1983

Comparison Of Four Models For Predicting Person Reliability

Geziena Cynthia Fekken

COMPARISON OF FOUR MODELS FOR PREDICTING PERSON RELIABILITY

by

Geziena Cynthia Fekken

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
September, 1983

ABSTRACT

The purpose of this research was to examine the efficacy of four models for predicting person reliability, which was defined as across-session item consistency. Two prediction models were based on item characteristics alone, namely, p-values and social desirability scale values. Two prediction models attempted to take into account both item and person characteristics. These models were based on individuals' latencies for responding to particular items and on individuals' thresholds for answering items in terms of some item characteristic. The latter model was derived from Jackson's (1968) threshold model for responding which describes the response process with reference to a threshold, marking an individual's transition from the tendency to reject to the tendency to endorse items relative to some item characteristic.

Models were evaluated in three separate studies. In Study 1, models were compared using test-retest responses to personality items relatively neutral in desirability and measuring normal personality content. Study 2 focussed on the role of item characteristics. Three models, the two based on item characteristics and one on threshold theory, were compared using test-retest responses to personality items measuring psychopathological content and ranging widely in social desirability scale values and p-values.

Study 3 emphasized decision difficulty. Decision difficulty was experimentally manipulated through a rating task designed to induce desirable responding. Decision difficulty was also examined by comparing prediction based on subjects' own ratings of item desirability to prediction based on independently derived desirability scale values.

Overall, the response latency model was found consistently to be the strongest predictor of exactly which items individuals changed on retest. Theoretically, this finding was interpreted as support for distinguishing between the influences of item and person characteristics on the process of responding to personality items. Practically, this finding has applications for constructing person reliability indices. Results failed to support the hypothesis that the threshold model, which also took into account person and item characteristics, was a better predictor than models based on item characteristics alone. The relative lack of success of this model was considered in light of the need to specify correctly the item property underlying responding.

ACKNOWLEDGEMENT

First and foremost, I would like to offer my sincerest thanks to my advisor, Dr. Douglas N. Jackson, for his continuing guidance and support during all the phases of this research. I would also like to thank the members of my various committees, Dr. E. Helmes, Dr. N. A. Kuiper, Dr. R. W. J. Neufeld, Dr. H. Skinner, Dr. N. Skinner and Dr. P. A. Vernon, for their valuable contributions to this research. The time and effort spent by all of these individuals on my behalf is greatly appreciated.

Finally, I would like to extend my thanks and appreciation to my husband, Ron. Not only did he offer endless academic and programming advice, but, most important of all, he offered the moral support I needed to finish this project.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Human behavior shows remarkable variation. The focus of much behavioral assessment has been the reliable measurement of consistent patterns of individual differences. However, the consistency or stability of behavior shown by a single individual also deserves attention. Behavioral differences among individuals can only be reliably assessed if individuals themselves are behaving lawfully or consistently. The consistency of behavior shown by one individual is referred to as person reliability. Person reliability has been assessed using overt behaviors (cf. Schneiderman, 1980), but, most frequently, person reliability is assessed in the context of structured inventories (cf. Fiske, 1957a; Jackson, 1977). This is not surprising: the reliability of structured inventories has long been recognized as critical for the measurement of individual differences. The logical extension of this concern is that an individual's responses to the inventory must also show reliability if individual differences are to be measured. Only if individuals answer an inventory lawfully or consistently can differences be reliably assessed.

The purpose of this research is to examine the stability of an individual's responses to structured inventories by comparing a number of prediction models. The four models considered in this research are based on: (1) the endorsement probabilities or p-values of individual items; (2) the social desirability scale values of individual items; (3) persons' latencies for responding to individual items; and (4) the proximity of items to persons' thresholds for responding to items in terms of some item characteristic. The evaluation of these models will proceed in the following manner. First, person reliability will be defined and evidence for the reliability and generalizability of various general indices of person reliability will be discussed. The argument will be made that it is appropriate and useful to attempt to predict the stability of individual item responses for a single respondent. The four models for predicting item stability will be presented. The empirical data supporting the conceptualization of each of these models will be reviewed. Finally, the efficacy of each of the models will be compared in three separate but complementary studies.

Defining Person Reliability

Traditionally, reliability has been conceptualized in terms of tests. That is, reliability coefficients are calculated for test scores based on the responses of a number of individuals to the same set of items. It is also possible to calculate reliability coefficients for persons based on their responses to a number of sets of items. Such indices of person reliability are of interest for a variety of reasons. Jackson (1976) has suggested that a coefficient of person reliability may be useful as an index of random responding, as an index of domain-articulation, and as a measure of response style. A person reliability coefficient may also be potentially helpful for identifying subgroups with different factor structures (Jackson, 1976) or subgroups for which intra-individual variability may be used as a moderator variable to improve prediction (Bem & Allen, 1974; Ghiselli, 1956; 1963). In addition, in settings where decisions regarding clinical treatment are based on psychological test data, a person reliability index may be important for establishing the dependability of a clinical respondent's data. Finally, when particular test items are of critical interest as, for example, in the domain of psychopathology, a person reliability coefficient may help determine how meaningful it is to interpret single critical items for a given individual.

Across-Session Person Reliabiltiy

Person reliability associated with responses to psychological test items has been conceptualized, basically, in two ways. First, person reliability has been defined as the tendency to respond consistently to an identical stimulus presented at two points in time (Fiske &

Rice, 1955). This kind of consistency will be referred to as across-session person reliability. An individual is asked to respond "true" or "false" to a particular item (e.g., "I enjoy being with people") at Time One and to respond again at Time Two. The responses are considered consistent if, for example, "true" or if "false" is answered on both occasions; the responses are considered inconsistent if "true" is answered on one occasion and "false" is answered on the other occasion. A variety of across-session person reliability indices have developed out of this conceptualization of consistency. They are based on item consistency, profile stability or response variability indices. Across-session item consistency measures are based either on the percentage of items or on a simple count of items answered identically at the two testing times. An across-session profile stability measure involves correlating scale scores obtained from the same test inventory at Time One and at Time Two. Such an index can be used to evaluate the consistency of an individual's profile of scale scores over two (or even more) occasions. Finally, across-session person reliability indices can be based on a measure of the variability in an individual's responses. One across-session response variability index is the ratio of the average within-item variance (over occasions) to the total variance of a person's responses to a scale's ~~items~~ over occasions (Bentler, 1964); a similar index is based on the sum of squares of scores on a number of tests subtracted from an individual's mean score on all tests across occasions (Berdie, 1961, 1969a, 1969b).

