identical responses and which models do predict with perfect accuracy.

Results and Discussion

The outcome of the predictions of the five invariant models is given in Table 4-1 for both samples and collapsed across samples. A point to be kept in mind in discussing all the prediction results concerns the statistical significance of the values of kappa. With a very large number of observations (over 32,000 for the UWO sample and over 138,000 for the high school sample), it is very easy to have relatively small values of kappa become significantly different from

Table 4-1. Results for Invariant Prediction Models

	Sample								
	UWO			High School			Pooled		
Model	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa
Random	15,884	49.05	-.019	52,388	49.45	-.011	68,272	49.35	-.013
Scoring Key	17,139	52.92	.058	53,493	50.49	.010**	70,632	51.06	.021
Disjunctive Content	16,491	50.92	.019	52,623	49.67	-.006*	69,114	49.96	-.001 †
Desirability	20,465	63.19	.262	62,922	59.39	.186	83,387	60.28	.204
P-value	21,256	65.63	.313	66,194	62.48	.250	87,450	63.22	.265

† NS
* p<.05
** p<.01
All others, p<.001

zero. Only one value of kappa in Table 4-1 is not different from zero at the .05 level (one-tailed), and that value is zero. All others are significant, some with values of the standard normal deviate in excess of 70, with accompanying remote probabilities.

The random key appears to be one of the keys possible by random chance that does significantly worse than chance. It should be noted however, that the accuracy level of 49.35% is not appreciably different from the expectation of 50%. In addition, this value provides an indication of the magnitude of random error of prediction.

Both aspects of item content as typified by the true-false scoring key and by the disjunctive scale content key fail to predict item responses very well, although the scoring key at least did somewhat better than chance.

As expected, the desirability key predicted over 60% of the responses over both samples, as did the p-value key. As the arithmetic mean, the p-values should be among the most accurate of possible predictors. This is true even with the consideration that the p-values used in composing the prediction key were from a population (the U. S. college group) distinctly different from those for which predictions were actually made.

In order to determine which models were the better predictors, a randomized-block analysis of variance was calculated with subjects as the blocking factor and value of kappa for each subject on each model as the dependent variable. This was done separately for each sample. In both cases, both subjects and models showed significant main effects at the .001 level. Neuman-Keuls tests showed that all models differed from each other at the .01 level for the UWO group, but that the random and disjunctive models did not differ at the .01 level in the high school group. Complete analysis of variance results are contained in Appendix 3.

These results would indicate the importance of social desirability in item responding. An indication of the pervasiveness of desirability can be obtained by examining Table 4-2, which reports intercorrelations among some of the keys used by the invariant prediction models and some item parameters for the 352 PRF-E items taken from Helmes et al. (1977). Note that only the true-false scoring key has even a degree of independence from the other variables. The high degree to which desirability contributes to p-values can be seen in the second column of Table 4-2. Scale values based on judgements (Helmes et al., 1977) account for 55% and 38% of the variance in p-values in the two samples reported.

The importance of social desirability in item responding would argue that a model based upon social desirability incorporating individual differences would prove to be a good predictor of responses. Such a model is the next to be described.

Table 4-2. Correlations among Bases of Invariant Models

	College p-values	HS p-values	Judged Dsy	Judged Freq.	P-value Key	Desirability Key	True-false Scoring Key
College p-values	-	83	74	63	80	59	07
HS p-values		-	62	74	61	47	03
Judged Dsy			-	45	61	77	18
Judged Freq.				-	37	25	-06
P-value Key					-	58	12
Dsy Key						-	21

Note: Decimals omitted.

Jackson's Threshold Model

Several years ago, Jackson (1968) formulated a model of the item response process which was within the general tradition of cumulative models but which differed in certain vital respects. Primary among these was that item content in relation to a scale construct was NOT the main determinant of item responses. Instead, he suggested that individuals differed in the degree of the salience or importance of social desirability for them and in their threshold or willingness to respond to items in terms of social desirability, and placed these concepts into a testable mathematical framework. (See Figure 4-1 and below for a definition of these parameters.) This model was described in detail by Rogers (1971) and evaluated through an examination of the factor structure of the MMPI using artificial data generated using Jackson's threshold model. It was concluded that differences in threshold and salience of social desirability as generated by the threshold model gave a very close approximation to the generally obtained factor structure of the MMPI. Later work by Voyce (1973; Voyce & Jackson, 1977) using the Differential Personality Inventory (DPI) confirmed these results, with the exception that the two components of the threshold model accounted for less variance than in the case of the MMPI.

The PRF was constructed in the same tradition as the DPI to assess Murray's (1938) needs. One would therefore expect a similar role of social desirability as found by Voyce and Jackson (1977). However, Voyce and Jackson analyzed the model in terms of the loadings of the model parameters on the first two DPI factors. In order to evaluate the threshold model in the present context, it was necessary to translate the model into a form suitable for prediction of item responses.

Parameters of the Model

The first step was to calculate saliences and thresholds for the respondents in the two samples. There are several possible methods of obtaining these values. Voyce (1973) found that a method based upon a linear regression procedure gave more stable and independent values than the method originally used by Rogers (1971). The original method involved computing the biserial correlation between the binary item responses and the item social desirability scale values as a measure of salience and computing ascending and descending estimates of the threshold by forming successive 20 item scales with items of increasing or decreasing desirability content until a mean scale score of 0.5 was reached. With Voyce's results in mind, it was decided to use a regression procedure. Voyce (1973) calculated the subject parameters on the basis of Desirability scale values for groups of items frcm the DPI, ordered by Desirability scale value, as the independent variablc, with the proportion of true responses to those item groups as the dependent variable in the regression equation. Salience for each respondent was taken as the slope of the best fitting least squares straight line for the regression of the respondent's proportion of true responses upon the item group mean scale values. The threshold was taken as the point on the abscissa at which the probability of a true response exceeded 0.5.

The method used here of obtaining estimates of salience and threshold combines the approaches taken by Rogers and Voyce. Salience was computed as the slope of the least squares regression line of the binary item responses upon the scale values for desirability. As the slope of this line is a direct function of the biserial correlation, this procedure is similar to that used by Rogers (1971) and Jackson (1968). The threshold was defined as the point on the abscissa (item scale values) at which the probability of a true response was 0.5. This involves rearranging the terms of the regression line to solve for a value of X when Y is 0.5. Thus the threshold is a function of the salience. This procedure is similar to that used by Voyce (1973), with the exception that in the latter case, the points in the regression line were means for groups of items rather than actual item scale values. Voyce (1973) and Rogers (1971) both found these two parameters of the model to be independent of one another in the groups of respc.dents which they studied.

In order to compare these procedures, Voyce's (1973) procedure was also used to calculate salience and threshold. Twenty-two groups of 16 items each were formed by placing the 16 items with the lowest Desirability scale values into group 1, the 16 items with the next highest scale values into group 2, and so on. The dependent variable was the number of true responses to each of these groups of items. There are thus 22 pairs of observations in the regression. Saliences (S_{ds}) and thresholds (T_{ds}) were calculated as described above.

Saliences and thresholds were also calculated directly using the binary item responses as described previously. In this case, the approximation of the regression to Figure 4-1 is much lower than the case with Voyce's procedure. As the previous chapter showed that Desirability scale values based upon judgements correlated only 0.6 with the basic structure content scores for Desirability, it was decided to calculate saliences and thresholds using both sets of scale values. Therefore, a regression line for each subject was calculated using each set of scale values and the binary responses to those items. Thus there is a salience based upon basic structure content scores for Desirability (S_r) and a corresponding threshold (T_r), and the equivalent parameters (S_j and T_j) calculated upon the successive intervals scale values for desirability reported by Helmes et al. (1977). These two sets of scale values for 352 items were the same as those intercorrelated in the previous chapter.

Correlations among the three sets of subject parameters are given in Table 4-3. The first three columns of Table 4-3 are based upon the saliences calculated from basic structure content scores (S_r), from successive intervals scale values from judgements (S_j) and from 16-item scales formed on the basis of the degree of item desirability (S_{ds}). The final three columns are based on the thresholds corresponding to the three saliences in the first three columns.

For both samples, the most important aspect of Table 4-3 is the independence of salience and threshold. The largest correlation between a salience and a threshold accounts for less than 4% of the total

Table 4-3. Correlations among Subject Parameters for Jackson's Threshold Model

	S_r	S_j	S_{ds}	Scale Score	T_r	T_j	T_{ds}
S_r	-	81**	81**	80**	-02	-18**	-13*
S_j	05	-	99**	77**	01	-13*	-07
S_{ds}	06	99**	-	75**	02	-13*	-07
Scale Score	16	75**	70**	-	05	-16*	-08
T_r	-18	06	06	01	-	03	-01
T_j	-06	04	04	02	-01	-	48**
T_{ds}	02	-23*	-23*	-35**	03	-01	-

Note: Decimals omitted.

Abbreviations: S_r - Salience based upon content scores from responses
S_j - Salience based upon judged scale values
S_{ds} - Salience based upon scores from desirability scales
T_r - Threshold based upon content scores from responses
T_j - Threshold based upon judged scale values
T_{ds} - Threshold based upon scores from desirability scales

Values above the diagonal are for the high school sample. Values below the diagonal are for the UWO sample.

* $p < .05$
** $p < .01$

variance. Even though the correlations are significantly greater than zero, they are of negligible importance and many of them fail to replicate in the two samples. This replicates the findings of both Rogers (1971) and Voyce (1973) for this aspect of Jackson's model. However, the presence of this number of statistically significant correlations does argue that the independence of these parameters is an empirical result and not an artifact of the computational procedures. In the high school sample, all methods of calculating salience are highly intercorrelated and correlate highly with the scores on the Desirability scale of the PRF. This is important, as the high correlations between S_{ds} and the other two saliences indicates that this parameter is highly stable. S_{ds} is calculated upon 22 observations with a multi-valued dependent variable. S_r and S_j use 352 observations and a binary dependent variable. The stability of a subject parameter across such differences in computational procedure is encouraging. Thresholds are largely uncorrelated, except for a moderate correlation between scales based upon item judgements and upon desirability scales. The latter finding is not replicated in the UWO sample, although the other correlations among saliences do replicate, as do the findings for thresholds. Another result unreplicated in the UWO sample is the correlation of salience based upon content scores with other saliences. The reason why these correlations are found only in the high school sample is unclear. Replication on a third sample of university students would appear to be necessary to clarify the relationships among these different methods of calculating the parameters.

In summary, the major finding with regard to the methods of determining the parameters is that the different thresholds are largely uncorrelated with one another. Saliences are more stable, correlating highly across methods of calculation, with the possible exception of the salience based upon content scores, which in turn are derived from item responses. In addition, Desirabiity scale scores correlate highly with saliences and may therefore be interpreted as a measure of salience.

Predicting Item Responses

To make predictions of item responses for an individual, Jackson's model was translated into a determinate (error-free) prediction form. Two sets of scale values for desirability were used, these having been shown to be somewhat different in the previous chapter. One set was the set of successive intervals scale values reported by Helmes et al. (1977). T_j was used as the respondent parameter to determine which items were predicted as true responses. The second was the set of basic structure content scale values, for which T_r was used as the respondent parameter.

In making predictions, the appropriate threshold for each individual was compared to a vector of the item desirability scale values as ranked by magnitude of scale value. The number of items above the threshold was determined and these items were then predicted as having a true response. Those items lying below the threshold were predicted as having a false response.

Results of the item predictions are given in Table 4-4. In the UWO sample, it can be seen that the two models predict equally well (p>.40). However, the accuracy drops for the high school sample when response scale values are used to calculate thresholds, whereas this does not occur if judged desirability scale values are used. As a result, the predictions based upon judged desirability scale values are greatly superior to those based upon response desirability scale values when the two samples are combined (p<.0001).

It should be noted that this version of Jackson's model predicts about as well as does the invariant model based upon desirability (p>.05). This is indicative of the potency of social desirability in determining item responses. A comparison of the accuracy for invariant and variable models will be discussed in more detail later.

The previous section showed that invariant models based upon content were relatively poor predictors of item responses. It remains to be seen if models based upon content which also incorporate individual differences would predict item responses significantly above the chance level. Accordingly, various individual differences models based upon content comprise the remainder of the prediction models to be described.

Table 4-4. Results for Cumulative Threshold Models based upon Desirability

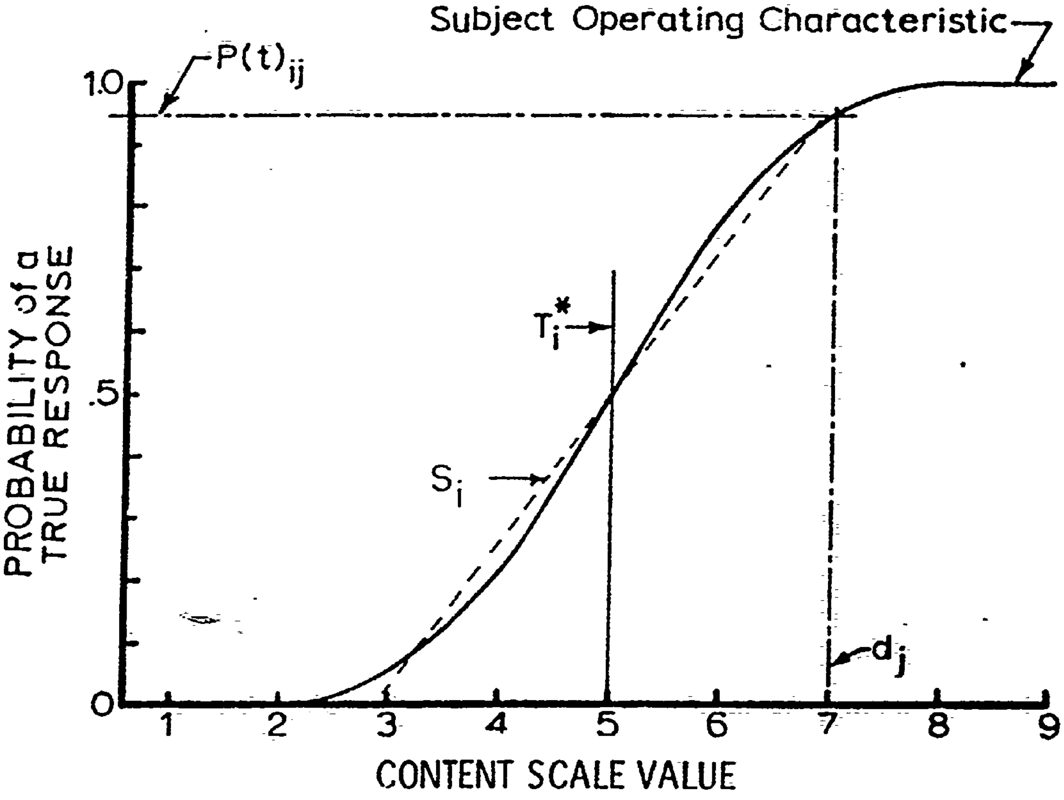
	Sample								
	UWO			High School			Pooled		
Model	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa
Response Values	18,250	56.36	.125	55,492	52.37	.046	73,742	53.31	.064
Judged Desirability	18,259	56.38	.126	64,738	61.10	.221	82,997	60.00	.199

All values of kappa are significantly greater than zero at the .001 level.

Cumulative Threshold Models for Content

Jackson´s (1968) model is based upon social desirability as the dimension underlying item reponses. Other work (e.g. Fiske, 1963; Kuncel, 1973) has indicated the utility of models using scale content as the underlying dimension in what can generally be termed a cumulative homogeneity model, following Loevinger´s (1948) tradition. A model of this sort (here called a cumulative threshold model, to distinguish it from Fiske´s (1963; 1966) model, from which it differs in some details) is illustrated in Figure 4-1. If one substitutes social desirability as the abscissa instead of scale content, this figure also illustrates Jackson´s threshold model (1968) for item responding. The two models are essentially the same, with the vital exception that Jackson´s model uses social desirability as the underlying dimension and the current models use dimensions of substantive content.

The probability of a true response is taken as being an increasing monotonic function of the item scale values for the trait construct in question. This function is termed the subject operating characteristic and has two basic parameters, salience and threshold. The salience of the characteristic curve defines the degree to which the content dimension is relevant to the respondent. Individuals with differing saliences are shown in Figure 4-2. Respondent A has a higher salience than Respondent B who in turn has a higher salience than does Respondent C. Therefore the content dimension is more important or relevant for Respondent A than for either other respondent in that the content dimension determines more of the item responses than do other factors.



Derivation of the Probability of a True Response on Item j, for Subject i with the following Parameters:

T_i^* = Content Responding Threshold = 5

d_j = Content Scale Value of Item = 7

$P(t)_{ij}$ = Probability of a True Response = .94

S_i = Salience of Dimension = 1.9

Figure 4-1. Cumulative threshold model of the response process

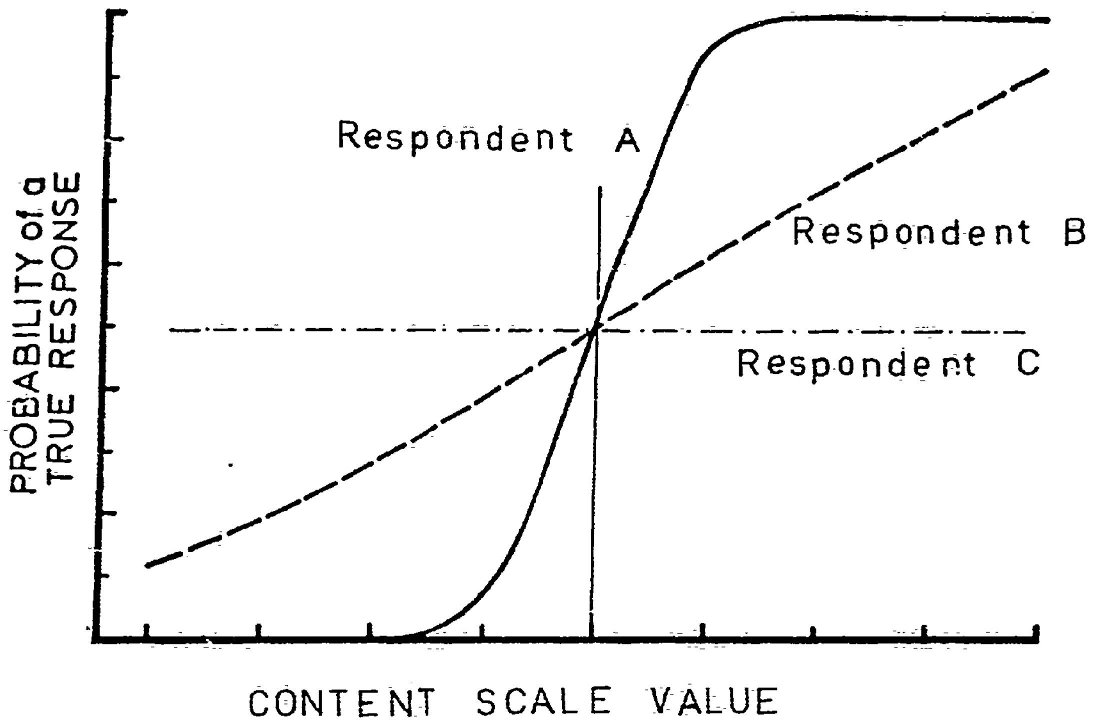


Figure 4-2. Subject characteristic curves with different saliences

Respondent C has an extremely low salience which indicates that the content dimension is essentially irrelevant in determining responses for this particular individual.

The second parameter of the characteristic curve is the threshold. This determines the point at which the probability of a true response for an individual exceeds 0.5. In other words, the threshold separates true responses from false responses. This point can occur at any point along the content dimension, as indicated in Figure 4-3. Here, both respondents have the same salience, but Respondent E has a higher threshold than does Respondent D and therefore responds true to fewer items.

Both parameters relate to a single dimension of content. Obviously, then, one must obtain a threshold and salience for each respondent for each scale of a multiscale inventory. Alternatively, one may prefer an approach which does not assume that each scale defines a separate construct. In this case, one might prefer the use of factors composed of linear composites of scales and calculate saliences and thresholds for the factor constructs rather than for the scale constructs. The basic structure scaling procedure described in Appendix 1 is flexible enough that scale values for either set of constructs can easily be determined.

The method of computing salience and threshold has already been mentioned for Jackson´s threshold model. Salience is the slope of the best-fitting straight line for the regression of the binary item responses upon item scale values. Threshold is the point at which the

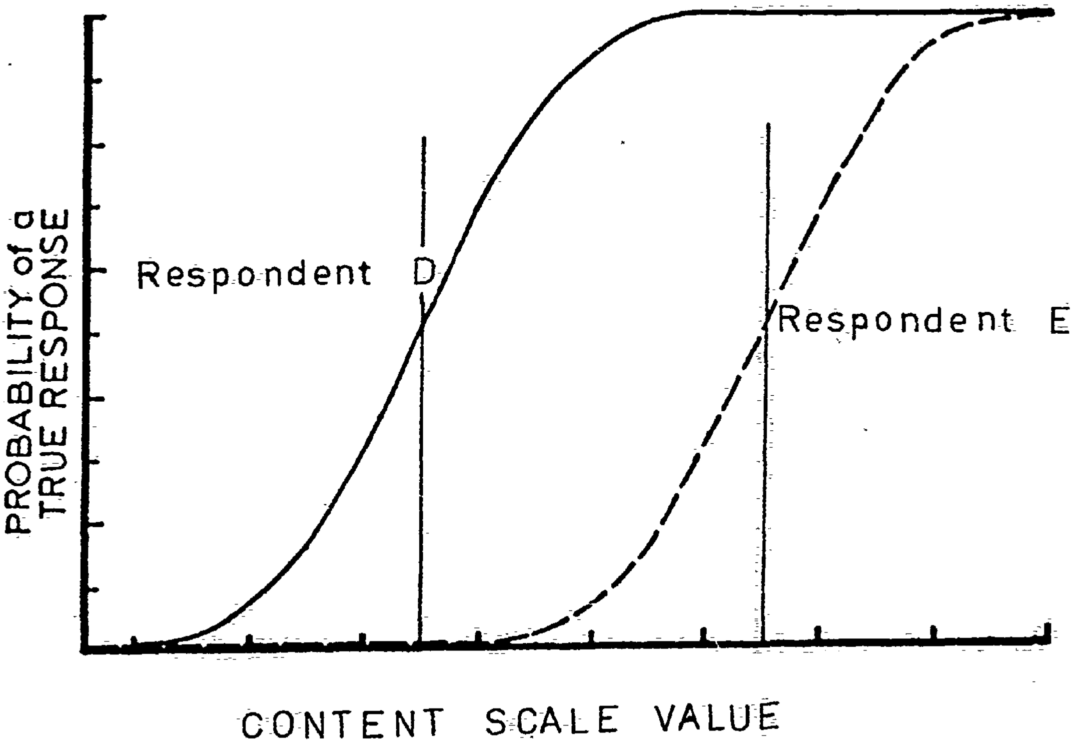


Figure 4-3. Subject characteristic curves with different thresholds

probability of a true response exceeds 0.5. This procedure was used over Voyce's (1973) regression method using scale scores as the previous section on Jackson's threshold model for desirability showed that the two methods gave very similar results. Saliences correlated 0.99 and thresholds correlated 0.48 for the high school sample (Table 4-3). In addition, in some variants of these models, there are insufficient items to form the groups of items required by the Voyce method. For example, one can calculate the respondent parameters on the basis of scale values for only the items keyed on a given dimension (n=16), or on the basis of the scale values for all items (n=352), regardless of whether they are relevant to the dimension or not. This latter possibility can increase the number of items on which the parameters are calculated severalfold. However, this may be done at the cost of introducing a large amount of random noise from the irrelevant items which may result in less useful values for salience and threshold than if the shorter set of scale values were used. Which set of item scale values would prove to be superior in predicting item responses thus becomes a question to be answered empirically.

Evaluation of the Respondent Parameters

In order to evaluate questions concerning the respondent parameters, it is necessary to calculate the parameters in each different manner outlined above. If this is done, one can then determine the extent to which saliences and thresholds are independent of one another, as well as the degree to which different methods of calculation give similar values of the respondent parameters. One of the virtues of the use of scale values from the basic structure scaling method of Appendix 1 is that scale values on a given dimension are indeed truly unidimensional and independent of other dimensions.

Therefore, saliences and thresholds were calculated both upon short vectors for only the 16 items keyed on a scale (S_s and T_s) and for long vectors of all 352 PRF-E items (S_1 and T_1), using the basic structure content scores described earlier. The computational procedures used to obtain salience and threshold for respondents were identical to those described in the previous section, with the exception that the Voyce (1973) method was not used. The four sets of respondent parameters were intercorrelated as before. In addition, PRF-E scale scores were included in the analysis. Scale scores are the most commonly used pieces of information about respondents which are obtainable from personality inventories. It would therefore be advantageous to know the relationship of scale scores to the parameters of the cumulative threshold model. This is particularly true in view of the correlation of scale scores with saliences found for Jackson's threshold model.

The relevant correlations are reported in Tables 4-5 and 4-6. It can be seen that there are no essential differences between the two sets of results. Therefore, they shall be discussed together except where noted.

The most important conclusion derived from Tables 4-5 and 4-6 is that saliences and thresholds are largely independent of one another. Although there are a few significant correlations of this sort in Table 4-6, these fail to replicate in the high school sample. This

Table 4-5. Correlations among Subject Parameters (High School Sample)

Scale	S_l-S_s	S_l-T_l	S_l-T_s	S_l-SS	S_s-T_l	S_s-T_s	S_s-SS	T_l-T_s	T_l-SS	T_s-SS
ABA	74	-03	01	74	-04	01	90	02	01	-00
ACH	84	-00	06	84	01	02	95	-04	05	07
AFF	85	-04	06	85	-07	06	99	-01	-05	06
AGG	81	-02	-02	81	-04	-02	97	03	-03	-00
AUT	87	-00	-03	86	-04	-03	97	-00	-02	-02
CNG	75	02	-10	73	02	-04	95	05	00	-03
CST	84	-02	-01	83	-03	02	98	-04	-03	02
DEF	84	-02	-00	83	-04	-01	98	04	-05	-01
DOM	94	-02	01	94	-03	-01	99	07	-03	-01
END	82	04	02	81	00	-02	97	-03	-00	-01
EXH	89	-01	-03	89	-02	00	99	00	-01	-01
HAR	93	-01	-04	92	-04	-01	99	-01	-04	-01
IMP	89	-01	01	89	02	-02	98	10	02	-01
NUR	86	-01	03	86	00	01	99	06	02	01
ORD	96	-01	03	96	-01	-00	99	17	00	00
PLY	82	-05	05	82	-05	01	97	01	-07	01
SEN	84	-02	-01	83	01	02	98	04	-00	00
SCR	86	-00	-09	86	-03	-04	99	00	-02	-04
SUC	90	-02	02	89	-02	-01	99	00	-02	-01
UND	91	03	00	90	03	03	98	-01	02	02
INF	34	07	-08	35	05	-10	98	10	05	-08
DSY	82	-02	-01	80	02	-04	94	-09	05	-06
Mean	833	-008	-006	828	-012	-007	974	021	-007	-004

Note: Decimals omitted.

Abbreviations: S_l - salience based on long vectors; S_s - salience based on short vectors; SS - scale score; T_l - threshold based on long vectors; T_s - threshold based on short vector.

A correlation of .181 is required to reject the null hypothesis of $\rho = 0$ at $\alpha = .01$ with 300 df.

Table 4-6. Correlations among Subject Parameters (University Sample)

Scale	S_1-S_s	S_1-T_1	S_1-T_s	S_1-SS	S_s-T_1	S_s-T_s	S_s-SS	T_1-T_s	T_1-SS	T_s-SS
ABA	76	-09	13	75	-10	-01	93	-01	-06	-02
ACH	82	02	07	81	07	08	95	-07	02	17
AFF	84	-08	-02	81	-18	-01	98	02	-19	00
AGG	84	00	14	83	-04	-04	98	03	-03	-05
AUT	84	04	07	84	-03	02	97	-04	00	01
CHG	81	-07	-02	76	-07	-07	96	09	-12	-11
CST	84	-05	10	82	-14	08	98	-07	-17	09
DEF	85	13	-00	83	07	-11	98	06	02	-11
DOM	94	05	-03	93	02	-04	99	-00	03	-07
END	80	-03	-07	78	-08	-03	97	-16	-10	-02
EXH	90	-03	07	90	-05	02	99	29	-05	03
HAR	95	-05	04	95	-10	01	99	15	-09	00
IMP	92	05	-01	92	-01	02	99	11	-00	-00
NUR	80	-12	11	82	-11	22	99	-01	-10	21
ORD	97	-04	04	97	-03	06	99	17	-04	08
PLY	87	06	-03	88	-01	-11	98	05	-03	-07
SEN	77	-13	-09	75	-32	-10	97	-03	-25	-07
SCR	89	-01	-16	89	-03	-09	99	-10	-05	-09
SUC	91	-03	-05	90	-04	-06	99	-03	-02	-06
UND	90	-03	03	89	-01	02	97	12	-03	02
INF	25	14	-12	23	08	-36	94	02	06	-38
DSY	77.	20	-18	76	05	-10	92	00	01	-11
Mean	829	-003	001	818	-047	-037	973	027	-054	-026

Note: Decimals omitted. Abbreviations as in Table 4-4.
A correlation of .267 is required to reject the null hypothesis of $\rho = 0$ at $\alpha = .01$ with 90 df.

independence of the two respondent parameters replicates the findings of Voyce and Jackson (1977) and those reported here earlier for the desirability threshold model. The next item of interest is that the saliences computed by both methods are highly correlated with each other and with the scale scores. Conversely, the two sets of thresholds are essentially uncorrelated. This indicates that the salience is a much more robust respondent parameter than is the threshold in that relatively similar estimates of it can be obtained from either long or short vectors of item scale values. Of more interest is that the scale score is correlated quite highly with both saliences and particularly with the saliences based upon the short vector of scale values for items keyed on that scale. This would indicate that in many cases there is no need actually to calculate the regression line between item scale values and responses if one is interested solely in the salience of a scale for a particular respondent.

As can be seen in Tables 4-5 and 4-6, there is little or no similarity among the thresholds calculated from vectors of different lengths of item content scale values. This lack of correlation implies that thresholds are extremely sensitive to the manner in which they are calculated, quite unlike the saliences. The fitting of a line to a set of data is very sensitive to the distribution of the data. In this case, the lines from which the concepts of salience and threshold are calculated are based upon quite different distributions. In the case of the long vectors, the distribution of scale values is approximately normal, that is, unimodal and fairly symmetric. In the case of the short vectors, the distribution is markedly bimodal. However, the similarity is great enough that these correlate quite highly. The threshold, however, is a function of the salience. A slight change in the salience may therefore result in a rather large change in the threshold. Table 4-6a gives the mean threshold and salience for the 22 scales of the PRF. It can be seen that the values of the salience are near 0, and, as a result, large changes in the threshold may easily occur. This argument is supported by the magnitude of the variances. The variance of the salience is one-tenth the value of the mean. The variance of the threshold averages approximately two to four hundred times the average value of the mean threshold. This factor could in itself account for the lack of correlation among thresholds calculated upon the vectors of different length. To completely eliminate differences in distribution in form from the calculation of thresholds, one would need a completely uniform distribution of the items across the content dimension and select the short vectors as a random sample from the full distribution. Of course, a uniform distribution is required for the most accurate fitting of a curve to any set of data.

Prediction Models

The instability of the thresholds in turn implies that they would not serve as highly accurate predictors of item responses. At the very least, however, it remains to be seen what degree of predictive accuracy can be obtained from thresholds and which set of thresholds proves to be the most accurate predictor. In order to determine this, four sets of thresholds were used in predicting item responses. The two distinguishing factors were the number of scale values used in

Table 4-6a. Mean Saliences and Thresholds for PRF-E Scales

	Salience, Long Vector		Salience, Short Vector		Threshold, Long Vector		Threshold, Short Vector	
	Sample		Sample		Sample		Sample	
Scale	A	B	A	B	A	B	A	B
ABA	-.02	-.02	-.04	-.02	-.52	-2.62	1.51	1.25
ACH	-.01	.02	.04	.05	.33	-2.66	.07	-.12
AFF	.03	-.00	.05	-.01	4.76	2.59	-1.33	-1.55
AGG	.01	.03	.01	.06	.74	3.01	1.21	-.05
AUT	-.00	-.02	-.01	-.05	1.31	5.31	.16	2.36
CHG	.04	.04	.04	.04	-.32	.14	2.29	2.85
CST	.01	.02	.02	.03	1.23	.30	-.61	-.54
DEF	-.03	-.04	-.02	-.04	-2.11	20.42	1.75	-2.29
DOM	.01	.03	.01	.04	.68	-12.21	1.33	4.74
END	.03	.04	.01	.02	-.79	-.03	2.42	10.08
EXH	-.00	.00	-.01	-.01	-.02	-.19	1.75	-.76
HAR	-.01	.00	-.02	.00	-1.01	-.10	-2.03	-.05
IMP	-.01	-.03	-.01	-.04	-.01	1.03	-.91	3.08
NUR	.03	.04	.04	.08	-.10	.36	2.44	-.48
ORD	-.02	.00	-.02	.01	-.37	.03	-2.40	1.54
PLY	.03	.02	.05	.03	2.80	-17.42	-1.17	1.58
SEN	.03	.04	.03	.06	.02	113.25	-.44	1.25
SOC	.01	.01	.02	.02	-.89	-.98	.25	-.07
SUC	-.01	-.00	-.01	-.01	-.91	-.80	2.38	1.68
UND	-.02	-.00	-.03	-.01	15.75	-1.17	1.32	16.02
INF	-.15	-.16	-.19	-.19	.04	.08	.14	.10
DSY	.03	.01	.04	.05	.08	12.12	-.03	-.95

Note: Sample A is the high school sample; Sample B is the UWO sample.

calculating thresholds and the type of scale value used. Either short vectors of only items keyed on a construct were used, or long vectors consisting of scale values for all items. The second factor was the use of either constructs defined by PRF-E scales or of factors as defined by the six factors found by Skinner et al. (1976). The basic structure scaling procedure was used to obtain the scale values for items on the factor constructs. These constructs have from 16 to 80 items keyed on them, with the normal range of scale values within the interval of -4 to +4.

In accordance with the model described earlier, the basic approach of the cumulative threshold model for predicting responses is this: assume that m scale values are defined for a dimension of content and can be ranked on that dimension in order of the magnitude of the scale value, as in Figure 4-1. Then determine the number and identity of the items which have scale values which are in excess of the individual's threshold. The items above the threshold are those which are predicted to be responded to as true. Items below the respondent's threshold are predicted as being responded to as false. In Figure 4-4 the length of the horizontal bars indicates the ranked scale values for a set of items relevant to a construct. Respondent A has a low threshold, and therefore it is predicted that he will respond true to the eleven items which have scale values above his threshold. Respondent B has a high threshold for this particular construct and is predicted to respond true only to the three items above his threshold. In addition to predictions for the four sets of thresholds, a set of predictions was made using thresholds calculated from vectors of 16 scale values for scale

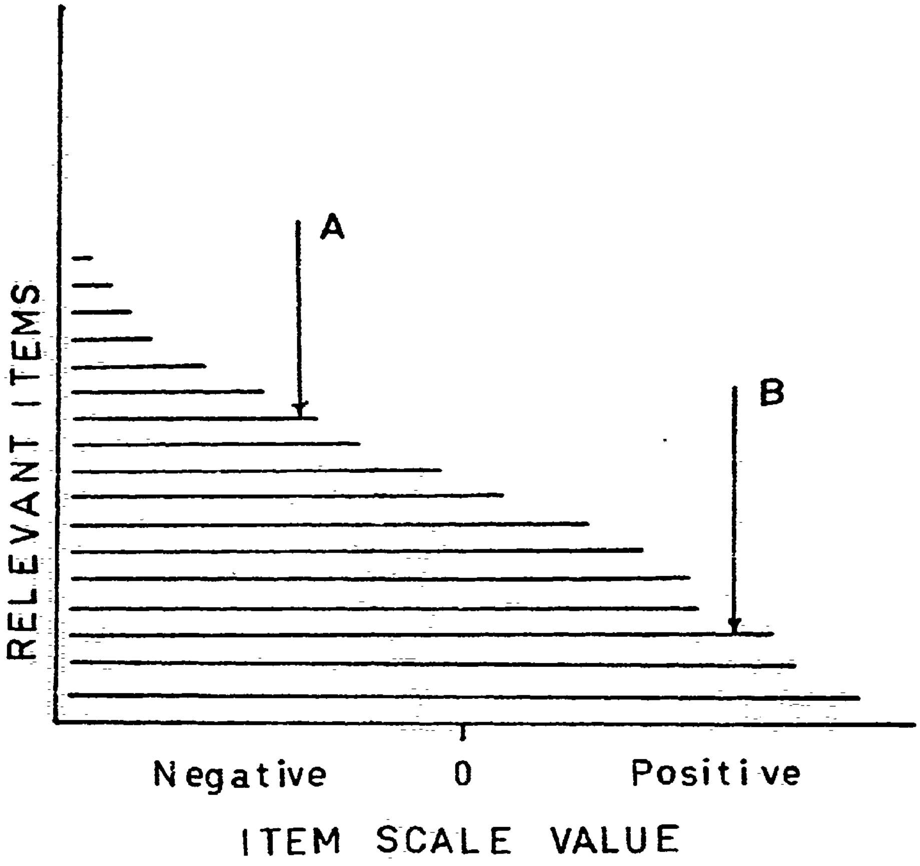


Figure 4-4. Illustration of thresholds and the number of items responded to as being true

constructs, but in which the scale values were not ranked, but ordered randomly. As this destroys the cumulative nature of the ordering of the items, this procedure leads to a set of predictions which are essentially random.