Within-Session Person Reliability

The second conceptualization of person consistency is the tendency to respond consistently to an identical or a similar stimulus presented in the same testing session. This kind of consistency will be referred to as within-session person reliability. Various within-session person reliability indices based on single session data have been calculated, using responses to repeated or "psychologically equivalent" items, scores on scales measuring consistency and within-session profile stability. First, within-session person reliability can be measured by examining the consistency of responses to items which have been repeated within a given psychological inventory (e.g., the TR index based on the 16 duplicated items on the MMPI). Similarly, responses to "psychologically equivalent" items can be compared to calculate a person reliability index (e.g., Greene, 1978; Raine & Hills, 1959). Another approach to within-session person reliability is through scales designed to measure consistency. Some of these scales (e.g., Chance, 1955; Mills, 1954; Pepper, 1964; Schofield, 1950) are made up of items empirically selected from existing inventories to discriminate between consistent and variable individuals; other scales are specifically designed to identify random responders (e.g., the Infrequency Scale on the Personality Research Form, Jackson, 1974) or careless responders (e.g., the F-scale on the MMPI). Finally, within-session person reliability indices can be calculated from single session item response data by evaluating the degree to which a respondent generates a similar profile of personality test scores, when different sets of items comprise the scales. Consider a multi-scale personality inventory having 12 20-item scales. By dividing the items from each of the 12 scales into parallel sets of 10 items each, a single respondent's responses could be used to generate two profiles, each based on one of the parallel sets. The similarity in shape between the two profiles might be termed within-session profile stability, evaluated by correlating the profiles, with N being the number of scales. For random responses, the expected value of this within-session person reliability index is 0.00 (Jackson, 1977). The range of values for such individual reliability coefficients has been assessed empirically (Jackson, 1976).

The Reliability of Person Reliability Indices

In an effort to evaluate person reliability indices, diverse studies have examined the reliability and generalizability of person consistency indices. In general, person reliabilty indices have shown at least moderate reliabilities. Across-session item consistency indices have shown adequate internal consistency, ranging from .34 to .70, when calculated on several interest measures and on an adjective checklist (Mitra & Fiske, 1956). When calculated on structured measures having dichotomous response alternatives or on relatively unstructured tests (Fiske, 1957b), these indices have ranged from .46 to .86. Further, item consistency indices have shown test-retest stability (Glaser, 1952). The stability of individuals' profiles across testing sessions has been found to be high (Holden, Helmes, Fekken & Jackson, 1981; Layton, 1954; Mauger, 1972), as has the reliability of response variability indices, calculated either across sessions (Berdie, 1961, 1969a) or across scales within a testing session (Bentler, 1964). Goldberg (1978; Goldberg & Rust, 1964) has demonstrated that both the internal consistency and test-retest stability coefficients for within-session person reliability indices based on items duplicated within a session could be as high as such reliability coefficients for the MMPI clinical scales if the indices were based on sufficiently large numbers of items. Goldberg and Jones (1969) have also reported that the internal consistency of scales based on items empirically selected from an existing inventory such as the MMPI (see Mills, 1954; Pepper, 1964; Schofield, 1950) can range from .85 to .93. Test-retest coefficients for such scales are somewhat lower, about .5 (Mills, 1954). The test-retest stability of scales designed to detect random responding, such as the PRF Infrequency Scale or the

MMPI F-scale, is also reported to be about .5 (Bentler, 1964; Hathaway & McKinley, 1967). Finally, within-session profile stability coefficients calculated on a multi-scale inventory measuring psychopathology have demonstrated at least modest homogeneity and test-retest stability (Holden et al., 1981).

<u>Generalizability of Person Reliability Indices</u>

Obviously, reliable indices of within-session and across-session person reliability can be calculated. However, the conceptual significance of these person reliability indices needs to be examined. One approach to this task has been to examine empirical relationships between various indices of person reliability. Evidence on the generalizability of these indices of person reliability is then interpreted as support for or against the hypothesis that person reliability is a meaningful and useful construct. As the evidence is summarized, it should become clear that a surface evaluation of the data does not allow definitive conclusions about the nature of person reliability to be drawn.

Much of the study of the generalizability of consistency indices has involved using a within-session person reliability index (calculated on single session data) to predict an across-session person reliability index. For example, Raine and Hills (1959) found empirical support (r = .55) for a relationship between a within-session consistency index based on "psychologically equivalent items" and an across-session profile stability index. Relatively strong correlations have also been found between a variety of within-session reliability coefficients and across-session person reliability coefficients on the MMPI (Schubert, 1975) and on other measures of psychopathology (Holden, Helmes, Fekken & Jackson, 1981). Finally, three studies have shown that the consistency of individuals' judgements (based on a circular triad score) can significantly predict across-session stability measures (Ace, 1969; Hendel & Weiss, 1970; Weksel & Ware, 1967).

Various scales measuring within-session consistency have been found predictive of across-session stability. For example, Bentler (1964) found that the PRF Infrequency scale correlated (r = .45) with across-session stability, suggesting that the within-session random responding aspect of response inconsistency is related to across-session stability. In an extensive study, Goldberg (1978) examined six scales (Mills, 1954; Pepper, 1964; Schofield, 1950), based on MMPI items empirically selected to predict response changes over two MMPI administrations. Goldberg found that all but one of these scales had a significant correlation with MMPI test-retest stability. Validity coefficients ranged from .61 to .14. These findings support the hypothesis that a person reliability index based on responses to subsets of items (i.e., scales) may be generalized to across-session person consistency.

Finally, the generalizability of person reliability indices has been studied in contexts other than an examination of the convergence of across- and within-session consistency. Consistency indices based on items duplicated within a session (i.e., the TR index) apparently can predict the validity of MMPI profiles as indicated by the F-scale (Greene, 1978, 1979). Further, within-session indices of person consistency calculated across trials on different behaviors (i.e., nonrating measures) show a strong relationship to each other (Schneiderman, 1980).

However, not all research findings on the generalizability of person reliability have supported the hypothesis that person reliability indices are empirically related. Some research examining the relationship of within-session response variability measures either across trials (Berdie, 1969a, 1969b; Fiske, 1957b; Fiske & Rice, 1955; Whitely, 1978) or across measures (Glaser, 1949) has yielded low correlations. Indeed, in one review article (Fiske and Rice, 1955) an average correlation of .2 between within-session response variability measures across trials is reported. Similarly, others have noted (Berdie, 1969a; Whitely, 1978) that while correlations between within-session response variability indices can have a wide range (e.g., .00 to .50; Berdie, 1969a), on the average, such correlations tend to be low. As well, Glaser (1949) found low correlations between across-session item consistency indices based on different inventories, leading him to hypothesize that person reliability indices are probably not generalizable.

Clearly, the empirical evidence for the generalizability of person reliability indices as support for the person reliability construct is somewhat difficult to interpret. Should the modest correlations between consistency indices be interpreted as support for the generalizability of person reliability or should these correlations be interpreted as failure to support the generalizability of person reliability? The data are hard to explain when a unidimensional view of the person reliability construct is maintained. However, the interpretation becomes clearer when person reliability is conceptualized as a multidimensional construct. In this view, person reliability is seen as comprised of different facets, which all tend to be related. Different indices of person reliability sample different facets of the construct and would not necessarily be expected to correlate highly. Consider the different implications of within-session and across-session person reliability. In order to obtain a high degree of within-session person reliability, responses to individual items need to be related in some consistent fashion. In particular, items which are exemplars of the same construct need to be answered lawfully if they are to have some organized relationship to each other. For example, when items repeated within a single test are used to measure within-session consistency, identical items must be answered identically. Similarly, when within-session reliability is assessed using a within-session profile stability coefficient, high person reliability will be obtained when responses to items which are all exemplars of the same homogeneous construct are related in some lawful way.

For a high degree of across-session person reliability, items need only to be answered identically at Time One and at Time Two. High within-session person reliability is not a necessary condition for high across-session consistency, although random responding either at Time One or at Time Two or at both times would yield no across-session person reliability. Still, within-session person reliability would seem to contribute to high across-session reliability: when item responses are lawfully related within a session (e.g., in terms of content, social desirability, acquiescence, etc.), use of the same response strategy on another occasion should contribute to both types of consistency. Nonetheless, high across-session consistency could still be obtained even when responses to items which are all exemplars of the same trait are not answered lawfully. Theoretically, item responses only need to be identical at Time One and at Time

Two. Indeed, empirical support exists for the hypothesis that within-session and across-session consistency form two relatively independent facets of person reliability (Holden, Helmes, Fekken, & Jackson, 1981). The view of person reliability as a multidimensional construct might help explain the seemingly surprising findings on the generalizability of person reliability. Different indices of person reliability sample different facets of the construct which are not necessarily strongly related.

CHAPTER II

VARIABLES INFLUENCING CONSISTENT RESPONDING

The literature review on person reliability thus far has suggested that reliable indices of consistent responding indeed can be constructed. Further, these indices do appear to have at least moderate generalizability, although the empirical evidence would suggest that different indices of person reliability might be viewed best as multiple facets of the person reliability construct. Thus, attempts to move beyond descriptions of consistency using various indices and to study the process of consistent responding would appear to be complicated by the multidimensional nature of the person reliability construct. Seeking to clarify the process of consistent responding through a systematic examination of variables influencing consistency would presumably yield findings dependent on and complicated by the exact indices used. One way to minimize such complexity would be to study the process of consistent responding in the context of a single definition of consistency. An argument can be made that the simplest definition of consistency would be the most appropriate to use. That is, person reliability should be studied by focussing on the stability of responses to particular items. The prediction of the stability of individual item responses surely has practical implications. Predicting the stability of particular responses gives a more refined index of person reliability than the overall within- or across-session indices. For example, such information might improve the evaluation of the appropriateness of individual item interpretation or the degree of domain articulation. Further, knowledge of the stability of item responses for an individual could be used to construct a tailored test having a minimum degree of stability associated with item responses. Indeed, knowledge of the stability of responses associated with items across individuals could be used to construct generally more reliable tests.

As well as practical implications, the prediction of the stability of responses to particular items also has theoretical implications. Predicting individual responses would seem by definition to require an examination of the actual factors which might influence response selection and, hence, the stability of individual responses.

Previous studies have attempted to predict the stability of individual responses to test items by considering variables such as aspects of the test-taking situation, items or persons either independently or simultaneously. Characteristics of the test-taking situation which have been examined are instructions for responding, type and number of response categories, item format and scoring and the interval between two testing sessions. Characteristics of items which have been studied include endorsement probability, social desirability and item content. Also, attempts have been made to relate person characteristics, such as scores on psychological inventories, to across-session item consistency. Studies which have taken into account both the characteristics of the item and the person have shown that the precision of measurement can be increased by considering these variables simultaneously. A review of the empirical evidence will clearly indicate that consistent responding is lawfully related to certain identifiable factors.

Effects of the Test-Taking Situation on Consistency

Various sources of evidence support the hypothesis that characteristics of the test-taking situation can influence person reliability. For example, when the similarity of two tasks (i.e., test and retest) is increased by familiarizing subjects with the items through prior item response (Schubert, 1975; Schubert & Fiske, 1973), across-session consistency is increased. Conversely, such consistency decreases when tasks are made less similar, such as by changing item format, item length (Fiske, 1957b) or scoring method (Ace, 1969). Indeed, a factor analysis of both within- and between-session person reliability indices collected on nine tests across eleven occasions suggested that much of the common variance is associated with the form of response required by the test (Fiske, 1957b). Finally, data collected on varying interval lengths suggest that the time interval between testing sessions also affects across-session item consistency. The percentage of responses changed on retest may be largest for short intervals but will stabilize around a particular value after a certain retest interval has been reached (Benton & Stone, 1937), even though test-retest scale reliability coefficients tend to be slightly higher for shorter rather than for longer intervals (Neprash, 1936; Pintner & Forlano, 1938).