Table 4-7 reports the outcome of predictions using the different variants of the cumulative threshold model. For the UWO sample, all variants of the cumulative threshold model predicted item responses slightly better than chance, including the random model. Post-hoc tests of mean values of kappa following a randomized-block analysis of variance showed no differences between the random model and the long vector, scale content model. The other three models were significantly better (p<.01), but they did not differ among themselves. Analysis of variance results are contained in Appendix 3. Predictability dropped for the high school sample, with the result that only predictions for the random model are above the chance level. The analysis of variance showed that the random model was superior to the other four models (p<.01), which did not differ significantly among themselves. The loss in predictability from the university sample is probably due to a lack of generalizability of the content scores derived from college students to the high school sample. This effect was also noted for the content scores for Desirability in the previous section. The particularly poor predictions for the high school students by the models using scale values for factors is noteworthy, these predictions being significantly worse than chance, although not significantly worse than the other models. Unpublished work has indicated that the factor structure of PRF-E for this sample of high school students is not identical to that

Table 4-7. Results for Cumulative Threshold Models based upon Content

	Sample								
	UWO			High School			Pooled		
Model	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa
Scale Content, Short Vectors	16,912	52.22	.045***	53,041	50.01	.001	69,953	50.57	.011***
Scale Content, Long Vectors	16,489	50.92	.019***	53,095	50.11	.002	69,584	50.30	.006*
Factor Content, Short Vectors	17,082	52.75	.055***	52,361	49.42	-.012***	69,443	50.20	.004
Factor Content, Long Vectors	16,964	52.38	.047***	52,442	49.50	-.010***	69,406	50.17	.003
Random Content, Short Vectors	16,498	50.95	.019***	53,853	50.83	.017***	70,351	50.86	.017***

* p<.05
*** p<.001

of the military applicants used by Skinner et al. (1976). The target matrix used in calculating the content scores for factors was based upon the latter group, which was composed entirely of males. This difference may account for some of the loss in predictability.

In summary then, it is apparent that predictions from all forms of the threshold model based on content are very low and indistinguishable from random predictions in many cases. This does not necessarily imply that the concept of threshold for content is useless. It is possible that the utility of the threshold is apparent only when attempts are made to produce changes in the level of the threshold through experimental manipulations. Another possibility is that the incorporation of another dimension such as social desirability into the threshold model would improve the accuracy of its predictions.

Two-dimensional Cumulative Threshold Models

In the prediction models used above, predictions are made for one scale or factor at a time, using orthogonal scale values for items. This procedure minimizes the role of social desirability in item responding. A more realistic view of the item response process would be that both item content and item desirability are relevant to an individual responding to the items of a personality inventory. This calls for a somewhat more complex model of the type illustrated in Figure 4-5. Here, instead of a subject characteristic curve, we have a subject characteristic surface, with the probability of a true response being a function of both item content and item desirability. As item content becomes more extreme in the positive direction, the probability of a true response increases. It also increases as item social desirability becomes more positive. Respondent saliences and thresholds can be calculated as before on the two dimensions of content and desirability. Thus a true response can result if either or both of the item scale values exceeds the relevant threshold for that individual.

Item Predictions using a Two-dimensional Model

The prediction procedure used for this model is essentially a combination of the models of the two previous sections. The thresholds calculated previous:, for both content for each scale and for desirability in each sample were used. The number and identity of items above each threshold were then determined and a true response was predicted for a given item if it was above either the individual's

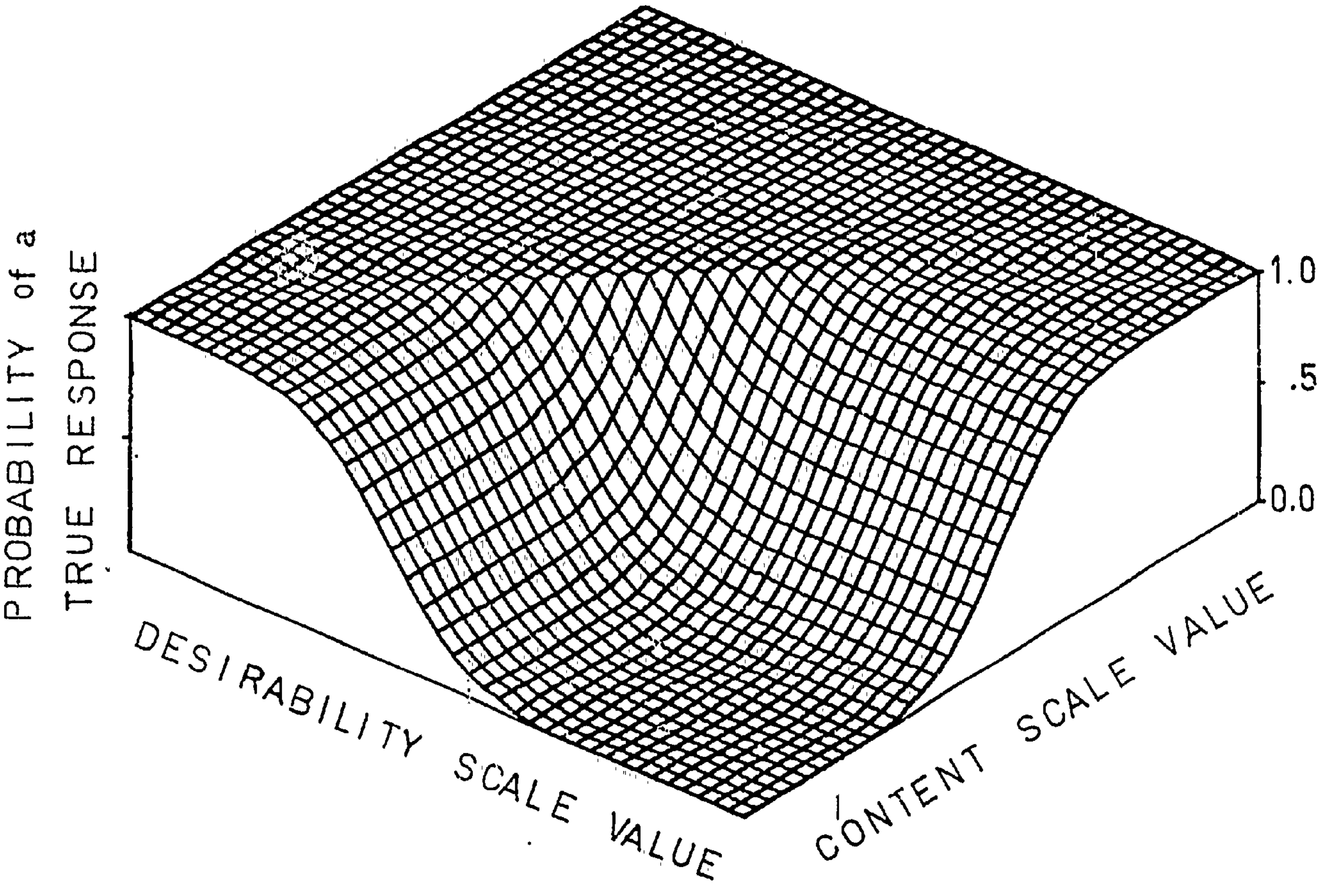


Figure 4-5. Illustration of a two-dimensional cumulative model

threshold for content or above his threshold for desirability. This was done for both content dimensions based upon factors and for content dimensions based upon scales.

The results of these prediction models for the two samples are given in Table 4-8. Both models are equally accurate for both samples ($p>.05$ in each case), although once again, predictions for the UWO sample are more accurate than those for the high school sample ($p<.001$ in both cases). This is also undoubtedly due to the shrinkage in predictability in transferring from one population to another. The gain in predictability of two-dimensional models over the simple cumulative threshold models is almost certainly due entirely to the superior predictability of desirability. A two-dimensional model using both desirability and content thresholds therefore has no advantage over the unidimensional model using desirability alone. In addition, it requires more parameters and yet achieves a lower level of accurate prediction.

At this point, it is perhaps relevant to point out that the prediction procedure used in these models and in Jackson's threshold model are not based on least-squares linear regression. Predictions are based upon the intermediate step of determining the number of items above and below the threshold and are made directly as either a true or false response. A regression procedure would give a continuous approximation to these binary values. Therefore, accuracy of prediction is not optimized through the inclusion of values being predicted or through capitalization upon chance. Accordingly, procedures such as

Table 4-8. Results for Two-Dimensional Models based upon Thresholds

Model	Sample UWO Correct Predictions	 Percentage	 Kappa	 High School Correct Predictions	 Percentage	 Kappa	 Pooled Correct Predictions	 Percentage	 Kappa
Scale Content, Desirability	19,273	59.51	.182	59,682	56.33	.122	78,955	57.07	.136
Factor Content, Desirability	19,320	59.66	.187	59,457	56.12	.118	78,797	56.96	.134

Note: All values of kappa are significantly greater than zero at the .001 level.

cross-validation or jack-knifing (Mosteller & Tukey, 1968) are not necessary.

It remains possible that a different type of model based upon content would predict responses more accurately than the threshold models just described. These models are discussed next.

Cumulative Scale Score Models

The previous section showed that models of the response process based upon thresholds were rather poor predictors of item responses. Therefore, it is appropriate at this time to turn to the other respondent parameter of the cumulative model, the salience. Saliences were stable across methods of calculation and so have this initial advantage over thresholds as predictors of item responses.

The use of saliences appears to be equivalent to rejecting the cumulative threshold model as a valid model of the response process. According to the formulation of this model, the threshold determines the point along the item continuum at which false responses end and true responses begin. Saliences are an integral part of the cumulative threshold model, but are not the determinant of the direction in which items are answered. To use the salience as a predictor of items means using it in an entirely different model.

Prediction Models using Saliences

It will be recalled that scale scores were correlated an average of 0.82 and 0.97 with saliences calculated upon long and short vectors of item scale values respectively (Tables 4-5 and 4-6). From this, one can argue that scale scores are a measure of the salience of a dimension for the individual. The use of scale scores rather than actual saliences thus maintains a connection of this model to the cumulative threshold model. In addition, the use of scale scores in prediction models has several advantages. First, scale scores are quicker and easier to determine than are saliences. A manual scoring template for a response sheet is all that is required, as opposed to the regression procedures required to calculate a salience. This is true at least for dimensions for scale constructs. For other constructs, such as those based upon factors, the regression procedures are usually required because the differential weighting of the scales on the factors will generally not be accurately known. Second, because the use of scale scores sidesteps regression procedures entirely, it thus avoids criticism on the grounds of optimized prediction through capitalization on chance. Third, scale scores are easily translatable into a format for item prediction.

In the cumulative threshold model, salience indicates the degree of relevance of the dimension to the individual. Therefore, if the salience is high, the dimension is important to the individual. If it is low, then the dimension is relatively unimportant to the individual. It would then follow that an individual for whom a dimension is important would tend to endorse more items in the keyed direction than would an individual to whom the dimension is less important. If we make the latter assumption, then we can predict that an individual with a high salience will answer more items in the keyed direction than will an individual with a lower salience. Therefore, if we utilize the connection between salience and scale score demonstrated in the previous section, we can use saliences indirectly in making predictions about item responses. If one individual has a scale score of 10 and another has a scale score of 8, a scale score or salience model would predict that the 8 items to which the second individual responds in the keyed direction are those with the most extreme scale values. The first individual would respond to those same 8 items in the keyed direction and also to the next two most extreme items. Notice that this is a quite specific prediction. It thereby avoids the simple tautology that an individual with a scale score of 10 would answer 10 items in the keyed direction. The prediction specifies exactly which 10 items would be answered in the keyed direction.

Obviously, such a model is only suited to scale constructs where the assumption of unidimensionality can be defended. The alternative is the use of dimensions from a Multidimensional Scalogram Analysis (Lingoes, 1972) which form Guttman scales. The use of basic structure content scores (Appendix 1) as the criterion for ordering items meets this requirement of unidimensionality. Each set of scale values is unidimensional, as it is based upon a single rotated component. In addition, the sets of scale values may also be orthogonal to one another, if the appropriate rotational procedure was used in obtaining the scale values.

Scale Score Prediction Models

Two scale score or salience models of this type were used to predict responses. The first was a simple model in which the scale scores for a given scale were used to predict the items on that scale. The second model was two-dimensional, in that the salience for desirability was also assumed to affect item responses, regardless of the salience of the content dimension. In this case, the salience for desirability was subjected to a linear transformation to expand its range to encompass all items and not simply those keyed on desirability. This was done by multiplying each scale score by 22, the number required to make a scale score of 16 equivalent to the maximum number of PRF-E items. This transformed salience was then used to determine what degree of extremity of item social desirability was salient for that individual over all items. If these extreme items were not predicted as true responses on the basis of content, they were then predicted true on the basis of desirability. All items not below the level of endorsement predicted by individuals´ saliences for content or for desirability were predicted as having false responses.

The outcome of predictions for these two models is given in Table 4-9. Overall, these models predict item responses quite well, over 60% of the predictions being correct. Although the addition of desirability to content significantly improves the accuracy of prediction for the UWO sample ($p<.001$), this does not occur for the high school sample ($p>.05$).

It would thus appear that scale scores (saliences based upon content) are fairly good predictors of item responses, superior to the best models based upon thresholds using item content (that for scales based on short item vectors, $p<.0001$). In addition, a scale score model based upon content is also more accurate than Jackson´s threshold model for responding on the basis of item desirability. Given the degree of attention given to the suppression of desirability responding during the construction of PRF-E (Jackson, 1974), such a result might be expected. For other inventories, in which such an effort had not been made, it is

Table 4-9. Results for Cumulative Models based upon Scale Scores

	Sample								
	UWO			High School			Pooled		
Model	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa
Scale Content	21,090	65.12	.303	67,448	63.66	.274	88,538	64.00	.281
Scale Content, Desirability	21,196	65.45	.310	67,697	63.89	.278	88,893	64.26	.286

Note: All values of kappa are significantly greater than zero at the .001 level.

very likely that desirability would be much more relevant.

The most complex models described to this point are two-dimensional with only one of these dimensions being a content dimension. There is no necessity of this restriction, or of the number of dimensions being limited to two. A model explicitly based upon the multidimensional nature of item content is the final model to be described and evaluated.

Multidimensional Spatial Models

A different approach to the cumulative models proposed here and by other workers such as Fiske (1966) is that of Cliff (1968), who has developed a multidimensional cognitive model of the item response process. Cliff´s model states that an individual´s response to an item is a function of the item´s location in a multidimensional cognitive space. The dimensions of that space constitute the dimensions of subjective meaning and the projections of the item upon those dimensions define its subjective meaning. In early work Cliff (1968) demonstrated a degree of utility for his model and has elaborated it primarily in the methods by which the multidimensional space is defined (Cliff, Bradley & Girard, 1973) and in using personality items rather than descriptive adjectives (Cliff, 1977). Note that in this model one can deal either with a space of common meaning of the items or develop a space that is unique to a given individual´s perception of the meaning of the set of items. Such a method thus has the potential for adapting to individuals whose perceptions of the meaning of words differs from the general consensus (Loehlin, 1961; 1967).

However, it should be noted that in the case in which one deals with a space of common subjective meaning, one is dealing with a straightforward multiple regression in which the item projections onto the dimensions of meaning are the independent variables and the item responses comprise the dependent variable. If the item scale values have any degree of validity, then a fairly good prediction is guaranteed, particularly with a large number of predictors that exhaust the space of subjective meaning. Testing for the significance of the multiple correlation in this case (as Cliff does) amounts to a test of whether or not item content is relevant to the item response and says nothing about the validity of a spatial model of subjective meaning.

Therefore, a comparison of Cliff´s approach with the cumulative model involves not only a comparison of two theoretical models but also a contrast of the predictive power of cumulative models with that of a powerful statistical technique for prediction.

Multidimensional Spatial Predictions

In this study, it was assumed that the basic structure content scores constituted a valid set of item scale values for the dimensions of subjective meaning. Cliff (1977) utilized a multidimensional scaling procedure based upon judgements to obtain scale values for items. The basic structure content scaling procedure (as outlined in Appendix 1) provides the response aspect of multidimensional scaling methods of this sort. The set of 22 scale values per item for the 352 PRF-E items was used for the first set of predictions. A second set of predictions was based on the set of scale values for the six factors found by Skinner et al. (1976). The scale values were used as the independent variables and item responses as the dependent variable in a conventional least-squares linear multiple regression procedure. One multiple regression was calculated for each set of scale values for each individual in each sample.

The results of these predictions are given in Table 4-10. Well over 70% of the responses are accurateiy predicted by both versions of the multidimensional spatial model. This level is achieved using content scores derived from the U. S. colleg. sample. This is a further demonstration of the generalizability of the basic structure content scores. The utility of these scores is further demonstrated by the fact that the 22 components accountel for only 50.8% of the variance during the calculation of the content scores. Six components accounted for 34.1%. Thus, more items can be accurately predicted than the proportion of reliable variance accounted for in the original decomposition would lead one to expect could be accurately predicted.

In both samples, 22 scales predicted r.sponses more accurately than did six factors (p<.001). This is hardly surprising, as 22 components will always account for an equal or larger percentage of variance than will six components. Therefore, the predictions based upon a larger propertion of the variance should be more accurate.

Once again, predictions for the UWO sample were more accurate than those for the high school sample (p<.001 in both cases). Again, this is probably due to the transfer effect of scores derived from one population being less valid for another population.