Effects of Item Characteristics on Consistency

Closely related to the study of characteristics of the situation is the study of the relationship of the characteristics of items to consistent responding. In one sense, properties of items also constitute aspects of the testing situation. The item properties which have been examined in relation to across-session item consistency most often are the endorsement probability (or p-value) and social desirability of items. Early researchers (Lentz, 1934; Neprash, 1936) noted that items with "low incidence" were very unlikely to be answered inconsistently on retest. It has since been demonstrated (Goldberg, 1963; Payne, 1974) that items with extreme p-values are less likely to be altered than items with moderate p-values. Further, inconsistency in responding to an item appears to be a function of a group's distribution of responses to that item (Goldberg, 1963; Mitra & Fiske, 1956). Items with a small distribution of responses are less likely to be altered on retest than items with a large distribution of responses. These findings indicate that the stability of responses to items is related to endorsement properties. As well as endorsement probability, the social desirability scale value of items has been considered in relation to item consistency. It has long been noted that items with extremely "positive" or "negative" connotations were unlikely to be changed on retest (Frank, 1936). Payne (1974) demonstrated that items extreme in social desirability were less likely to be answered inconsistently than items moderate in social desirability. Payne theorized that items extreme in social desirability and/or having extreme p-values structure the situation for respondents. When social desirability and p-value are moderate, the situation is ambiguous and the probability of inconsistent responding is increased. A study by Bentler (1964) yields support for Payne's hypothesis. Initially, Bentler hypothesized that the social desirability of an item would be related to item consistency such that retest changes would occur in the socially desirable direction. Bentler, however, did not find such a relationship between across-session person reliability and social desirability even when social desirability was made salient or anchored on the first occasion. This finding was explained by noting that the personality inventory used was comprised of items relatively neutral in desirability. Restricting the range of items' social desirability may well have attenuated the relationship between across-session item consistency and desirability.

Payne's theory minimized the role of the content of items. While radical empiricists agree (Horn, 1950), research has shown that the content of items does influence the consistency of responding. Early work suggested the items referring to "factual data" such as age and sex were found least likely to be changed while "subjective personal" items such as "Do you daydream a lot?" were found most likely to be changed (Bain, 1931; Smith, 1933). Modern personality inventories, of course, consist of items of the "subjective" type and attempts to uncover the content of items which are most and least variable have met with some success. In an extensive study by Goldberg (1978), the content of items comprising six scales which were empirically derived from the MMPI to predict MMPI across-session profile stability was examined. There was some suggestion that items having desirable content were least likely to be altered while items which appeared to reflect moodiness, anxiety and distrust were most likely to be altered. Goldberg, however, was unable to draw any firm conclusions regarding the content of these scales. This is not surprising since the scales he studied all were empirically derived from the MMPI item pool and, thus, cannot be expected to reflect a homogeneous construct.

Effects of Characteristics of the Person on Consistency

A somewhat different procedure for relating the content of items to the across-session consistency of item responses has been to correlate scores on a scale with person reliability indices. This actually shifts the focus from characteristics of the item to characteristics of the person. Research in which the number of items changed on retest was correlated with scale scores based on the same instrument has uncovered a general trend for more consistent responders to obtain scores indicative of "better adjustment" (Eisenberg & Wesman, 1941; Lentz, 1934; Neprash, 1936; Pintner & Forlano, 1938). Indeed, the hypothesis that individuals who scored as more neurotic and emotionally variable were less consistent in responding has received support from an extensive review which examined over 50 samples of individuals who were tested twice on a variety of structured inventories. Windle's (1954, 1955) main finding was that on scales measuring dimensions of psychopathology, scores tended to move in the direction of adjustment, particularly if the retest interval was short. However, strong arguments have been made to refute the conclusion that better adjusted individuals tend to change fewer responses in a test-retest situation (Bentler, 1964; Glaser, 1949, 1952; Whitely, 1978). Glaser has argued that a spurious mathematical relationship exists between total scale score and some measure of person reliability based on the same psychological inventory. Glaser provides empirical support for his reasoning (1949, 1951, 1952), concluding that any relationship between across-session item consistency and the content of a set of items is merely measurement artifact. Thus, this line of research has not proven very successful for relating person reliability and characteristics of the person.

Response Processes as Characteristics of the Person

Another point of view, however, argues that characteristics of the person are indeed relevant to consistent responding. This approach to relating person characteristics to person reliability has been through a series of studies which considered the relationship between response strategies and across-session item consistency. Rather than looking for personality correlates of person reliability indices, response processes are analyzed in terms of their relationship to the consistency of response. In the broadest sense, response processes refer to what takes place between the presentation of an item and the indication of a response. Lumsden (1977, 1978) has taken this general notion of response processes and has hypothesized that all test unreliability is due to the individual. In this theory, tests and items are both perfectly reliable. Any inconsistency within or across testing sessions is due to random fluctuations or "tremors" associated with the individual. Other researchers have examined general response strategies associated with answering items. The evidence suggests that an individual will be more likely to answer items consistently on retest if responses are selected to reflect a stable self-image (Johnson, 1981), if items can easily be compared to relevant information stored in memory (Rogers, 1974), or if the individual is aware of and attentive to his or her own internal states (Underwood & Moore, 1981). Such evidence suggests that individual differences in general response strategies can help account for differences in across-session item consistency.