A further note of caution regarding these predictions is in order. Cliff's model does use least-squares regression to predict values that were used in deriving the original regression, unlike the threshold and scale score models discussed earlier. This of course does capitalize on chance and enhances the accuracy of prediction above the level which

Table 4-10. Results for Multidimensional Spatial Models

	Sample								
	UWO			High School			Pooled		
Model	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa	Correct Predictions	Percentage	Kappa
Scales (22)	25,093	77.49	.550	78,892	75.40	.508	104,985	75.89	.518
Factors (6)	24,087	74.38	.487	75,883	71.62	.432	99,970	72.27	.445

Note: All values of kappa are significantly greater than zero at the .0001 level.

would be expected if this had not occurred. One solution would be to cross-validate - to use the regression weights derived from one set of items to predict another set and vice versa. Another alternative would be to jack-knife (Mosteller & Tukey, 1968). As the goal of the above section is to compare the maximum level of predictive accuracy of the various models, the optimal predictions are reported.

However, an estimate of the magnitude of the optimizing effects of the regression procedure was obtained by cross-validating the predictions for the UWO sample. Two sets of predictors and criterion were formed by taking half the items of each scale, with equal numbers of true- and false-keyed items, into each set. Prediction weights were determined for each set of items and were used to predict the item responses for each set of data. The UWO sample with scale predictors was used for cross-validation, as this combination showed the highest level of predictive accuracy (Table 4-10). It was based upon the most similar sets of data, and would therefore give a conservative estimate of the amount of shrinkage to be expected in the other cases.

When each set of scale values is used to predict its matched set of 176 responses for each individual, 25,860 responses (79.85%) are predicted correctly (kappa = 0.597). This level of predictive accuracy is significantly higher (p<.01) than the case in which all items are used to predict. This effect is probably due to there being fewer irrelevant items (items not keyed on the same scale as the items being predicted, with scale values more subject to error) in the case when half the items are used than in the case when all items are used.

When the regression weights fcr respondents are used to cross-validate the predictions, 23,408 (72.28%) responses are correct (kappa = 0.446). This represents a drop of over 5% in predictive accuracy over the non-cross-validated case in which all the items are represented in the predictors. If the predictions for the multidimensional spatial model based upon factors had also been cross-validated, it is likely that the shrinkage would be even greater, as there is less reliable variance being used in the predictors. Nevertheless, the 5% figure gives a conservative estimate of the amount of optimization involved in the predictions made by the multidimensional spatial model.

Evaluation of the Models

The results of the various prediction models are summarized in Table 4-11 from the appropriate tables in previous sections.

Under the assumption that the most accurate prediction models are the most valid, Table 4-12 reports the accuracy levels for the seven models with predictive accuracies in excess of 60 percent. Using this as a minimum criterion of accuracy of prediction, no other models will be discusssed further. The most accurate models are the two multidimensional spatial models, followed by those two based upon scale score models. There is a sharp drop in accuracy between these pairs, from over 70% to 64%. After this point, there are no major differences. The randomized block analysis of variance (reported in Appendix 3) showed all models differed from one another at the .01 level. The sole consistent exception to this was the two scale score models, which did not differ from each other in both samples.

The spatial models therefore appear to be the most valid of those models considered here, as they are substantially more accurate predictors than the next most accurate model (p<.0001). However, it will be recalled that Cliff's (1977) formulation uses multiple regression as its predictive model. This approach places other models at a severe disadvantage in that it is the only model which requires the knowledge of the item response in order to predict that same response. No other model has this requirement, which gives Cliff's approach an exaggerated accuracy. If this requirement is dropped, as for example, if cross-validated predictions are made, the accuracy of this model

Table 4-11. Summary of Results of Prediction Studies

Model	Number Correct	Percentage	Kappa
Invariant			
Random	68,272	49.35	-.013
Scoring Key	70,632	51.06	.021
Disjunctive Content	69,114	49.96	-.001
Judged Desirability	83,387	60.28	.204
P-value	87,450	63.22	.265
Cumulative Threshold, Desirability			
Response Desirability	73,742	53.31	.064
Judged Desirability	82,997	60.00	.199
Cumulative Threshold, Content			
Random	70,351	50.86	.017
Scale Content, Short	69,953	50.57	.011
Scale Content, Long	69,584	50.30	.006
Factor Content, Short	69,443	50.20	.004
Factor Content, Long	69,406	50.17	.003
Scale Content + Dsy	78,955	57.07	.136
Factor Content + Dsy	78,797	56.96	.134
Scale Score			
Scale Content	88,538	64.00	.281
Scale Content + Dsy	88,893	64.26	.286
Multidimensional Spatial			
Scales	104,985	75.89	.518
Factors	99,970	72.27	.445

Table 4-12. Summary of the Seven Most Accurate Prediction Models

Rank	Model	Correct Predictions	Percentage
1	Multidimensional Spatial, Scales	104,985	75.89
2	Multidimensional Spatial, Factors	99,970	72.27
3	Scale Score, with Desirability	88,893	64.26
4	Scale Score, Unidimensional	88,538	64.00
5	Invariant, p-values	87,450	63.22
6	Invariant, Desirability	83,387	60.28
7	Cumulative Threshold, Judged Desirability	82,997	60.00

would be much lower. Using the conservative figure of 5% shrinkage arrived at in the previous section, the accuracies of the two multidimensional spatial variants would be reduced to approximately 71% and 67% for the scale and factor variants respectively. These levels of accuracy are not as dramatically higher than the figures for the next most accurate models. In addition, the multidimensional spatial model requires more parameters for each item prediction than does any other. The factor variant has six independent variables for each item, the scale variant has twenty-two. In contrast, the next most complex models, the two-dimensional cumulative scale score and threshold models, have two item parameters and two subject parameters, for a total of four. These factors would argue against automatically accepting the multidimensional spatial model as the most valid or useful of those discussed.

Turning to the other models, the two invariant models can be dismissed because of their invariance. It is known that people do not all answer personality inventories identically. There is little point in considering models that cannot at least duplicate this feature. The point of interest for these models is that such limited models do in fact predict so well.

This leaves three models to be considered: Jackson´s threshold model and the two cumulative models using scale scores. There are no significant differences in accuracy between the latter two (p>.05), but they are somewhat more accurate than the former model (p<.0001). As the addition of desirability to content in the two-dimensional model does not improve predictive accuracy and requires more parameters as well, one would tend to prefer the simpler of the two. One would therefore conclude that the simple cumulative model using scale scores is to be preferred over the other models.

It might be noted at this point that references to Jackson's (1968) threshold model are not fully accurate. It should be understood that this model involves two subject parameters, only one of which is used in the prediction form used here. It is entirely possible that a prediction model using both subject parameters of salience and threshold would have a higher level of predictive accuracy then the form which uses only one parameter. This likelihood is supported by the finding reported earlier in this chapter that different methods of calculating salience gave figures which correlated highly, whereas this did not happen for thresholds (Table 4-3). The same situation was found for content models (Tables 4-5 and 4-6), in which salience predicted better than did threshold. It is therefore quite possible that a combination of salience and threshold in a prediction model would prove more accurate than threshold alone.

To this point we have considered the number of model parameters and predictive accuracy as the sole criteria. The latter is a property of the subjects, but item properties are also of interest. Therefore, let us now examine the models in terms of how well they reproduce the distribution of item p-values. These data are summarized in Table 4-13, which presents the first four moments of the distributions of p-values, plus the correlation of the predicted p-values with the actual p-values.

Table 4-13. Comparison of the Distributions of P-values of Predicted Responses

	Prediction Model							
Measure	Actual Distribution	Spatial-Scales	Spatial-Factors	Scale Score	Scale Score + Dsy	Jackson's Threshold	P-values	Judged Desirability
UWO Sample								
Correlation	-	.929	.906	.359	.381	.279	.723	.610
Mean	.510	.511	.514	.469	.475	.556	.486	.571
Variance	.047	.081	.097	.131	.132	.152	.250	.246
Skewness	-.002	.032	-.010	.177	.151	-.134	.057	-.287
Kurtosis	-.480	-1.021	-1.269	-1.504	-1.517	-1.702	-1.997	-1.918
High School Sample								
Correlation	-	.905	.862	.307	.322	.598	.654	.492
Mean	.506	.512	.514	.473	.483	.547	.486	.571
Variance	.037	.070	.084	.130	.130	.116	.251	.246
Skewness	.013	.075	.037	.124	.099	-.105	.057	-.287
Kurtosis	-.072	-.803	-1.071	-1.526	-1.528	-1.613	-1.997	-1.918

A positive skewness is indicated by a value greater than zero, a negative skewness by a negative value. A leptokurtic distribution is indicated by a positive value and a platykurtic distribution by a negative value (Snedecor & Cochran, 1967). The correlations between actual and predicted p-values alone are summarized in Figure 4-6. The results are broadly similar to those already discussed. The multidimensional spatial models have the highest degree of resemblence, but the invariant models have the next highest correlations. However, these models produce a bimodal distribution of p-values which cannot be mistaken under any circumstances for a set of real data. The scale score models and Jackson's threshold model are approximately at the same level of resemblence. In the UWO sample, the two are indistinguishable (differences between correlations, p>.05). In the high school sample, the threshold model does somewhat better than the scale score models (p<.01). Although it may be that Jackson's threshold model does hold across samples better than the scale score models, it is perhaps safer to note that this was an unreplicated result and not place a great deal of weight upon it. This would then not alter our original conclusions.

There is one other attribute of the models which may have a bearing upon the selection of the best model, and that is the internal consistency of the predicted responses. Reliable measurement is highly desirable, and one would prefer not to use a model which predicts unreliable scales. Table 4-14 presents the internal consistencies (coefficient alphas) for the PRF scales for responses of the UWO sample, as predicted by Jackson's threshold model for desirability, the scale score model and Cliff's multidimensional spatial model. Both optimized

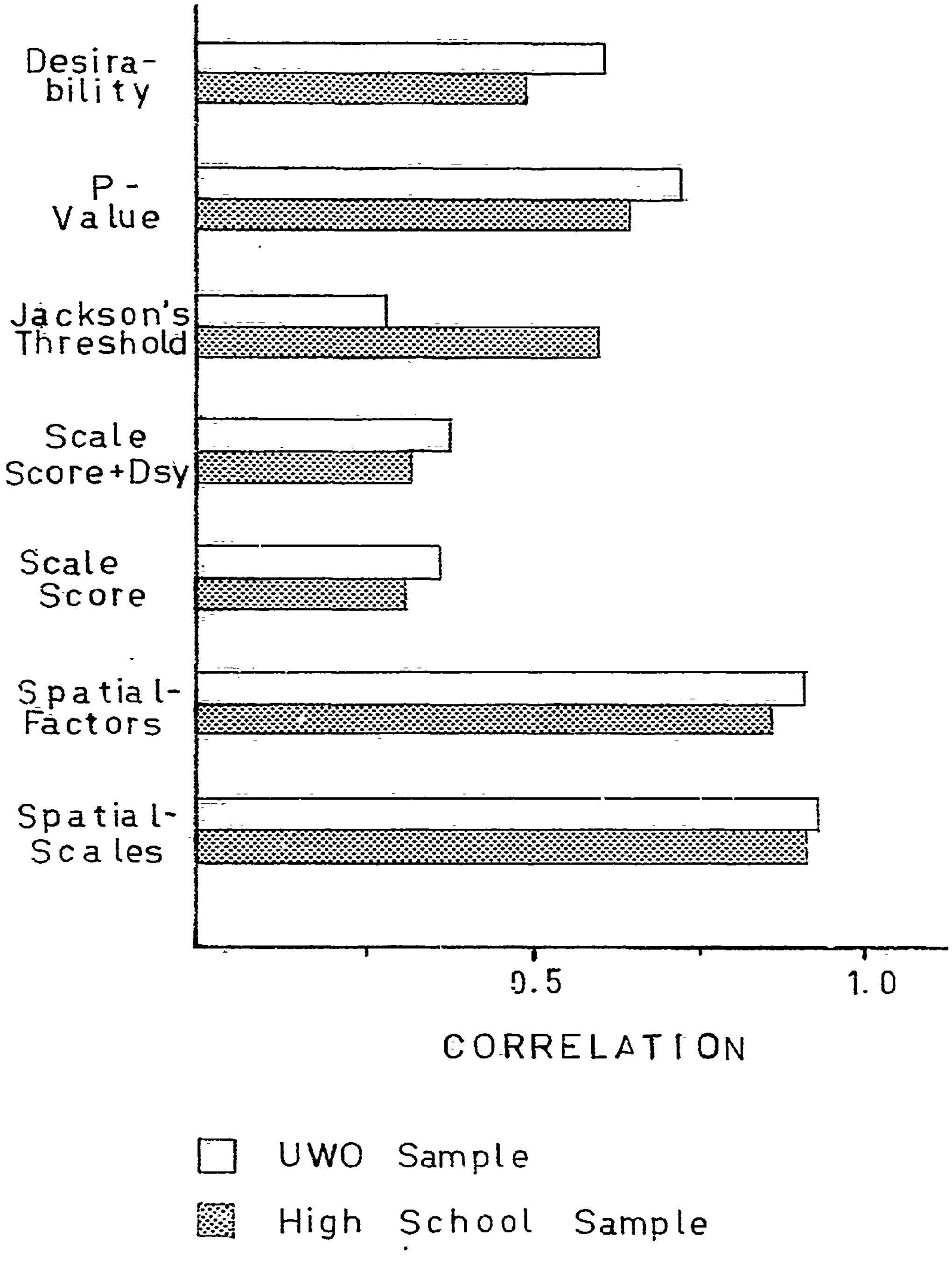


Figure 4-6. Similarity of predicted p-values to actual p-values

Table 4-14. Internal Consistency of PRF-E Scales for Responses Predicted by Three Prediction Models

	Prediction Model			
Scale	Jackson's Threshold[1]	Scale Scores	Spatial, Optimal.	Spatial, Cross-Validated
Abasement	00	86	74	70
Achievement	77	90	86	83
Affiliation	77	90	81	80
Aggression	48	91	82	81
Autonomy	00	87	85	85
Change	66	86	76	76
Cognitive Structure	00	87	85	82
Defendence	00	87	83	82
Dominance	27	92	90	88
Endurance	63	88	81	83
Exhibition	00	92	91	88
Harmavoidance	15	93	92	91
Impulsivity	63	92	91	91
Nurturance	80	87	83	83
Order	02	94	94	93
Play	46	89	81	81
Sentience	59	82	65	67
Social Recognition	00	90	87	85
Succorance	00	91	88	87
Understanding	53	87	80	79
Infrequency	67	34	64	53
Desirability	91	84	78	75
Mean	38	86	83	81

Note: Decimals omitted.

[1] Five scales had coefficient alphas less than zero. These have been changed to 0.00.

and cross-validated predictions were used for the latter. The scale score model has the highest level of internal consistency, followed by the multidimensional spatial models and then the desirability model. It should be noted that the highest internal consistency for the latter model is that for the Desirability scale. The mean internal consistency for the former two models is above the mean scale internal consistency for actual data on the PRF (Jackson, 1974). As neither model incorporates any random processes, as real respondents undoubtedly do, such a result might be expected. The high level of internal consistency of the scale score model brings to mind the finding of Turner and Fiske (1968) that scale homogeneity was correlated with the use of relevant response processes. The scale scores model uses only such a process in making its predictions and produces highly reliable scales.

One final note to be made is that none of the prediction models produced a distribution of p-values that could not be distinguished from that of the real data. All predicted overly high p-values for some items. As all the models were determinate and deliberately avoided the use of random processes, such a result is perhaps to be expected. It also indicates that our understanding of the response process is far from complete. However, the simple models used here do fairly well in reproducing the actual pattern of responses.

Discussion

On the basis of the previous section, we have tentatively concluded that the best prediction model is the cumulative scale score model. It has the great virtue of simplicity. It operates with unidimensional constructs, makes no distributional assumptions, has but one parameter for each item and one for each subject on each construct and can accurately predict 64% of the item responses. It has a demonstrated empirical relationship to the salience of the cumulative threshold model, which in turn has theoretical significance. This takes the model above the level of a purely mechanistic approach. The question now becomes: is this level of prediction adequate or is it unsatisfactory?

In answering this question, we must first decide what level of accuracy can be expected. The naive answer is 100%; all of the items for all of the people. This, however, ignores the inadequacies of both the test and of those responding to its items. The mean internal consistency (coefficient alpha) of the PRF-E for the sample of 214 U. S. college students (Helmes & Jackson, 1977) across scales is 0.72. This is exactly the same as the odd-even reliability reported in the manual (Jackson, 1974). The mean value of Bentler's (1972) theta for these data is 0.87. These figures are reliabilities for the scales. The item reliability is, of course, much lower, as is indicated in the following chapter. The average item inter-correlation on the PRF is below 0.2, a figure more consistent with item reliabilities. Accepting that PRF scales are not strictly unidimensional and so using the latter figure of 0.87 for the reliability of the PRF as a whole, we can determine that approximately 75% of the total scale variance of the PRF is reliable. It will be recalled that the scale variant of the multidimensional spatial model accurately predicted about this many item responses. One could then argue that this is the practical limit of accurate item prediction for the PRF. This level would of course be quite different for a less reliable inventory. This figure may even be generous, as the predictions were made at the item level, whereas the reliability used is based upon scales.

If the multiple regression models reach the limit imposed by the reliability of the test, we can now turn to the respondents and see if there is a similar limit in this case. Turner and Fiske (1968) reported that only 57% of their subjects used response strategies consistent with a cumulative approach to test responding. This figure is somewhat lower than the level of accuracy reached by the cumulative scale score model. As the PRF has better item properties than the tests used by Turner and Fiske, it is likely that this model was predicting close to the limit imposed by the degree of use of irrelevant response strategies.

It should be noted that these two figures for the proportion of predictable responses are not inconsistent or contradictory. Prediction by multiple regression uses all the reliable and consistent variance in the predictors which is linearly related to the criterion. The more structured scale score model uses the proportion of variance relevant to that model. In an imperfect test, there remains variance that is predictable but does not conform to the properties required by a structured model such as the scale score model.

It would be comforting to end with the conclusion that the models developed are predicting the entire amount of predictable variance in the test. What is less than comforting is that this level is so low. If the models developed are indeed valid and individuals use relatively simple processes (such as those inherent in a cumulative homogeneity model) in responding to personality inventories, there is a great deal to be done in uncovering the characteristics of items that promote the use of strategies used by the models.

The accuracy of the various prediction models is a function of the adequacy of the values of the parameters used. One method of improving the predictive accuracy of such models would be to determine what properties of items are associated with predictive accuracy. This knowledge could then be used to develop tests which could be more accurately predicted. As models which produce highly accurate predictions also produce highly reliable scales (Table 4-14), tests developed on the basis of such information might also prove to have other good properties.

CHAPTER FIVE

CORRELATES OF ITEM PREDICTIVE ACCURACY

Having determined which models are best capable of predicting item responses, we shall now examine the items in somewhat more detail. The only item parameters used by the accurate prediction models are scale values for content and for social desirability. Obviously, items have more attributes than these. Presumably, respondents are sensitive to such other attributes. Therefore, by looking at properties of items, we may be able to determine which, if any, are related to the predictability of the items. If items are predictable and we know which item properties are related to predictability, we can perhaps use this knowledge to improve future test instruments.

The previous chapter showed that it is possible to develop prediction models capable of predicting substantially all the reliable variance of a modern personality inventory. It would be desirable to increase further the reliability of such tests, and the proportion of respondents using the response processes relevant to the prediction models. At present we know relatively little about the factors promoting reliability, and even less about those dealing with predictability. Here we can examine factors related to the predictability of test items. This knowledge may be useful in the search for increased reliability of assessment methods.

Item Properties and Predictability

The first step is to select a set of item properties for study. An item can have as many properties as one is capable of devising and reliably differentiating within a collection of items. One could thus examine items in terms of their length, whether self-referent or not, the presence or absence of modifiers, or the number of occurrences of the letter ´q´. One could also examine the more traditional numeric properties, such as item popularity or social desirability scale value. For the sake of convenience, we shall take the latter, more conservative, approach and examine some of the more easily obtained item properties.

The item properties analyzed include the item popularity or p-value, successive intervals scale value for social desirability, the content scale value for the scale on which an item is keyed, the biserial correlation between the item and the total scale score for the scale on which the item was keyed and the direction of keying. This choice of properties is somewhat arbitrary and is based largely on the availability of these values. However, it does include the major statistical properties of items as used in most test construction programs.

The item p-values used were for the U. S. college and high school samples as reported by Helmes et al. (1977). The high school sample used in the prediction studies was part of the latter sample. The social desirability scale values were obtained from the same source. The content scale values for scale content were the basic structure content scores for the U. S. college sample obtained as described in Appendix 1. Item-total biserials were those for the U. S. college sample. This measure was used by Turner and Fiske (1968) as their measure of item homogeneity. For the item content and desirability scale values, negative scale values were reflected, as accuracy is presumed to be related to the extremity of the scale value and not to its polarity. Predictive accuracy was assessed by calculating Cohen's (1960) kappa for the 352 items of PRF-E for the UWO and high school samples used by the prediction models. These values of kappa were then correlated with the above six item properties.

The correlation of predictive accuracy with item properties was done for the three prediction models of most interest. The first such model is the multidimensional spatial model variant using scales. It is of interest because of its overall high degree of predictive accuracy. The scale score model and Jackson's threshold model are included because they combine a reasonably high degree of accuracy with a higher degree of parsimony than has the multidimensional spatial model.

Both the multidimensional spatial and scale score models should demonstrate an association between item predictive accuracy and item content. Scale content was the dominant emphasis in PRF construction, coupled with efforts to minimize desirability (Jackson, 1974). This effort should be reflected in the predictive accuracy statistics. The effect of item reliability is more doubtful. This was also emphasized in PRF construction, but the effect may not appear at the item level, as the emphasis in the PRF is upon the scale constructs.

Jackson's threshold model is not expected to show an association between any item property and item accuracy, except for social desirability. This is the only item property relevant to this model. Even this may show a limited relation, as the PRF is not particularly suited to such a model. Jackson's threshold model was demonstrated to be of high utility in studies by Rogers (1971) and Voyce and Jackson (1977). However, both cases dealt with psychopathological content for normal respondents. This is perhaps the case most prone to elicit socially desirable responding. In dealing with a range of normal content, sharply limited in social desirability, Jackson's model is of limited applicability for the PRF. This was previously demonstrated in the low predictive power of this model in the previous chapter. If operating in a situation for which it was designed, the model would undoubtedly have a higher predictive capability.

The results are reported in Table 5-1. Note that in no case are item properties based on the same sample as that on which the item predictions are made. The most notable aspect of Table 5-1 is the generally low magnitude of the correlations. Those which are significantly greater than zero account for at most approximately 30% of the variance.

If we turn to individual prediction models, we find that there is a strong relationship between accuracy of prediction and both item content and item-total biserial in both prediction samples for the multidimensional spatial model. The low correlations of the true-false scoring key with accuracy for the multidimensional spatial models would

Table 5-1. Relation of Prediction Accuracy to Item Properties

	Model					
Property	Multidimensional Spatial		Scale Score		Jackson's Threshold	
	UWO	High School	UWO	High School	UWO	High School
Item Content Scale Value	33	35	13	14	-09	-13
True-false Scoring Key	08	13	02	-01	07	-11
Desirability Scale Value	06	-23	23	-02	08	-20
U. S. College P-value	01	-05	-04	-05	-14	-14
High School P-value	01	-04	-00	01	-07	-05
Item-total Biserial	23	54	04	37	-09	06

Note: Decimals omitted.

tend to argue that the polarity or direction of item content is a component of the predictability of the items, but not a major one. Obviously, item content itself is a more substantial component than simple directionality. It is interesting to note the negative correlation of desirability for the high school sample, implying a higher predictive accuracy for items with a low desirability scale value. Although this finding is not replicated in the university sample, it is consistent with the importance of content, as high desirability and high content saturation are not usually compatible in the same item in a well-constructed test. Similarly, item content is presumably behind the high correlations for the biserial, an index of the item´s contribution to the scale construct.

For the scale score model, there is a consistent low correlation of item content with the predictive accuracy. This was predicted and recalls Kuncel´s (1973) findings in the test-retest situation. Extreme item content is associated with large subject-item distances. This result therefore extends Kuncel´s findings to the predictive case. In the UWO sample, desirability scale values are also correlated with accuracy. The correlation between desirability scale values and content scale values for items is only 0.13. Therefore, a correlation of content and desirability is unlikely as an explanation of this outcome, although it does remain as a possibility. If such a correlation were to be the cause, one would expect that it would appear in both samples, and it does not. Its appearance remains puzzling. Similarly, the occurrence of a substantial correlation between item biserials and predictive accuracy is found only in the high school sample. One might expect such a correlation on the grounds that the biserial, as an index of item internal consistency, should be related to accuracy. Turner and Fiske (1968) found that the biserial was correlated with use of response processes relevant to models such as the scale score model. A highly homogeneous and internally consistent scale should also be easier to predict accurately than a less internally consistent scale. This idea is also supported by the finding in the previous chapter that the scale score and multidimensional spatial models predicted responses which formed scales with high internal consistencies. The emphasis placed on the biserial during construction of the PRF (Jackson, 1974) may contribute to the lack of correlation in college students, through a reduction of the range of the biserials. As this population was used in its development, this is possible, but this would appear insufficient to account for the lack of replication. Because of the failure of replication of a theoretically expected result, we shall examine this factor of internal consistency again later.

For Jackson's threshold model for desirability, the results are also rather odd. The only significant correlation common to both samples is a negative correlation of accuracy with p-values. This indicates a tendency for higher predictive accuracy to be associated with items with a low popularity. However, as half the items are false-keyed, reversing these items to present the p-values as the proportion of responses in the keyed direction would give a rather different picture.

One important point to be considered in discussing accuracy of prediction is that the predictive accuracy at the item level is rather low. This is illustrated by Table 5-2, which gives the mean level of predictive accuracy for the three preferred models. The values of kappa are highest (about 0.4) for the multidimensional spatial model. This is entirely understandable, as this model requires knowledge of the actual item response in order to make a prediction. Values of kappa for the scale score model are moderate (about 0.22), with a notable degree of variability. Values here ranged from a low of 0.10 for Sentience to a high of 0.40 for Order in the UWO sample and from 0.13 (Change) to 0.36 (Orer) in the high school sample. Values of kappa for Jackson's threshold model are all quite low, indicative of the lower predictive accuracy of this model as compared to the other two models. In addition, it should be pointed out that the emphasis of Jackson's threshold model is more upon the respondent than the items, whereas the other models tend to concentrate more upon the items.

Item Reliability and Predictability

A major point to be kept in mind in discussing and interpreting item properties is that one major property of individual items is their unreliability. This is not reflected in the item-total biserials. Table 5-3 contains three different estimates of item reliability of PRF-E items, obtained for three different samples. Each estimate was obtained by scaling down the reliability coefficient for the scale to the level of the single item by use of the Spearman-Brown formula. The odd-even reliability reported by Jackson (1974, Table 21) agrees well

Table 5-2. Mean Values of Kappa for PRF-E Scales for Three Best Prediction Models

	Model					
	Multidimensional Spatial		Scale Score		Jackson's Threshold	
Scale	UWO	High School	UWO	High School	UWO	High School
Abasement	29	27	15	15	02	07
Achievement	41	41	24	21	05	11
Affiliation	43	42	22	21	04	09
Aggression	42	37	29	20	02	07
Autonomy	47	39	19	18	03	07
Change	31	25	15	13	10	09
Cognitive Structure	38	38	18	20	02	08
Defendence	34	33	15	16	03	10
Dominance	49	47	28	24	04	08
Endurance	31	33	21	23	06	07
Exhibition	48	49	29	29	01	08
Harmavoidance	48	41	37	26	02	03
Impulsivity	49	45	30	25	04	08
Nurturance	31	39	24	26	08	06
Order	58	52	40	36	05	08
Play	38	37	20	15	01	09
Sentience	22	27	10	17	05	07
Social Recognition	48	32	29	19	05	09
Succorance	47	43	24	25	04	11
Understanding	36	41	19	28	05	06
Desirability	32	37	13	17	04	08
Mean	38	37	22	21	04	08

Note: decimals omitted.

Table 5-3. Estimates of Item Internal Consistency

	Estimates			
Scale	Odd-even	Alpha	Theta	KR-20 PRF-AA
Abasement	13	10	23	10
Achievement	08	18	36	14
Affiliation	38	18	34	17
Aggression	10	20	39	17
Autonomy	11	14	31	12
Change	10	09	21	07
Cognitive Structure	12	14	29	14
Defendence	11	12	28	12
Dominance	11	28	49	26
Endurance	16	16	31	18
Exhibition	26	25	54	17
Harmavoidance	39	29	49	23
Impulsivity	26	23	39	11
Nurturance	10	14	28	14
Order	34	34	60	26
Play	06	16	31	12
Sentience	13	06	17	12
Social Recognition	15	20	36	20
Succorance	15	21	39	18
Understanding	17	12	34	10
Infrequency	13	01	06	03
Desirability	12	08	25	09
Mean	16	17	33	15

Note: Decimals omitted.

with coefficient alpha calculated on the UWO sample, and with the KR-20 (coefficient alpha) reported in the manual (Jackson, 1974, Table 8) for the longer scales of PRF-AA. Bentler's (1972) theta is substantially higher than the other, unidimensional estimates of internal consistency, as would be expected. However, even this is not very high, reaching a mean level of 0.33, with the highest level of internal consistency (.60) again being for the Order scale.

It will be recalled that Turner and Fiske (1968) noted that item homogeneity tended to be associated with the use of relevant response categories. Relevant response categories are those congruent with the use of cumulative prediction models such as that using scale scores. The scale score model, based upon such a process, predicted highly internally consistent scales (Table 4-14). Turner and Fiske (1968) were dealing with response stability in test-retest situations. In addition, homogeneity and internal consistency are not synonymous, but these results provide sufficient incentive to examine the relationship between predictive accuracy and internal consistency. This is particularly true given the suggestive results for item-total biserials in the previous section. Therefore, the mean values for PRF-E scales (Table 5-2) were correlated with the estimates of item internal consistency (Table 5-3). These results are reported in Table 5-4.

It will be noted that there is a substantial degree of association between the predictive accuracy and scale internal consistency for both multidimensional spatial and scale score models. This degree of association is completely lacking for the threshold model. Such a

Table 5-4. Correlations between Predictive Accuracy for Scales and Item Internal Consistency

		Measure of Internal Consistency			
Model	Sample	Odd-even	Alpha	Theta	KR-20 (Form AA)
Multidimensional Spatial	UWO	45	88	75	75
	High School	40	80	73	68
Scale Score	UWO	60	96	84	83
	High School	56	80	73	72
Jackson's Threshold	UWO	-12	-04	-06	03
	High School	-23	25	11	23

Note: Decimals omitted.

pattern of correlations of accuracy and internal consistency is not surprising. Multiple regression will predict more accurately the greater the amount of non-error predictive variance there is in the data. An increase in internal consistency which maintains the same relationship to responses will automatically reduce the proportion of random or non-predictable variance. Hence, the more reliable scales are predicted more accurately by the multidimensional spatial model, which is based upon multiple regression. The high level of correlation for the scale score model is of special interest. This finding extends the previous work of Turner and Fiske (1968) who showed a similar relation between test-retest stability and the use of relevant response processes. The scale score model is based upon such relevant response processes and shows a very high degree of association with predictive accuracy for PRF-E scales. This supports the general idea of producing scales with as high a degree of internal consistency as is possible without producing scales of trivial breadth. It would be predicted that scales and items with a high degree of internal consistency would show both a high degree of test-retest stability and a low level of use of irrelevant response processes. A lack of association between predictive accuracy and scale internal consistency for Jackson´s threshold model is not surprising. Scale or item reliability is irrelevant to such a model and so no relation is expected. The low level of predictive accuracy and limited applicability in this context already noted for this model are also relevant to this point.

In addition to scale internal consistency, there are other item properties worthy of investigation in this regard. These are less easily dealt with than those properties which have already been discussed, but are perhaps more important, as they are frequently more apparent to the respondent. Such properties might include whether the item is self-referent or not. Rogers (1977) showed that such items require more time in making a response. It might thus be expected that such items are more subject to error and less predictable. Loehlin's (1961; 1967) work has shown the undesirability of using adjectives for assessment. However, the advantages of short and simple items are well-known. Perhaps the use of only words of unambiguous meaning in short items would prove advantageous. This is only one aspect of item length that is worthy of exploration. Item ambiguity itself is another property deserving of more study, as is the role of qualifying terms. These aspects of items are among those which might be termed the grammatical and lexical aspects which most warrant further research.

In summary, the most important item property in high predictive accuracy is the internal consistency of the relevant scale. The other major factor is item content saturation, particularly as this is relevant to high internal consistency. This factor is probably an indirect one. A closer relationship would probably be found in a direct measure of subject-item distance for the relevant construct. Low social desirability is also intermixed in this, as high content saturation and social desirability should not be found in the same item for purposes of purity of assessment. These aspects of items should at least provide a relatively concrete background for work on the improvement of personality assessment methods.

CHAPTER SIX

SUMMARY AND CONCLUSIONS

To recapitulate briefly the findings of this work, various prediction models of responses to the Personality Research Form (PRF-E) were evaluated. This process requires the use of scale values for all 352 items on the 20 content scales of the PRF. This in turn required the use of scale values from the recently-developed basic structure content scaling procedure (Jackson & Helmes, 1976). Therefore, it was first demonstrated that there is a high degree of agreement between scale values based upon item responses as derived from the basic structure scaling procedure and scale values derived from traditional scaling methods based upon judgements of item content. This agreement allows the use of the basic structure content scores in the various prediction models.

The prediction models were evaluated on two samples of respondents, one of 92 university students and the other of 301 high school students. Five invariant models were evaluated first to determine which attributes of items were most effective in predicting item responses. Results showed that models based upon item p-values and social desirability ratings were the most accurate predictors. Invariant models based upon content did not predict item responses well. Two threshold models were described and evaluated, each in several variants. Unlike the invariant models, these models allow for individual differences in two parameters, the salience of the particular dimension for that individual, and the individual's threshold for selecting a true response. One of the threshold models, that first described by Jackson (1968), is based upon social desirability as the dimension upon which items differ and are scaled. The other model uses the content of the items as the latent dimensions. In the latter case, thresholds are not useful in predicting item responses above the chance level. Thresholds for the social desirability dimension correctly predict approximately 60% of the item responses. For both these models, analysis of the parameters of the models for the individual respondents showed a strong correlation between the scale score on the relevant scale and the salience of the dimension for the individual. In addition, the salience and threshold were largely independent of one another. The correlation of scale scores and saliences for content led to the development of another model based upon content which used the scale score in a cumulative homogenity model. This model accurately predicted 65% of the item responses. A multidimensional spatial model using multiple regression procedures accurately predicted over 70% of the item responses.

Further evaluation of the models was restricted to those seven model variants which accurately predicted at least 60% of the item responses in the two samples. Of these, the two invariant models, those based upon social desirability and p-values, were eliminated from further consideration because of their inability to account for individual differences. A variant of the scale score model which predicted on the basis of both content and social desirability was rejected because it did not predict significantly more accurately than the variant based upon content alone and required more parameters. This left three models, one of which had two variants. In order of decreasing accuracy of prediction, these were: multidimensional spatial using 22 scale content dimensions, multidimensional spatial using 6 factor content dimensions, scale scores and Jackson's threshold model using social desirability.

The seven most accurate models were also evaluated in terms of the distributions of p-values of the predicted items. The distributions produced by all prediction models could be distinguished from that of the actual responses. The most similar distributions were produced by the multidimensional spatial model, followed by the invariant p-value model and the scale score models. The considerations of accuracy of prediction and similarity of distributions of p-values generally led to the same conclusions as to the three best models. Cross-validation of the predictions of the multidimensional spatial model showed shrinkage of about 5% in the number of items accurately predicted. This figure allowed this model to retain its position as the most accurate prediction model. Examination of the reliability of the predicted responses for the three best models showed that the scale score model had the highest level of internal consistency, with a mean across scales of 0.86. The multidimensional spatial models were slightly less reliable, and the desirability threshold model had the poorest level of internal consistency.

The final decision as to the best model therefore took into account the factors of predictive accuracy, similarity of distribution of predicted item p-values to che actual distribution, reliability of the predicted scales and parsimony of the models in terms of the number of item parameters required to make a prediction. On these grounds, it was concluded that the scale score model was the best of those evaluated.

Presumably, the accurate prediction models have some relationship to processes actually used by respondents. Research by Kuncel (1973; 1977) has shown that responses which are stable over time do in fact involve respondent processes similar to those used by the scale score model. Therefore, it may be helpful to our understanding of item responding and to the development of better measurement instruments to know what properties of items are related to highly accurate predictions. This analysis was performed by correlating various numerical attributes of PRF-E items with accuracy statistics for the same items for the three best prediction models. For both multidimensional spatial and scale score models, predictive accuracy was associated with item content scale values and with indices of item reliability.

These results would argue that methods of scale development aimed at improving scale reliability and the intensity of item content would be most likely to lead to improved assessment of personality. Knowledge of attributes of items related to these factors is limited at this time. It is suggested that detailed analysis of the grammatical and lexical structure of items would be one starting point.

APPENDIX 1. BASIC STRUCTURE CONTENT SCALING

Note: This material has also been reproduced as U. W. O. Psychology Department Research Bulletin Number 442, by D. N. Jackson & E. Helmes.