Response strategies specifically related to responding to psychological test items have also been used to predict the stability of responses to items on retest. In particular, assessing the appropriateness of an individual's response selection has shown promise for predicting the stability of item responses on retest. Appropriate responding requires that the individual, in responding to an item, interpret the item in a particular way. Appropriate responding might depend on interpreting items in the same manner as one's peers or in accordance with the intentions of the item writer. To elicit response processes, individuals can simply be asked to write out their interpretation of each item. The interpretations of each item are then rated for similarity across individuals (Eisenberg, 1941; Eisenberg & Wesman, 1941). Alternatively, response processes can be elicited by having individuals describe their item responses in relation to specific factors, such as difficulty in selecting a response, applicability and meaningfulness of the item, influence of recent events and introduction of new elements. It has been repeatedly demonstrated that across-session item consistency indices of person reliability are negatively related to the use of inappropriate response categories (Eisenberg, 1941; Kuncel, 1973, 1977; Kuncel & Fiske, 1974; Turner & Fiske, 1968). Thus, individuals who characteristically evaluate items inappropriately are more likely to change their responses on retest. Although an individual's tendency to use inappropriate response processes appears to be stable across trials (Kuncel, 1973), apparently there is no reliability to the irrelevant aspect of the item on which attention is focussed or to the irrelevant element which might be introduced in interpreting the item.

CHAPTER III

PERSON AND ITEM CHARACTERISTICS IN INTERACTION

Clearly, across-session item consistency is a function of the variables associated with the test-taking situation and with the icem. In addition, variables associated with the person, in terms of either general or specific response processes, show a relationship to the stability of individual items. Predicting the stability of individual item responses by taking into account relevant variables simultaneously would seem to have the advantage of increasing the precision of measurement and hence, improving prediction. Two strategies characterize the interaction approach to predicting across-session item stability. These strategies reflect psychologists' increasing interest in identifying the theoretical mechanisms which might underlie responding to items (Embretson, 1983).

One strategy has attempted to predict item stability by defining person-item interactions using procedures borrowed from item response theory (cf. Hambleton & van der Linden, 1982) to scale persons and items. The other strategy has studied person-item interaction by examining individuals' response latencies to specific items. In the first group of studies, persons and items were scaled using a Rasch model (Rasch, 1960), and the distances among persons and items were defined in terms of the probability of item endorsement. The Rasch model involves scaling subjects and items on the same continuum. Very simply, subjects are ordered according to the total number of items they have endorsed while items are ordered in terms of frequency of endorsement. When the continua are superimposed, a certain orderliness is expected to emerge, with subjects endorsing items in order of p-value. The point where the subject stops endorsing items is called the threshold and theoretically all items with p-values above the subject's threshold will be endorsed while all items below will be rejected. The Rasch model is a log linear test model used to estimate subject and item scale values. The advantages of the Rasch model are that it enables comparisons to be made among individuals which are not dependent on the instrument; and it functions independently of what was measured so that individuals' results (i.e., scale values) are not simply a function of the group the individual happened to be in (Rasch, 1960). Empirical results of these studies have demonstrated that person-item distance is negatively related to across-session item consistency (Fiske, 1968; Kuncel, 1973, 1977; Kuncel & Fiske, 1974). These findings are interpreted as follows. When an item is far from a person on a continuum, the decision to endorse or to reject is relatively easy. However, when an item is near a person, the decision is relatively more difficult. Less across-session item consistency is expected to be associated with difficult than with easy decisions. As further support for these ideas, Kuncel (1973) shows that the probability of the emergence of "inappropriate" response processes varies inversely with the distance between items and persons. This research points to the value of examining person characteristics as well as item characteristics in determining which items are likely to be unstable. Although item stability is still related to the mean distance of items from a group of people (Fiske, 1968; Glaser, 1949) and to the extremity of endorsement for a group (Kuncel, 1977), item stability is also a function both of characteristics of the item and of the person.

The second strategy for predicting item response stability by taking into account a person-item interaction has been through the examination of reaction times or response latencies. Response latency is simply the amount of time an individual takes to make a particular response to a given stimulus. Research in cognitive psychology has shown that response latencies are a function of such obvious variables as stimulus intensity, stimulus complexity, the number of response choices and the confidence level associated with the judgements (Brebner & Welford, 1980). Similarly, personality psychologists have attempted to show the relationship of reaction time to variables relevant to the personality domain, especially the characteristics of structured personality inventory items. Individuals apparently take longer to respond to items which tend to be long, ambiguous, controversial and have a wide dispersion of social desirability ratings associated with them (Dunn, Lushene, & O'Neil, 1972; Goldberg, 1963; Hanley, 1962; Rogers, 1973a, 1973b). However, in addition to considering response time a function of item properties, response time can itself be an index of the difficulty of an item for a particular person. This individual differences approach to response latencies forwards the hypothesis that items with shorter latencies for an individual may reflect easier or more confident decisions. Such an hypothesis is supported by Kuncel's (1973) finding that the latencies for responding to items scaled near an individual's threshold derived from the Rasch model tended to be longer than for distant items, presumably because the decision was more difficult. Other empirical studies have shown moderate self-ratings (e.g., on adjectives) involved longer response latencies (Rogers, 1973b, 1974; Rogers, Kuiper, & Rogers, 1979) than extreme self-ratings. These data also suggest certain decisions, particularly for "close" items, are relatively more difficult for the individual. ~~Indeed~~, Rogers et al. (1979) found that subjects' self-ratings on trait descriptors could predict virtually all of the variance in response latencies. Such findings clearly support the position that some items may be significantly and reliably more difficult for some individuals.

Four Models For Predicting Across-Session Item Consistency

Empirical studies investigating the influence of such variables as aspects of the test-taking situation, items or persons have yielded two general conclusions: one, responding is lawful; and two, the across-session item consistency can be influenced by manipulating such aspects. Empirical studies concerned with an interaction between aspects of items and persons has yielded promising results for the prediction of an individual's consistent responses to items (Eisenberg, 1941; Eisenberg & Wesman, 1941; Kuncel, 1973, 1977; Kuncel & Fiske, 1974).

The purpose of this research is to investigate the stability of individual item responses by evaluating four prediction models.

The four models are as follows:

1) Model PV, which is based on the p-values of items;

2) Model SD, which is based on the social desirability scale values of items;

3) Model TH, which is based on items' proximity to a

threshold for responding to items in terms of some item characteristic;

4) Model RL, which is based on persons' latencies for responding to individual items.