Abstract

A basic structure approach is proposed for obtaining scale values for attitude, achievement, or personality items on a number of dimensions from response data. Unlike multidimensional scaling methods, the scaling of large sets of stimuli is practical, and judgments of items are obviated. In attitude and personality item scaling, the technique permits the unconfounding of scale values due to response bias and to content. It also permits the partitioning of item indices of popularity or difficulty among a number of relevant dimensions, a property of possible relevance to tailored testing.

Basic Structure Content Scaling[1]

Ever since Thurstone's (1929) classic work in attitude measurement, there has been a recurring need for item scale values, representing the locations of items on underlying content dimensions. Such information is useful for a number of practical and theoretical applications: for example, to construct attitude, personality, and ability measures so that different points along content dimensions are appraised; to identify the polarity of items or other stimuli with respect to one or more dimensions; and as a basis for investigating processes underlying responses. Although both judgment and response methods have been used for scaling stimuli (Torgerson, 1958), unidimensional and multidimensional judgment scaling procedures have been much more widely applied than have response methods, and have frequently been found to yield results of high reliability. However, there are reasons for sometimes preferring methods based on responses for certain purposes: (a) some judgment methods are impractical for large data sets; (b) there are economies with response methods for psychological scale development in that they obviate the initial collection and analysis of judgmental data; and (c) if the focus is upon identifying scale values associated with processes underlying responses to the stimuli, more accurate and more relevant scaling may be obtained from responses. For example, there is evidence (Boyd & Jackson, 1966) that multivariate response methods applied to a set of attitude items yield a markedly different structure from that provided by multidimensional scaling of the same items, with response bias substantially represented in the former.

Here we outline a scaling procedure based upon item responses, which frequently can be obtained more quickly and easily than equivalent judgments. This procedure is intended for the scaling of the items of large multiscale personality and attitude questionnaires, although it is potentially useful for other types of assessment devices as well. This emphasis differs from the majority of multidimensional scaling studies, in which the emphasis is upon the identification of the prominent dimensions of the space of perceived similarity, rather than the scaling of objects.

The procedure described here employs a data matrix generated from responses to a set of items, rather than the more familiar methods of scaling items or objects from judgments. There are similarities, however, between certain proposed models for scaling objects from preference judgments (Bennett, 1956; Coombs & Kao, 1960; Slater, 1960; Tucker, 1955) or dominance judgments (Carroll, 1972) and the method for scaling response data proposed here. There are also similarities to previous work by Hill (1974). Essentially, the current method is analogous to these judgment methods in that item responses are interpreted as a kind of judgment regarding the presence or absence of the attribute represented by the item for a particular respondent. This is somewhat different from the way in which response data are traditionally analyzed by some form or combination of correlational, factor, or component analysis (Shepard, 1972). However, both judgments and component analysis responses may yield similar information from a respondent. For example, to determine the dimensions of preference for flavors of ice cream, one could present an individual with all possible pairs of flavors of ice cream and ask for a judgment of which of each pair was preferred. Alternatively, c.ie could ask for a response as to whether or not the individual liked each flavor of ice cream. The problem lies in determining scale values for ice cream flavors for the latter type of data.

The problem of representing different response patterns among respondents may be approached by considering each respondent as a vector in a multidimensional space (cf. Jackson & Messick, 1963). Each respondent is assumed to respond to each item with respect to a one dimensional attribute, namely, how true or characteristic the item is of the person. Respondents may differ among themselves in their pattern or ordering of true and false responses. Rather than seeking similarities among items or tests, the similarities are sought among respondents. Assume a set of personality items was drawn from scales for aggression, dominance, and exhibition and that these items varied in desirability. Persons responding to these items might present different patterns of true and false responses, depending on their perception of these attributes in themselves and upon the weight accorded desirability in responding. Different response patterns permit the isolation of different types or clusters of individuals with regard to the personality traits assessed. A type or cluster is represented by a vector extending in some particular direction within the multidimensional array of item responses such that the vector best represents that type. Item projections on each such vector represent scale values for each identified type. Thus item responses can be used to scale those same items. A similar rationale underlies vector models of preference judgment or of factor analysis of ratings of traits with respect to a single multidimensional criterion, such as desirability (Messick & Jackson, 1972). Unlike classical unidimensional scaling models, such a vector model does not require that the relevant dimension(s) be specified in advance. Unlike some multidimensional scaling models, judgments of inter-stimulus distances are not required. It is well-known that judgment data can be scaled using a vector model, just as response data can be scaled using a point model. The advantages of the basic structure approach lie in its use of responses. Normally, this means less demand upon the subject's time and attention, as fewer responses are required than judgments. In addition, analysis of responses to items may provide different information than that provided by judgments.

Most metric or nonmetric multidimensional scaling procedures begin with pairwise judgments of similarity or other measure of distance between objects. In dealing with psychophysical or perceptual data, such a concept is directly applicable to the stimuli. However, with complex stimuli, it is apparent that the subjects' judgments of similarity can add cognitive dimensions to those that provided the physical basis for constructing the stimuli (Torgerson, 1965). This provides a concrete illustration that the processes involved in judging a stimulus may be somewhat different from those involved in responding to the same stimulus. Boyd and Jackson (1966) provide another example in the scaling of attitude statements. In this case, both factor analysis and multidimensional scaling of the same attitude statements led to the same dimensions, with the addition of an acquiescence dimension when responses were used. The difference between judgments and response processes can be thought of as being the difference between the distance between an item and an ideal concept and the distance between the item and the person. Depending upon the concept and the person, these two distances may or may not be the same. In the previous example, if asked to judge his preference for ice cream, an individual might apply dimensions of color, sweetness, and presence of fruit. If asked whether or not he liked the same flavors, a novelty factor might be introduced.

Nevertheless, one would normally expect much the same dimensions to appear from both types of data. The appearance of the same dimensions from both sources of data, judgments and responses, has been taken as being a necessary condition for accepting the generalizability of the dimensions (Stewart, 1974).

The rationale for the scaling procedure is developed formally as follows.

Definition of Notation

In general, the notation follows that of Horst (1965).

X is an entity (N) by attribute (n) matrix.

Z is X, column standardized, with mean 0 and unit variance, scaled by $1/N$ so that $R = Z'Z$ Items are entities, persons are attributes of items in both Z and X. This is different from conventional practice, in which persons are entities and items or scales are interpreted as representing attributes of persons.

P and Q are the left and right basic orthonormals, respectively, of Z.

Δ is the basic diagonal of Z.

T is a k by k orthonormal transformation matrix, where $T'T = TT' = I$, and where k is the number of dimensions.

Y is an N by k matrix of item scale values with respect to k dimensions.

Description of the Method

The procedure is a form of conventional components analysis. It is most parsimoniously described in terms of a singular value or Eckart-Young decomposition, but this is not necessary. In this case, we are interested in the projections of items upon axes of the person space. In traditional usage, this is equivalent to obtaining component scores for items, with component loadings associated with individuals.

Using a singular value decomposition routine (Businger & Golub, 1969),

$$Z = P\Delta Q' \quad . \tag{1}$$

Perform an orthogonal procrustes transformation of k columns of P (e.g., Schönemann, 1966; Ten Berge, 1977).

(2) $$Y = PT$$

where Y is the best least-squares approximation to the hypothesis or target matrix. Generally, this matrix will consist of the scoring key for the test or the best estimation of the allocation of the items to a priori dimensions. Alternatively, an oblique rotation may be preferred, or some other form of analytical or graphical rotation may be used. If one so wishes, the scale values can be rescaled to unit variance following rotation.

If a singular value decomposition routine is not available, the same solution may be obtained by conventional procedures (Kaiser, 1962; Horst, 1965).

(3) $$R = Z'Z$$
(4) $$R = Q\Delta^2Q'$$
(5) $$A = Q\Delta$$
(6) $$B = AT$$
(7) $$Y = ZB(B'B)^{-1}$$

Again it should be noted that this Y represents component scores for items which we interpret as content scale values for items. Y is not the matrix of component scores for individuals.

Alternatively, the rotation may be deferred and carried out directly upon the unrotated item content scores:

(8) $$P = ZA\,(A'\,A)^{-1} \quad ,$$

which may be entered into equation (2) to obtain item scale values.

Illustration

To illustrate the scaling procedure, a set of 16 items was selected from four scales of the Jackson Personality Inventory (JPI) (Jackson, 1976). The responses of 82 college students (provided by P. M. Bentler) were used to determine the scale values.

The first step, after arranging the items into groups by scales, was to standardize the binary response matrix to remove respondent means and yield unit variance for each respondent. In accordance with the equations, we shall assume an item by respondent data matrix, and that standardization is by columns.

Next the standardized data matrix is decomposed according to equation (1) and the four largest singular vectors of the left-hand basic orthonormal matrix, P , are retained. As there are four scales involved, four vectors are retained as the expected number of dimensions.

In other circumstances, a different criterion for the number of dimensions might have been employed.

In this case, scale values were desired for the four dimensions corresponding to the scales. Therefore, the hypothesis matrix for the targetted rotation consisted of +1 for an item keyed true on a scale, -1 for an item keyed false on a scale and 0 for an item not keyed on a scale. There was thus only one non-zero entry in any given row of the hypothesis matrix, and that non-zero entry denoted the scale on which the item was keyed. The best least squares approximation to this hypothesis matrix was obtained through Schöneman's (1966) orthogonal Procrustes procedure. The resulting rotated matrix was then rescaled to unit variance by columns, giving the final scale values, reported in Table 1.

Insert Table 1 about here

Examination of the scale values reveals that, in general, extreme scale values are associated with items on their keyed scale and that the scale values invariably reflect the direction of keying. Exceptions to this are congruent with item content. For example, item 100 has a rather extreme scale value for Self Esteem, but is keyed on Conformity. Item content, which deals with being uncomfortable if dressed differently in a social setting, is congruent with low Self Esteem as defined by the JPI. Other conformity items, not highly related to social settings, do not have scale values as extreme on the self esteem dimension.

One point of concern in the use of this method is the determination of the number of dimensions to be retained. One possibility would be the use of the number of a priori scales in the inventory, or secondarily, the number of factors in the test, as determined by separate factor analytic studies. The use of the number of scales has conceptual advantages in that the relationship of an item to the scales of well-constructed tests or questionnaires is well defined. If a test displays convergent and discriminant validity and the suppression of irrelevant sources of variance, then each item should be related most strongly to its own scale.

For personality and attitude items, high content saturation on their keyed scale, together with a reduction in ambiguity of wording in items, are also desirable item properties in well-constructed questionnaires. These two properties act to produce extreme scale values for items by encouraging that individuals possessing similar levels of the trait will respond similarly to the same item. Analogously, ability and achievement items having univocal properties in the sense that they assess skills or knowledge relevant to a single factor are considered desirable in multifactor test batteries. Ideally, such items will differentiate persons possessing particular levels of ability or achievement. It would thus follow that the relative difficulty of items univocally associated with a particular factor would be associated with their scale values for dimensions corresponding to that factor.

We suggested earlier that the hypothesis matrix might be based on the original item keying. But in exploratory work, such a key may not exist. The use of factors either at the scale or item level provides an alternative source for the construction of a hypothesis matrix. This will normally involve the retention of substantially fewer components, which account for a smaller amount of total variance, but which may have higher utility for actual applications of the content scores. A conservative recommendation would be that the determination of the number of factors be based upon analysis of a different sample than that upon which content scores are to be derived.

Item factor analyses provide another source for the hypothesis matrix. A prime example is the MMPI, from which several investigators have constructed scales not corresponding to the MMPI clinical scales. In cases such as these, the hypothesis matrix could be based upon the keying for the constructed scales, or hypotheses about item clusters, rather than the original scales.

The question now arises as to the conditions under which this technique will yield useful scale values. Ideally, the use of this technique assumes that the scale meets certain minimum standards of internal consistency and of content homogeneity. In addition, scales should meet the same criteria as those required of modern construct-oriented approaches to personality scale construction (Wiggins, 1973). Ideally (a) there should be a strong and demonstrable substantive and empirical relationship of an item to its own scale and to no other; (b) the scales should possess a degree of convergent and discriminant validity; (c) irrelevant sources of variance, whether due to irrelevant content or method variance, should be suppressed as much as possible; (d) the items should have a moderate frequency of endorsement.

This final point is of particular importance for the scaling procedure. The presence of items with extreme endorsement proportions in an analysis of this nature may be unstable and lead to interpretative difficulties through the introduction of components associated with differences in p-values. This caution is tempered somewhat by the report by McDonald and Ahlawat (1974) that such components or factors tend to appear only with extreme p-values and non-linear item trace lines. Nevertheless, items with extreme p-values should be avoided if at all possible.

Although originally intended for use with personality and attitude inventories, where the content domain is clearly multidimensional, this method is potentially useful in cognitive domains as well. For example, a problem in trigonometry might involve ability factors for reading, vocabulary, spatial visualization, general reasoning, and number, as well as others. Through the inclusion of appropriate subtests in a battery, the contribution of each of these components of ability to item difficulty could be assessed through examination of the appropriate content scale values. This in turn would imply that different item characteristic curves might be generated for a single item on each of the relevant ability dimensions.

At this point we should note a difference in interpretation between the scale values for items as obtained here, and the item factor or component loadings obtainable from an item factor analysis. An item loading can be interpreted as representing the correlation between a set of responses and component scores. A content scale value defines the projection of the item upon the latent personality or ability dimension. These are logically distinct, in that they are geometrically separate, and also mathematically different in that they involve different combinations of the triple product resulting from the decomposition of the data matrix. In practice, component scores and loadings do not in general yield the same ordering of items on dimensions. This is true even though the two sets of values may be based upon the same original data matrix, subject to different scalings. This does not deny that the two matrices are related in that, in our equations, they represent different scalings of the left-hand orthonormals.

Partial p values

In a perfect Guttman scale, scale values are a function of item popularities (Green, 1956). But even with errorless data, if item responses are attributable to more than one dimension, unidimensional p values may not be interpreted accurately as representing a single underlying content dimension.

A given *a priori* scale or test may contain a number of sets of items for which the item p values in each set are attributable to different content or process dimensions, or, alternatively, item p values which are complex, in that their endorsements or difficulties are each attributable to a number of dimensions in different proportions. Many investigators and test constructors implicitly make the convenient assumption of unidimensionality, but in certain instances, the attribution of p-value variance to distinct processes may be of some critical importance. For example, it is well known that the probability of endorsement of a personality item is a function of its desirability. Efforts to identify popularities attributable to their putative content would at least require separate identification of that portion due to desirability. In tailored testing, item difficulties are employed as a basis for different item sequences. But if the test is designed to assess some unitary aptitude or achievement domain, it might be appropriate to employ item difficulty indices which are not due to the effects of extraneous sources of difficulty.

The scale values described in this paper can be conceived of as standarized item p values. For a unidimensional, errorless test, the two would be highly correlated. For a multidimensional test, the standardized p value could be reproduced by the sum of the item scale values on retained dimensions plus the residual components. Of course, one does not normally standardize item popularities or difficulties, since it is more convenient and relevant to retain the metric of the unstandardized proportion of endorsements. If item p values are required for separate dimensions, all of the formulas for item scale values apply, except that the unstandardized entity by attribute matrix, X , should be substituted for Z in equation (1). In this case, the elements of Y in (2) and (7) would represent "partial p values." The sum of the elements in a given row of Y plus the residual components would be equal to the proportion of the N individuals endorsing that item or attribute. These elements would represent the popularities or difficulties of different items apportioned among the processes or content represented by the separate dimensions.

Large Data Matrices

The method is intended for large multiscale inventories in cases where judgment methods require an excessive number of judgments. However, with a very large number of respondents one can arrive at a data matrix too large for central storage in even a relatively large computer. In this case, the attribute by attribute (respondent by respondent) matrix will almost certainly exceed computer capacity, assuming more respondents than items. Alternative procedures do exist. For example, the decomposition may take place on the entity by entity covariance matrix and the equivalent results obtained by well-known transformations (Horst, 1965, p. 326; Messick & Jackson, 1972; Tucker & Messick, 1963). But even this may be too large to be practical for say, large multiscale personality inventories, for example, our example below of the 352 items of form E of the Personality Research Form (PRF-E), as administered to 969 respondents. An alternative is to extend the item matrix into a components space defined by scales and persons.

Based upon scale keys, or a priori clustering of item content, construct a data matrix X_S of L subscales and n persons. There should be at least three items in each subscale and three subscales per scale. For example, each of the 22 scales of the PRF-E could be divided into four subscales of four items each. This would yield an 88 x n data matrix.

Standardize this matrix, X_S , by columns, yielding Z_S . Note again that this removes the person mean, not the item mean. Proceed as described above to find a set of content scores for the subscales:

(9) $$Y_s = P_s T_s$$

In addition, form the correlation matrix of items and subscales, R_{ys} ,

(10) $$R_{ys} = 1/n Z_y Z_s'$$, where Z_y is the same as Z in the previous section.

In a manner analogous to Dwyer extension procedures (Dwyer, 1937; Tucker, 1971), proceed to obtain extension scale values for items on the scale constructs. Note however that this procedure involves the use of the left basic orthonormals rather than the more common right basic orthonormals,

(11) $$Y_E = R_{ys} Y_s(Y_s'Y_s)^{-1}$$

Thus item scale values may be obtained on a large item x subject matrix without the direct decomposition of this matrix. With relatively homogeneous scales, these values of Y_E will be similar, but not in general identical to those derived from equation (2). There are two reasons for this. The first is due to the possible differences in the manner in which item and scale component analyses identify components. Only if items on a scale form a perfect scale will the two be identical. The second lies in the differences in rotation. That is, the transformation matrices T and T_S do not rotate axes to precisely the same orientation.

Example of Analysis of Large Data Matrix

Form E of the PRF (Jackson, 1974) is composed of 352 items comprising 22 scales. The sample used to illustrate the scaling of this large inventory consisted of 214 college students drawn from U.S. colleges as described in Helmes and Jackson (1977). These data were first scaled directly by means of equation (2). The hypothesis matrix consisted of the True-False scoring key with +1 for true-keyed items, -1 for false-keyed items and 0 for items not keyed on a scale.

Complete results are not presented here for reasons of space,[2] but Tables 2 and 3 give the scale values for items of the Achievement and Dominance scales. The values in Tables 2 and 3 have been taken from a matrix of content scores which have been rescaled to unit variance. As can be seen, the scale values are quite in accordance with item content and with direction of keying, as expected. PRF-E items were written to reflect high content saturation for the scale on which they are keyed (Jackson, 1974). Therefore, the scale values for items keyed on a scale should be substantially more extreme than for items not keyed on that scale. This is in fact the case, as seen in Table 4. The procedure normally assigns negative scale values to false-keyed items. These have been reflected in Table 4.