Two of these models are based on item characteristics alone while two models take into account both item and person characteristics. The two models based on item characteristics are included for comparison. Strong empirical evidence indicates that item characteristics are useful for predicting response stability; however, it is hypothesized that models based both on item and person characteristics will be better predictors.

Model PV

The first model, Model PV, is based on the endorsement properties of items. Previous research has clearly indicated that items with extreme endorsement properties are more likely to be answered consistently on retest than items with moderate endorsement properties (Frank, 1936; Goldberg, 1963; Lentz, 1934; Mitra & Fiske, 1956; Neprash, 1936; Payne, 1974). Model PV is based on the p-value of items, which is simply the proportion of "true" responses to a dichotomous item. Model PV predicts that individuals are more likely to change retest responses to items having moderate rather than extreme p-values.

Model SD

The second model, Model SD, is based on the tendency for an item to elicit a socially desirable response. The social desirability scale value of an item is simply its mean judged social desirability. Previous research has indicated that items moderate in social desirability are likely to be less stable than items with extreme social desirability scale values (Frank, 1936; Goldberg, 1963; Lentz, 1934; Neprash, 1936; Payne, 1974). Model SD predicts that individuals are more likely to change their responses on retest to items having moderate rather than extreme social desirability scale values.

Model TH

The third model, Model TH, is based on the threshold model for responding (Jackson, 1968, 1982). A threshold model of responding was originally proposed to describe a set of unified, hypothetical processes which might characterize responding, especially stylistic responding (Jackson, 1968). The model proposes that the responding pattern of any individual can be described in terms of a subject operating curve (see Figures 1, 2, & 3). This curve is a plot of the relationship between an item property and the endorsement probability for the particular respondent. Jackson (1982) has proposed that, conceptually, the subject operating curve is similar to the

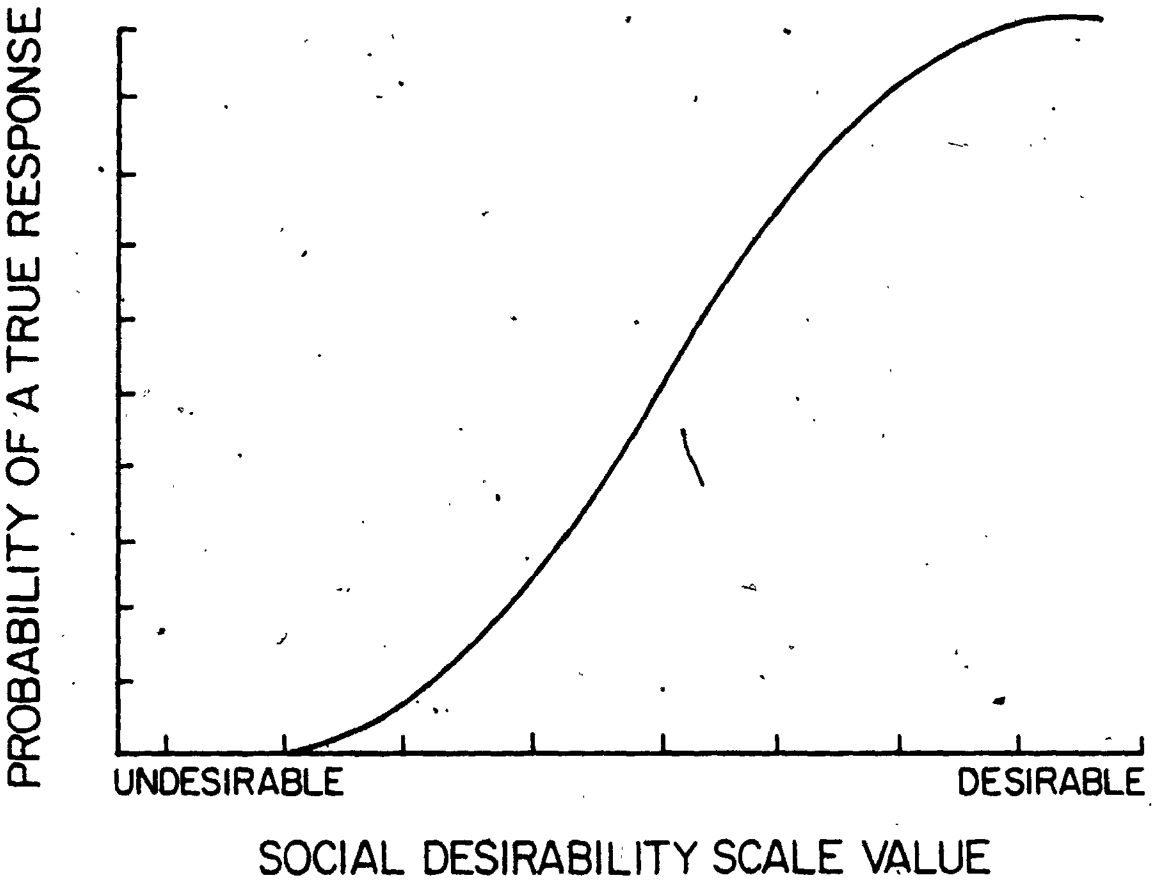

Figure 1. Subject operating curve described in the threshold model for responding