Insert Tables 2, 3, & 4 about here

Item scale values were also obtained using the extension procedure in equation 11 on the same set of subjects in order to evaluate the degree of similarity of the two sets of content scores. Four subscales were constructed for each PRF-E scale, each containing either four true-keyed items or four false-keyed items. This led to the decomposition of an 88 x 88 cross-product matrix, rather than a 352 x 214 data matrix. The reduction in the amount of computer central storage required is appreciable. The hypothesis matrix (consisting of +1, -1 and 0 as before) was then 88 x 22 rather than 352 x 22 as was the case when scaling items directly.

Simple linear correlations were calculated between the two sets of content scores for items keyed on a scale. Due to the high content saturation of these items, they would be most salient in the scaling procedure and have the most stable content scores. These correlations range from .73 to .99 with a median of .96.

This method of scaling provides a flexible means of scaling very large sets of stimuli such as are typically found in modern multiscale personality inventories. Even with large modern computers, it very often may be the only method of obtaining multidimensional scale values, given the difficulties of obtaining valid judgments from subjects with a large number of stimuli.

References

Bennett, J. F. Determination of the number of independent parameters of a score matrix from the examination of rank orders. Psychometrika, 1956, 21, 383-393.

Boyd, J. E., & Jackson, D. N. An empirical evaluation of judgment and response me^hods in multivariate attitude scaling. American Psychologist, 1966, 21, 718 (abstract).

Businger, P. A., & Golub, G. H. Algorithm 358. Singular value decomposition of a complex matrix. Communications of the Association for Computing Machinery, 1969, 12, 564-565.

Carroll, J. D. Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), Multidimensional scaling: Theory and applications in the behavioral sciences. Vol. 1. New York: Seminar Press, 1972.

Coombs, C. H., & Kao, R. C. On a connection between factor analysis and multidimensional unfolding. Psychometrika, 1960, 25, 219-231.

Dwyer, P. S. The determination of the factor loadings of a given test from the known factor loadings of other tests. Psychometrika, 1937, 2, 173-178.

Green, B. F. A method of scalogram analysis using summary statistics. Psychometrika, 1956, 21, 79-88.

Helmes, E., & Jackson, D. N. The item factor structure of the Personality Research Form. Applied Psychological Measurement, 1977, 1, 185-194.

Hill, M. O. Correspondence analysis: A neglected multivariate method. Journal of the Royal Statistical Society, Series C. Applied Statistics, 1974, 23, 340-354.

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

Jackson, D. N. Personality Research Form Manual. Revised edition. Port Huron, Michigan: Research Psychologists Press, 1974.

Jackson, D. N. Jackson Personality Inventory Manual. Port Huron, Michigan: Research Psychologists Press, 1976.

Jackson, D. N., & Messick, S. Individual differences in social perception. British Journal of social and clinical psychology, 1963, 2, 1-10.

Kaiser, H. F. Formulas for component scores. Psychometrika, 1962, 27, 83-87.

McDonald, R. P., & Ahlawat, K. S. Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 1974, 27, 82-99.

Messick, S., & Jackson, D. N. Judgmental dimensions of psychopathology. Journal of consulting and clinical psychology, 1972, 38, 418-427.

Schönemann, P. H. A generalized solution of the orthogonal Procrustes problem. Psychometrika, 1966, 31, 1-10.

Shepard, R. N. A taxonomy of some principal types of data and of multidimensional methods for their analysis. pp. 21-47 in R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), Multidimensional scaling: Theory and applications in the behavioral sciences, Vol. 1. New York: Seminar Press, 1972.

Slater, P. The analysis of personal preferences. British Journal of Statistical Psychology, 1960, 13, 119-135.

Stewart, T. R. Generality of multidimensional representations. Multivariate Behavioral Research, 1974, 9, 507-519.

Ten Berge, J. M. F. Orthogonal Procrustes rotation for two or more matrices. Psychometrika, 1977, 42, 267-276.

Thurstone, L. L. Theory of attitude measurement. Psychological Review, 1929, 36, 222-241.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

Torgerson, W. S. Multidimensional scaling of similarity. Psychometrika, 1965, 30, 379-393.

Tucker, L. R. Description of paired comparison preference judgments by a multidimensional vector model. Princeton, N.J.: Educational Testing Service, Research Memorandum, 55-57, 1955.

Tucker, L. R. Relation of factor score estimates to their use. Psychometrika, 1971, 36, 427-436.

Tucker, L. R., & Messick, S. An individual differences model for multidimensional scaling. Psychometrika, 1963, 28, 333-367.

Wiggins, J. S. Personality and prediction: Principles of personality assessment. Reading, Mass.: Addison-Wesley, 1973.

Footnotes

Supported by Canada Council Research Grant NO. S75-0495-X2.

[1] An earlier version of this paper was presented to the Society for Multivariate Experimental Psychology, State College, Pennsylvania, November 11, 1976. We thank D. Chan for his comments on an earlier draft.

[2] Complete tables of orthogonal and oblique content scores for PRF-E have been deposited with NAPS. See NAPS document No. 03198 for 53 pages of supplementary material. Order from ASIS/NAPS c/o Microfiche Publications, P.O. Box 3513, Grand Central Station, New York, New York, 10017. Remit in advance for each NAPS accession number. Institutions and organizations may use purchase orders when ordering. However, there is a billing charge of $5.00 for this service. Make cheques payable to Microfiche Publications. Photocopies are $12.25. Microfiche are $3.00. Outside of the United States and Canada, postage is $3.00 for photocopy or $1.00 for fiche.

TABLE 1

Illustration of Content Scaling of 16 JPI Items

Item No.	Scale	Keying	Item	Content Scale Values Cny	Rkt	Ses	Vlo
36	Cny	T	In most situations, I usually agree with the opinions of the group.	1.06	.25	.56	.61
52	Cny	F	When I want to purchase something, I rarely consider other people's opinion of it.	-2.12	.30	.02	.60
100	Cny	T	It makes me feel uncomfortable to be dressed ddfferently from those around me.	1.93	.22	-1.08	- .13
212	Cny	F	I do what I please, not what others say I should do.	- .70	- .74	.50	-1.37
42	Rkt	T	I would enjoy bluffing my way into an exclusive club or private party.	.75	2.34	- .29	- .60
218	Rkt	F	I consider security an important element in every aspect of my life.	.92	-1.22	1.13	.70
170	Rkt	T	I enjoy taking risks.	.10	1.89	.90	.02
314	Rkt	F	I rarely make even small bets.	- .97	- .82	-1.46	- .49
59	Ses	T	I rarely feel self-conscious in a strange group.	-1.40	.44	.78	1.46
11	Ses	F	I make a better follower than a leader	.40	-1.08	-1.44	.27
315	Ses	T	I find it easy to introduce people.	- .10	- .55	1.91	- .75
43	Ses	F	I have never been a very popular person.	- .85	.02	-1.63	.41
63	Vlo	T	My values might seem a little old-fashioned by modern standards.	.99	-1.05	- .04	1.03
175	Vlo	F	Married people who no longer love each other should be given a divorce.	- .02	- .98	.71	-1.85
223	Vlo	T	People today don't have enough respect for authority.	.12	.18	.10	1.60
303	Vlo	F	People respect tradition more than necessary.	- .12	.90	- .68	-1.52

Abbreviations: Cny - Conformity; Rkt - Risk Taking; Ses - Self Esteem; Vlo - Value Orthodoxy

T - True keyed; F - False keyed

TABLE 2

Scale Values and Items for the Achievement Scale

Item Number	Scale Value	Item
220	3.94	I often set goals that are very difficult to reach.
90	3.46	I will not be satisfied until I am the best in my field of work.
178	2.44	My goal is to do at least a little bit more than anyone else has done before.
310	2.35	I don't mind working while other people are having fun.
2	2.33	People should be more involved with their work.
46	2.20	I enjoy difficult work.
134	1.85	I would work just as hard whether or not I had to earn a living.
266	0.60	As a child I worked a long time for some of the things I earned.
332	0.01	I am not really very certain what I want to do or how to go about doing it.
156	-0.33	I do not let my work get in the way of what I really want to do.
68	-1.23	I have rarely done extra studying in connection with my work.
244	-1.71	People seldom think of me as a hard worker.
200	-1.78	In my work I seldom do more than is necessary.
112	-1.86	I try to work just hard enough to get by.
24	-3.19	I seldom set standards which are difficult for me to reach.
288	-4.13	It doesn't really matter to me whether or not I become one of the best in my field.

TABLE 3

Scale Values and Items for the Dominance Scale

Item Number	Scale Value	Item
317	4.73	I would like to be an executive with power over others.
229	4.41	The ability to be a leader is very important to me.
9	3.68	I feel confident when directing the activities of others.
97	3.68	I try to control others rather than permit them to control me.
53	3.47	I would like to be a judge.
141	3.35	I would like to play a part in making laws.
185	2.34	In an argument, I can usually win others over to my side.
273	2.06	I am quite effective in getting others to agree with me.
295	-2.24	I am not very insistent in an argument.
251	-2.44	Most community leaders do a better job than I could possibly do.
339	-2.56	I would not want to have a job enforcing the law.
31	-3.48	I would make a poor military leader.
119	-4.03	I don't like to have the responsibility for directing the work of others.
75	-4.23	I avoid positions of power over other people.
207	-4.38	I feel uneasy when I have to tell people what to do.
163	-4.89	I have little interest in leading people.

TABLE 4

Mean Scale Values of Keyed and Non-Keyed Items

Scale	16 Keyed Items	336 Non-Keyed Items
Abasement	1.88	.71
Achievement	2.08	.69
Affiliation	2.25	.69
Aggression	2.05	.68
Autonomy	2.30	.66
Change	2.43	.63
Cognitive Structure	2.28	.68
Defendence	2.40	.71
Dominance	3.51	.51
Endurance	2.24	.67
Exhibition	2.26	.62
Harmavoidance	3.32	.55
Impulsivity	2.49	.65
Nurturance	2.48	.65
Order	3.64	.47
Play	2.42	.67
Sentience	2.67	.65
Social Recognition	2.66	.63
Succorance	2.67	.64
Understanding	2.80	.59
Infrequency	2.47	.69
Desirability	1.75	.69

Note: False keyed items have been reflected.

APPENDIX 2. EXPERIMENTAL MATERIALS

How Well Do You Judge Personality?

Statements about individuals differ in the degree to which they accurately describe those individuals. Consider the fictional character Tarzan, who is generally thought of as courageous, free and sportsmanlike. Now look at the statements below in terms of how characteristic a 'True' response to them is of Tarzan.

	Extremely Uncharacteristic				Neither Characteristic Nor Uncharacteristic				Extremely Characteristic
A. I enjoy being outdoors.	1	2	3	4	5	6	7	8	[✓ 9]
B. People think I am often afraid.	1	[✓ 2]	3	4	5	6	7	8	9
C. My work is always done on time.	1	2	3	4	[✓ 5]	6	7	8	9

One individual thought that a 'True' response to the statement "I enjoy being outdoors." by Tarzan was extremely characteristic of Tarzan and circled 9. On the other hand he thought that a 'True' response by Tarzan to "People think I am often afraid." was very uncharacteristic of Tarzan, and therefore circled 2. In the case of "My work is always done on time." he thought a 'True' response by Tarzan was neutral, that is, it was neither characteristic nor uncharacteristic of Tarzan, and so circled 5.

In a similar manner, we would like you to judge a series of statements for each of several target individuals. Read the description of each individual carefully twice. Then judge the degree to which a 'True" response to each statement by the target is characteristic of the target and circle the appropriate number as indicated in the examples.

For each statement, circle the one number that best indicates the degree to which the statement is characteristic of the individual. Try to use all the categories. Do not skip any statements.

George Franklin

George works for a large advertising agency as a personal assistant to one of the vice-presidents. His job involves arranging appointments for his boss and generally ensuring that everything is done according to his superiors' wishes.

If anything goes wrong, George usually gets the blame, but he never complains about any injustice. He is deferential to his superiors and is always willing to carry out their wishes. He tends to minimize his importance in the agency and is resigned to his role in life.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Carl Bates

Carl is an aspiring lightweight boxer. He took up boxing on the advice of friends as a way to direct his combative nature. In the ring he is aggressive and is noted for attacking his opponent.

He often says he doesn't let anyone push him around, and is known as having a hot temper. In conversation with others, Carl is often argumentative and is prone to making biting and caustic comments about others.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Tom Harrison

Tom works as an electrician. He is currently looking for a new job because he says the one he currently works at has become boring. He has had several jobs in the last few years, and moves frequently as well. "I get tired of doing the same thing and seeing the same people all the time," he says. He fits in quickly whenever he gets a new job.

Tom has had many hobbies, but never stays at any of them very long. He frequently gives up doing something to start something new.

His friends say he is unpredictable and changeable.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Frank Harris

Frank is the general manager of a small corporation. He is known for his tight control over his staff and his unwil'ingness to delegate any authority to others.

He has a forceful personality and is very persuasive in an argument. His friends note that Frank usually dominates the conversation wherever he is and seems to enjoy telling people what to do.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Harry Mason

Harry works in a small manufacturing plant where he is noted for his perseverance. He never gives up on something he has started.

However, Harry's main pleasure in life is marathon running. He has taken part in marathons for the last 8 years and trains by running at least 15 miles per day. He runs every day, regardless of the weather.

His friends describe him as patient, determined and a steadfast friend.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Tom Bradley

Tom Bradley has a reputation as one of the best TV stuntmen. His work involves jumping from buildings and cliffs, car crashes and fight scenes. "I get a kick out of it all" he says.

His main hobby is sport fishing for sharks and he is an avid fan of the thrill rides at amusement parks. He says everyone should get a good scare every now and then to add a bit of spice to life.

His friends say he is adventurous to the point of recklessness.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Murray Wilson

Murray Wilson is an accountant. Most of his work involves making sure his clients' books are in order and balanced. He checks new accounts carefully and immediately files them in their place. Murray is noted for his tidiness and the methodical way he does his work.

He is always on time and his wife oten remarks that she never has to remind Murray of anything. She describes him as deliberate and systematic.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

George Burgess

George is a chef in a large hotel. He is noted for the delicate seasonings he uses and the general excellence of his meals.

He is known as an art critic but spends a good deal of time at his home in the country. There he goes for walks in the woods, enjoying the "simple delights of life," as he puts it. He delights in pointing out small changes in the plants and wildlife of the woods to his less observant friends. His close friends say he is sensitive and perceptive.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Robert Jansen

Robert works for a large corporation where he is known for his impeccable manners and courtesy toward everyone. Nevertheless, he worries about what people think of him. "I try hard to make people like me" he says. He is careful to dress conventionally and never does anything out of the ordinary.

He rarely disagrees with anyone and generally goes along with whatever is said. His friends describe him as agreeable, proper and sensitive toward others.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Alfred Anderson

Alfred Anderson is a science student at a large university. He is fascinated by the many aspects of science and spends as much time as he can reading and working at his labs.

He enjoys reading history books and biographies of historical figures. His friends report that Alfred has an endless curiosity about almost everything. They describe him as logical and thoughtful.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are are to judge how characteristic a 'True' response is of the target individual.

PRF-E Items Used for Judgement Task

Target	George Franklin	Carl Bates	Tom Harrison	Frank Harris	Harry Mason
Scale	Abasement	Aggression	Change	Dominance	Endurance
Item Order					
1	80	334	72	339	343
2	111	79	266	269	118
3	45	184	138	46	32
4	293	202	116	273	10
5	35	95	292	327	340
6	309	290	6	112	296
7	89	136	94	75	98
8	96	312	314	119	274
9	265	224	231	229	318
10	67	246	160	295	63
11	221	158	204	141	252
12	133	70	226	317	227
13	23	268	50	97	186
14	199	114	182	251	76
15	1	26	248	163	230
16	177	48	336	53	120
17	331	92	113	207	142
18	287	4	270	31	164
19	243	3	58	185	208
20	155	180	28	9	54

Note: John Tyler target (negative Abasement) was judged on the items following George Franklin.

PRF-E Items Used for Judgement Task

Target	Tom Bradley	Murray Wilson	George Burgess	Robert Jansen	Alfred Anderson
Scale	Harm-avoidance	Order	Sentience	Social Recognition	Understanding
Item Order					
1	298	323	99	245	328
2	12	279	104	150	196
3	102	81	127	216	350
4	78	169	193	194	341
5	100	59	303	304	130
6	56	15	203	40	64
7	333	345	61	172	217
8	276	257	39	238	16
9	188	125	149	85	240
10	144	191	281	62	313
11	232	235	83	282	306
12	210	37	171	260	284
13	320	213	14	84	42
14	122	211	325	106	152
15	140	147	347	326	174
16	166	170	259	128	20
17	115	103	215	187	86
18	342	139	237	302	108
19	254	332	105	18	262
20	34	301	17	348	218

John Tyler

John Tyler works as a doorman at a luxury hotel. He is fairly content in his job, but dislikes having to run errands for guests. His superiors have noted that John rarely accepts the blame for any mistakes, but is willing to accept credit for good work.

He takes pride in his appearance in his uniform and likes to use his position to his own advantage.

His brother says he is one of the most self-confident people he knows.

Read the description of the target individual twice. Then write his name at the top of an answer sheet and judge the statements on the following page in the same manner as the examples. Remember that you are to judge how characteristic a 'True' response is of the target individual.

Target Name ________ __________________

	Extremely Uncharacteristic				Neither Characteristic Nor Uncharacteristic				Extremely Characteristic
1	1	2	3	4	5	6	7	8	9
2	1	2	3	4	5	6	7	8	9
3	1	2	3	4	5	6	7	8	9
4	1	2	3	4	5	6	7	8	9
5	1	2	3	4	5	6	7	8	9
6	1	2	3	4	5	6	7	8	9
7	1	2	3	4	5	6	7	8	9
8	1	2	3	4	5	6	7	8	9
9	1	2	3	4	5	6	7	8	9
10	1	2	3	4	5	6	7	8	9
11	1	2	3	4	5	6	7	8	9
12	1	2	3	4	5	6	7	8	9
13	1	2	3	4	5	6	7	8	9
14	1	2	3	4	5	6	7	8	9
15	1	2	3	4	5	6	7	8	9
16	1	2	3	4	5	6	7	8	9
17	1	2	3	4	5	6	7	8	9
18	1	2	3	4	5	6	7	8	9
19	1	2	3	4	5	6	7	[illegible]	9
20	1	2	3	4	5	6	7	8	9

Part Two

In this task we would like you to judge the desirability of each of the target individuals. Look over the descriptions of each of the targets again. Then rate how desirable you think each individual target is in the space provided on the right.