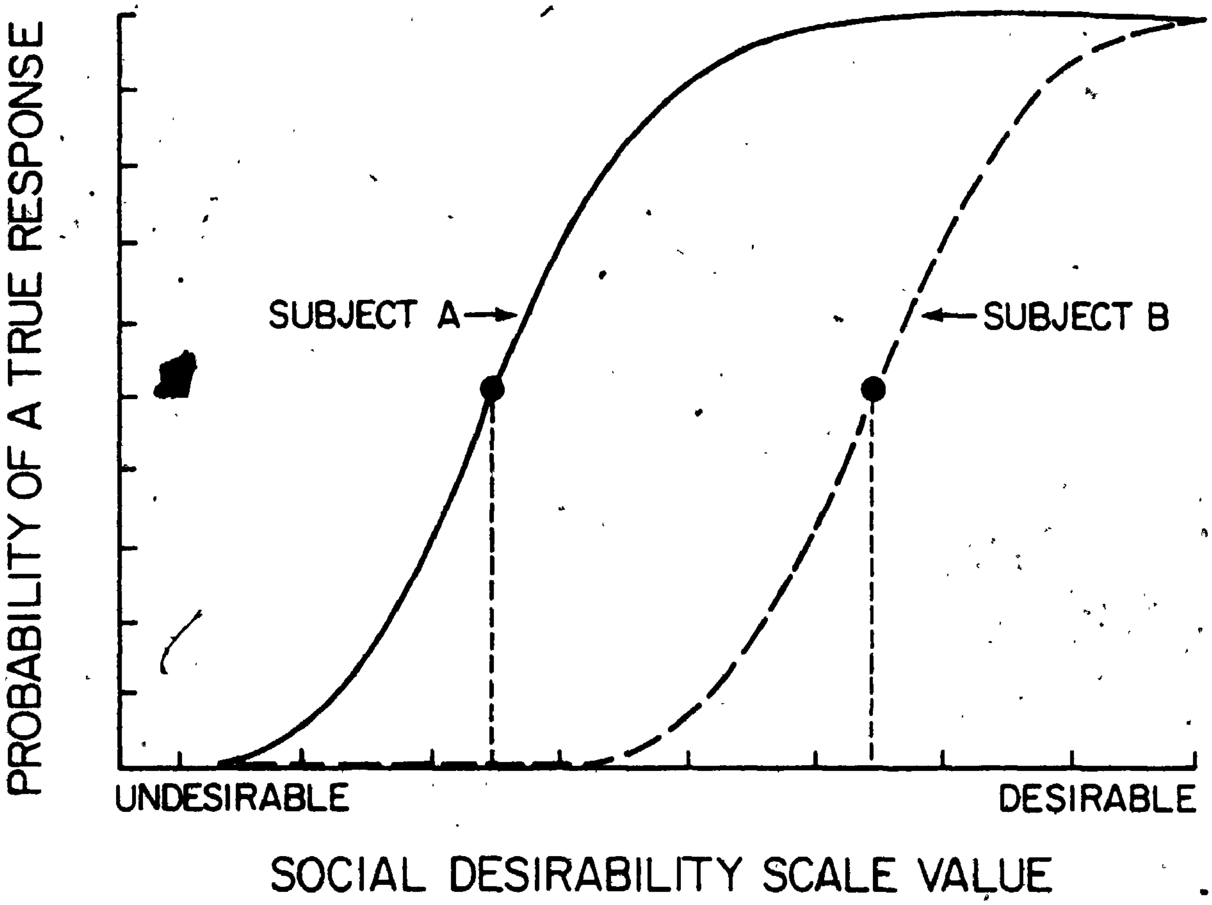

Figure 2. Subject operating curves with different thresholds

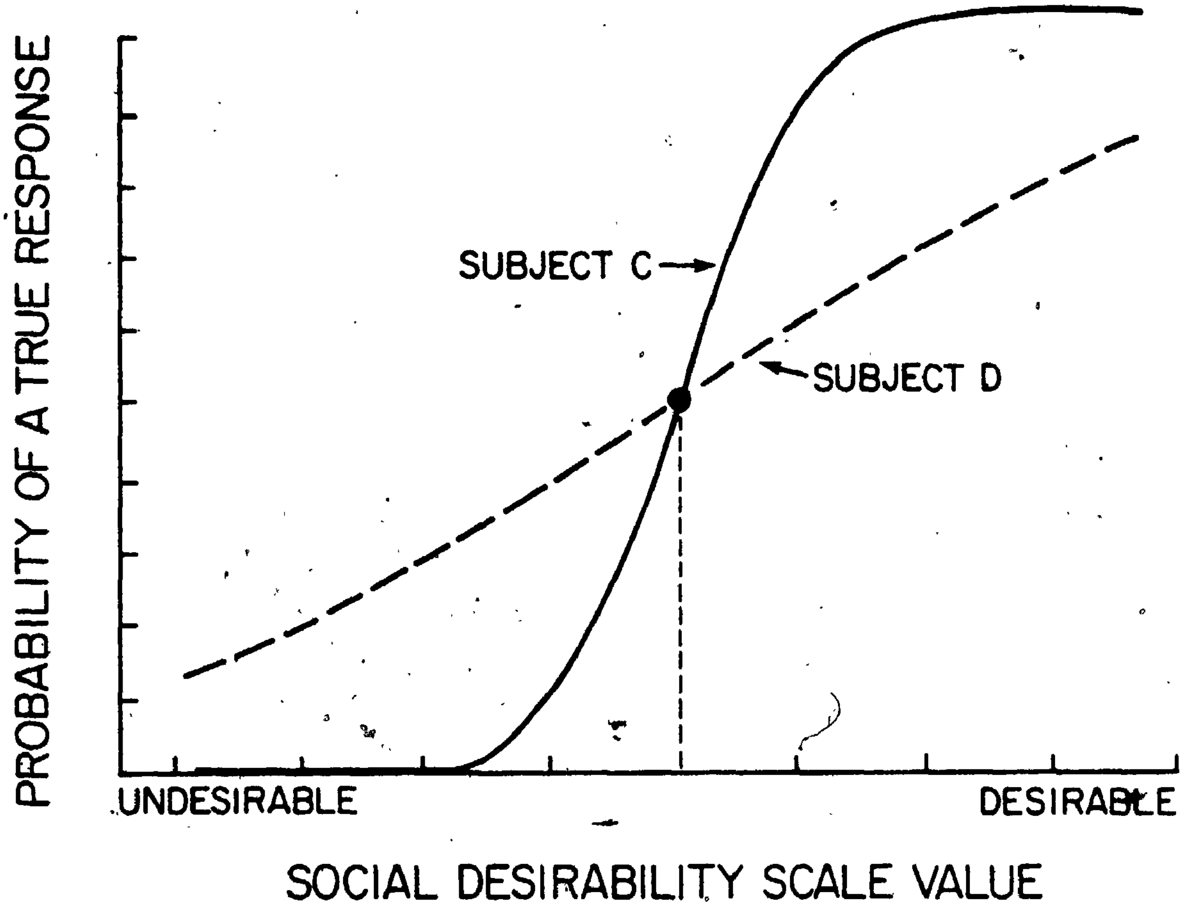

Figure 3. Subject operating curves with different saliences

item operating curve which relates an item and its properties to some underlying attribute. The subject operating curve is, in a sense, the "complement of the operating curve proposed for items" (Jackson, 1982, p.2) and describes the individual in terms of an underlying dimension using many item responses. The subject operating curve is considered to be a normal ogive. The curve and, hence, the response process, is described in term of two response parameters, threshold and salience. The mean of the normal ogive is considered to be the individual's threshold while the variance of the curve is considered to be the salience parameter. The threshold marks the transition from the tendency to reject items to the tendency to endorse items. The salience reflects the degree to which the item property (e.g., social desirability, if one is evaluating stylistic responding) is determining the individual's responses to items. Empirical investigations have demonstrated that when the item property evaluated is social desirability, individuals have different thresholds and saliences (Jackson, 1968) and that these parameters show test-retest stability (Rogers, 1970, 1971).

From this outline of threshold theory, it follows that when items are far from an individual's threshold (in terms of some item property), the probability of endorsing an item is close to either 0.0 or to 1.0. However, when items are near to a subject's threshold, the probability of endorsing an item deviates considerably from 0.0 or from 1.0, falling in the intermediate range. Thus, Model TH would predict that items falling near a subject's threshold should show maximum inconsistency on retest. As well, this effect should be most evident for individuals with high saliences.

There are a number of advantages to using the threshold model to conceptualize item-person interactions for use in the prediction of item stability. First, there is empirical research (Helmes, 1978) that suggests a threshold model of responding is indeed useful for predicting specific item responses. Next, the threshold model is like a Rasch model, because it also takes into account individual differences in the perception of item characteristics. However, the threshold model further describes this person-item interaction in terms of two relevant parameters, threshold and salience. Finally, the theory underlying the threshold model readily yields predictions concerning which items should be most difficult for an individual to respond to, and, hence, which items are most likely to be unstable on retest. Model TH predicts that individuals will be more likely to change their responses on retest to items which are near their individual thresholds rather than far from them. Thresholds will be determined for individuals' tendencies to respond to items in terms of social desirability. Thresholds will be based on desirability for the following reasons. The threshold model for responding was originally proposed to describe stylistic responding. Social desirability has been empirically related to the stability of individuals' item responses (e.g., Payne, 1974). Research on predicting item responses per se with the threshold model (Helmes, 1978) and with models based on multidimensional scaling techniques (Cliff, 1977; Cliff, Bradley & Girard, 1973; De Boeck, 1980) suggests social desirability is a stronger determinant of correct response prediction than item content or meaning.

Model RL