		Extremely Undesirable				Neither Desirable Nor Undesirable				Extremely Desirable
1.	Alfred Anderson	1	2	3	4	5	6	7	8	9
2.	Carl Bates	1	2	3	4	5	6	7	8	9
3.	Tcm Bradley	1	2	3	4	5	6	7	8	9
4.	George Burgess	1	2	3	4	5	6	7	8	9
5.	George Franklin	1	2	3	4	5	6	7	8	9
6.	Frank Harris	1	2	3	4	5	6	7	8	9
7.	Tom Harrison	1	2	3	4	5	6	7	8	9
8.	Harry Mason	1	2	3	4	5	6	7	8	9
9.	Robert Jansen	1	2	3	4	5	6	7	8	9
10.	Murray Wilson	1	2	3	4	5	6	7	8	9
11.	John Tyler	1	2	3	4	5	6	7	8	9

Judged Desirability of Targets

Target	Scale	Mean Rating
AA	Understanding	5.68
CB	Aggression	2.65
TB	Harmavoidance (negative)	5.29
GB	Sentience	7.39
GF	Abasement	4.77
FH	Dominance	3.68
TH	Change	5.45
HM	Endurance	6.87
RJ	Social Recognition	5.39
MW	Order	5.87
JT	Abasement (negative)	4.65
	Mean	5.24
	Standard deviation	1.27

APPENDIX 3. ANALYSIS OF VARIANCE RESULTS

Invariant Models (University Sample)
Invariant Models (High School Sample)
Threshold Models (University Sample)
Threshold Models (High School Sample)
Seven Best Prediction Models (University Sample)
Seven Best Prediction Models (High School Sample)

Note: The following convention has been used in reporting the results of the Neuman-Keuls tests. Any means underlined by a common line do not differ among themselves at the .01 level of significance. Any means which do not share a common line are significantly different at the .01 level

```
TWO-WAY ANOVA FOR      UWO INVARIANT  PREDICTION MODELS

TREATMENT GROUP MEANS   GROUP     TFKEY    DISJ    JDSY    PVAL    RAND
                                  .0585    .0189   .2618   .3132   -.0194

     SOURCE      DF      SUM SQUARES     MEAN SQUARE     F-RATIO

  TREATMENT      91         2.22968          .02450        3.21883 ***
      BLOCK       4         8.34154         2.08539      273.95740 ***
      ERROR     364         2.77080          .00761
      TOTAL     459        13.34202
```

Results of Neuman-Keuls test

*** p<.001

Model	RAND	DISJ	TFKEY	JDSY	PVAL
Mean	-.0194	.0189	.0585	.2618	.3132

Abbreviations: TFKEY - True-false scoring key; DISJ - Disjunctive content; JDSY - Judged desirability; PVAL - P-values; RAND - Random.

```
TWO=WAY ANOVA FOR  INVARIANT MODELS FOR HIGH SCHOOL SAMPLE = KAPPA

TREATMENT GROUP MEANS   GROUP     TFKEY    DISJ     JDSY   P=VAL    RAND
                                  .0098   =.0064    .1862   .2498   =.0113

     SOURCE       DF       SUM SQUARES       MEAN SQUARE      F=RATIO

   TREATMENT     300         9.56490            .03188       4.39489 ***
       BLOCK       4        18.27350           4.56837     629.72455 ***
       ERROR    1200         8.70547            .00725
       TOTAL    1504        36.54387
```

Results of Neuman-Keuls test

Model	RAND	DISJ	TFKEY	JDSY	P-VAL
Mean	-.0113	-.0064	.0098	.1862	.2498

*** p<.001

Abbreviations: TFKEY - True-false scoring key; DISJ - Disjunctive content; JDSY - Judged desirability; P-VAL - Pvalues; RAND - Random.

```
TWO-WAY ANOVA FOR  THRESHOLD MODELS FOR   UWO       SAMPLE - KAPPA

TREATMENT GROUP MEANS    GROUP      RAND    LGFA    LGSC    SHFA    SHSC
                                   .0176   .0457   .0150   .0545   .0432

     SOURCE       DF       SUM SQUARES       MEAN SQUARE      F-RATIO

   TREATMENT      91          3.32214           .03651        5.97747 ***
       BLOCK       4           .11639           .02910        4.76410 ***
       ERROR     364          2.22311           .00611
       TOTAL     459          5.66163
```

Results of Neuman-Keuls test

Model	LGSC	RAND	SHSC	LGFA	SHFA
Mean	.0151	.0176	.0432	.0457	.0545

*** p<.001

Abbreviations: RAND - Random; LGFA - Long vectors, factor content; LGSC - Lond vectors, scale content; SHFA - Short vectors, factor content; SHSC - Short vectors, scale content.

```
TWO-WAY ANOVA FOR  THRESHOLD MODELS FOR HIGH SCHOOL SAMPLE - KAPPA

TREATMENT GROUP MEANS   GROUP       RAND     LGFA     LGSC     SHFA     SHSC
                                   .0151   -.0086   -.0031   -.0092   -.0002

    SOURCE     DF       SUM SQUARES       MEAN SQUARE     F-RATIO

  TREATMENT   300        14.04789           .04683       8.05292 ***
      BLOCK     4          .11645           .02911       5.00671 ***
      ERROR  1200         6.97779           .00581
      TOTAL  1504        21.14213
```

Results of Neuman-Keuls test

*** p<.001

Model	SHFA	LGFA	SHFA	SHSC	RAND
Mean	-.0092	-.0086	-.0031	-.0002	.0151

Abbreviations: RAND - Random; LGFA - Long vectors, factor content; LGSC - Long vectors, scale content; SHFA - Short vectors, factor content; SHSC - Short vectors, scale content.

```
TWO-WAY ANOVA FOR  BFST PREDICTION MODELS FOR     UWO     SAMPLE = KAPPA

 TRFATMENT GROUP MFANS   GROUP     DSYTH   DYINV   P-VAL   SCORE   2DSCR    MD-6   MD-22
                                   .0967   .2618   .3132   .3034   .3098   .4751   .5406

     SOURCE       DF      SUM SQUARES     MEAN SQUARE     F-RATIO

  TREATMENT       91        2.68820          .02954       5.72319 ***
      BLOCK        6       11.57954         1.92992     373.90217 ***
      ERROR      546        2.81822          .00516
      TOTAL      643       17.08597
```

Results of Neuman-Keuls test

Model	DSYTH	DYINV	SCORE	2DSCR	P-VAL	MD-6	MD-22
Mean	.0967	.2618	.3034	.3098	.3132	.4751	.5406

*** p<.001

Abbreviations: DSYTH - Jackson's threshold; DYINV - Invariant desirability; P-VAL - Invariant p-values; SCORE - Scale score; 2DSCR - Two-dimensional scale score; MD-6 - Multidimensional spatial, factors; MD-22 - Multidimensional spatial, scales.

```
TWO-WAY ANOVA FOR  BEST PREDICTION MODELS FOR HIGH SCHOOL SAMPLE - KAPPA

 TREATMENT GROUP MEANS   GROUP     DSYTH   DYINV   P-VAL   SCORE   2DSCR   MD-6   MD-22
                                   .1891   .1862   .2498   .2735   .2781   .4151   .4953

    SOURCE    DF     SUM SQUARES      MEAN SQUARE     F-RATIO

  TREATMENT   300      10.82274          .03608       5.07765***
      BLOCK     6      24.17275         4.02879     567.05086***
      ERROR  1800      12.78867          .00710
      TOTAL  2106      47.78417
```

Results of Neuman-Keuls test

Model	DSYTH	DYINV	P-VAL	SCORE	2DSCR	MD-6	MD-22
Mean	.1891	.1862	.2498	.2735	.2781	.4151	.4953

*** p<.001

Abbreviations: DSYTH-Jackson's threshold; DYINV - Invariant desirability; P-VAL - Invariant p-values; SCORE - Scale score; 2DSCR - Two-dimensional scale score; MD-6 - Multidimensional spatial, factors; MD-22 - Multidimensional spatial, scales.

REFERENCES

Bejar, I. I. An application of the continuous response level model to personality measurement. Applied Psychological Measurement, 1977, 1, 509-521.

Bentler, P. M. A lower-bound method for the dimension-free measurement of internal consistency. Social Science Research, 1972, 1, 343-357.

Berg, I. A. Response set in personality assessment. Chicago: Aldine, 1967.

Bishop. Y. M. M., Fienberg, S. E. & Holland. P. W. Discrete multivariate analysis: Theory and practice. Cambridge, Mass.: MIT Press, 1975.

Blackmore, W. R. Some comments on "Computer simulation of a model of neurotic defense processes". Behavioral Science, 1972, 17, 229-232.

Boyd, J. E. & Jackson, D. N. An empirical evaluation of judgment and response methods in multivariate attitude scaling. American Psychologist, 1966, 21, 718. (Abstract)

Bruner, J. S. & Taguiri, R. The perception of people. pp. 634-654 in G. Lindzey (Ed.) Handbook of Social psychology. Vol. 2. Reading, Mass.: Addison-Wesley, 1954.

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

Cliff, N. Adjective check list responses and individual differences in perceived meaning. Educational and Psychological Measurement, 1968, 28, 1063-1077.

Cliff, N. Further study of cognitive processing models for inventory response. Applied Psychological Measurement, 1977, 1 41-49.

Cliff, N., Bradley, P. Girard, R. The investigation of cognitive models for inventory response. Multivariate Behavioral Research, 1973, 8, 407-425.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Colby, K. M. Computer simulation of a neurotic process. pp. 165-179 in S. S. Tomkins & S. Messick (Eds.) Computer simulation of personality. New York: Wiley, 1963.

Colby, K. M., Weber, S. & Hilf, F. D. Artificial paranoia. *Artificial Intelligence*, 1971, *2*, 1-25.

Coombs, C H. *A theory of data*. New York: Wiley, 1964.

Cranton, P. A. Computer models of personality: Implications for measurement. *Journal of Personality Assessment*, 1976, *40*, 454-463.

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

Damarin, F. A latent-structure model for answering personal questions. *Psychological Bulletin*, 1970, *73*, 23-40.

Diederich, G. W., Messick, S. J. & Tucker, L. W. A general least squares solution for successive intervals. *Psychometrika*, 1957, *22*, 159-173.

Dutton, J. M. & Briggs, W. G. Simulation model construction. pp. 103-126 in J. M. Dutton & W. H. Starbuck (Eds.) *Computer simulation of human behavior*. New York: Wiley, 1971.

Edwards, A. L. *The measurement of personality traits by scales and inventories*. New York: Holt, Rinehart & Winston, 1970.

Fiske, D. W. Homogeneity and variation in measuring personality. *American Psycholcgist*, 1963, *18*, 643-652.

Fiske, D. W. Some hypotheses concerning test adequacy. *Educational and Psychological Measurement*, 1966, *26*, 69-88.

Goldberg, L. W. A model of item ambiguity in personality assessment. *Educational and Psychological Measurement*, 1963, *23*, 467-492.

Green, B. F. A method of scalogram analysis using summary statistics. *Psychometrika*, 1956, *21*, 79-88.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. pp. 319-348 in P. Horst (Ed.) *The prediction of personal adjustment*. New York: Social Science Research Council, 1941.

Guttman, L. The basis for scalogram analysis. pp. 60-90 in S. A. Stouffer et al. (Eds.) *Measurement and prediction*. Princeton , N. J.: Princeton University Press, 1950.

Hays, W. L. An approach to the study of trait implication and trait similarity. pp. 289-299 in R. Taguiri & L. Petrullo (Eds.) *Person perception and interpersonal behavior*. Stanford: Stanford University Press, 1958.

Helmes, E. & Jackson, D. N. The item factor structure of the Personality Research Form. *Applied Psychological Measurement*. 1977, *1*, 185-194.

Helmes, E., Reed, P. L. & Jackson, D. N. Desirability and frequency scale values and endorsement proportions for items of Personality Research Form-E. *Psychological Reports*, 1977, *41*, 435-444.

Helmstadter, G. C. Procedures for obtaining separate set and content components of a test score. *Psychometrika*, 1957, *22*, 381-393.

Horst, P. *Personality: Measurement of dimensions*. San Francisco: Jossey-Bass, 1968.

Jackson, D. N. A threshold model for stylisitic responding. In C. Hanley (Chm.) New interpretations of response style and content in personality assessment. Symposium presented at the meeting of the American Psychological Association. San Francisco, 1968.

Jackson, D. N. A sequential system for personality scale development. In C. D. Spielberger (Ed.) *Current topics in clinical and community psychology*. New York: Academic Press, 1970.

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

Jackson, D. N. *Personality Research Form manual*. Rev. ed. Goshen, New York: Research Psychologists Press, 1974.

Jackson, D. N. The appraisal of person-reliability. Paper presented at the meeting of the Society of Multivariate Experimental Psychology, State College, Penn., 1976.

Jackson, D. N. & Helmes, E. A factor analytic approach to content scaling. Presented at the meeting of the Society of Multivariate Experimental Psychology, State College, Penn., 1976.

Jackson, D. N. & Messick, S. A distinction between judgments of frequency and of desirability as determinants of response. *Educational and Psychological Measurement*, 1969, *29*, 273-293.

Kuncel, R. B. Response processes and relative location of subject and item. *Educational and Psychological Measurement*, 1973, *33*, 545-563.

Kuncel, R. B. The subject-item interaction in itemmetric research. *Educational and Psychological Measurement*, 1977, *37*, 665-678.

Kuncel, R. B. & Fiske, D. W. Stability of response process and response. *Educational and Psychological Measurement*, 1974, *34*, 743-755.

Lazarsfeld, P. F. Latent structure analysis. pp. 476-543 in S. Koch (Ed.) *Psychology: A study of a science*. Vol. 3. New York: McGraw-Hill, 1959.

Lingoes, J. C. A general survey of the Guttman-Lingoes nonmetric program series. pp. 49-68 in R. N. Shepard, A. K. Romney & S. K. Nerlove (Eds.) *Multidimensional scaling: Theory and applications in the behavioral sciences*. Vol. 1. New York: Seminar Press, 1972.

Loehlin, J. C. Word meanings and self-descriptions. *Journal of Abnormal and Social Psychology*, 1961, *62*, 28-34.

Loehlin, J. C. Word meanings and self-descriptions: A replication and extension. *Journal of Personality and Social Psychology*, 1967, *5*, 107-110.

Loehlin, J. C. *Computer models of personality*. New York: Random House, 1968.

Loevinger, J. The technic of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, 1948, *45*, 507-529.

Lord, F. M. Some test theory for tailored testing. pp. 139-183 in W. H. Holtzman (Ed.) *Computer-assisted instruction, testing, and guidance*. New York: Harper and Row, 1970.

Lord, F. M. & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Lumsden, J. Test theory. *Annual Review of Psychology*, 1976, *27*, 251-280.

Lumsden, J. Person reliability. *Applied Psychological Measurement*, 1977, *1*, 477-482.

Moser, U. von Zeppelin, I. & Schneider, W. Computer simulation of a model of neurotic defence processes. *International Journal of Psycho-Analysis*, 1969, *50*, 53-64.

Moser, U., von Zeppelin, I. & Schneider, W. Computer simulation of a model of neurotic defense processes. *Behavioral Science*, 1970, *15*, 194-202.

Mosteller, F. & Tukey, J. W. Data analysis, including statistics. pp. 80-203 in G. Lindzey & E. Aronson (Eds.) *Handbook of social psychology*. (2nd ed.) Vol. 2. Reading, Mass.: Addison- Wesley, 1968.

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

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danmarks Paedagogiske Institut, 1960.

Rogers, T. B. The process of responding to personality items: Some issues, a theory and some research. Multivariate Behavioral Research Monographs, 1971, 6(2).

Rogers, T. B. Ratings of content as a means of assessing personality items. Educational and Psychological Measurement, 1973, 33, 845-858.

Rogers, T. B. An analysis of the stages underlying the process of responding to personality items. Acta Psychologica, 1974, 38, 205-213. (a)

Rogers, T. B. An analysis of two central stages underlying responding to personality items: The self-referent decision and response selection. Journal of Research in Personality, 1974, 8, 128-138. (b)

Rogers, T. B. Self-reference in memory: Recognition of personality items. Journal of Research in Personality, 1977, 11, 295-305.

Scott, W. A. Attitude measurement. pp. 204-273 in G. Lindzey & E. Aronson (Eds.) Handbook of social psychology. (2nd. ed.) Vol. 2. Reading, Mass.: Addison-Wesley, 1968.

Shepard, R. N. A taxonomy of some principal types of data and of multidimensional methods for their analysis. pp. 21-47 in R. N. Shepard, A. K. Romney & S. K. Nerlove (Eds.) Multidimensional scaling: Theory and applications in the behavioral sciences. Vol. 1. New York: Seminar Press, 1972.

Shepard, R. N. & Carroll, J. D. Parametric representaion of nonlinear data structures. pp. 561-592 in P. R. Krishnaiah (Ed.) Multivariate analysis. New York: Academic Press, 1966.

Skinner, H. A., Jackson, D. N. & Rampton, G. M. The Personality Research Form in a Canadian context: Does language make a difference? Canadian Journal of Behavioral Science, 1976, 8, 156-168.

Snedecor, G. W. & Cochran, W. G. Statistical methods. (6th. ed.) Ames, Iowa: Iowa State University Press, 1967.

Spence, I. & Domoney, D. W. Single subject incomplete designs for nonmetric multidimensional scaling. Psychometrika, 1974, 39, 469-490.

Stewart, T. R. Generality of multidimensional representations. Multivariate Behavioral Research, 1974, 9, 507-519.

Tomkins, S. S. and Messick, S. Computer simulation of personality. New York: Wiley, 1963.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

Turner, C. B. & Fiske, D. W. Item quality and appropriateness of response processes. Educational and Psychological Measurement, 1968, 28, 297-315.

Tyler, T. A. Response stability, person-item distance and homogeneity. Unpublished doctoral dissertation, University of Chicago, 1968.

Voyce, C. D. An evaluation of a threshold theory for personality assessment based on the Differential Personality Inventory. Unpublished M. A. thesis, University of Western Ontario, 1973.

Voyce, C, D. & Jackson, D. N. An evaluation of a threshold theory for personality assessment. Educational and Psychological Measurement, 1977, 37, 383-408.

Wright, B. D. & Mead, R. J. CALFIT: Sample-free item calibration with a Rasch measurement model. Research Memorandum 18, Statistical Laboratory, Department of Education, University of Chicago, 1975.

Wright, B. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.