The fourth model, Model RL, is based on an individual's response latencies associated with specific item responses. Previous research has indicated that response latencies associated with individuals' responses may reflect the difficulty of an item for individuals (Dunn et al., 1972; Goldberg, 1963; Hanley, 1962; Kuncel, 1973; Rogers, 1974; Rogers et al., 1979). While some researchers have explained individual response latencies mainly as a function or item characteristics (Dunn et al., 1972; Hanley, 1962), many other formulations have explicitly focussed on individual differences in response latencies. These latter formulations have suggested that the response process involves comparing the item to the self's position on an underlying content dimension (Ebbesen & Allen, 1979; Kuiper, 1981; Rogers, 1974, 1978). The difficulty of this self-referent decision is negatively related to the distance between the item and the self and is reflected in the individual's latency for responding to the item. Thus, Model RL predicts that individuals will be more likely to change their responses on retest to items which have relatively long response latencies. Specifically, a given individual will be more likely to change his or her responses to items on which he or she took a relatively long time relative to his or her own other responses.

Model Evaluation

Three studies are proposed to evaluate the usefulness of the four models for predicting the stability of individual item responses. The first study will compare the models' ability to predict the stability of individual items reflecting normal personality content and relative neutrality in social desirability. Items were adopted from a set of personality scales developed using modern principles of test construction (Jackson, 1970, 1971) which tends to yield questionnaires with scales showing specific psychometric properties.

In the second study, the influence of item characteristics on the predictive success of the three models (Models PV, SD and TH) explicitly based on a single item property will be addressed. The three models' ability to predict the stability of individuals' item responses will be compared for items having a considerable range of social desirability scale values and p-values.

The third study will assess the role of decision difficulty. In this study, subjects' awareness of certain item properties will be experimentally manipulated using a rating task. Making certain item characteristics salient is expected to affect individuals' strategies for responding. Thus, individuals are expected to respond to items in terms of the salient item characteristic and response decisions are expected to be easier, since a response strategy has been suggested. Each of the four models' ability to predict an individual's item responses will be compared given these assumptions.

CHAPTER IV

STUDY 1 : INITIAL MODEL COMPARISON

Method

Subjects. Subjects were 40 (20 males, 20 females) introductory psychology students who received experimental credit for their participation. The average age of the subjects was 19.65 years (standard deviation, 2.79).

Materials. Ten content scales and two validity scales (Infrequency and Desirability) were included in the study. These were comprised of 192 items taken from Jackson's (1974) Personality Research Form (PRF). As well, all subjects completed the Extended Range Vocabulary Test developed by Educational Testing Service (French, 1962).

Procedure. Each subject was tested twice with approximately one week separating sessions. In the first session, all subjects responded to the 192 PRF items in a manner allowing the collection of responses latencies. Then subjects were asked to respond to the timed Extended Range Vocabulary Test. Finally, subjects were asked to provide general demographic information such as sex, age and Grade 13 average.

In the second session, 20 subjects were selected randomly within sex to respond to the 192 items so that response latencies could be collected while the other 20 subjects responded to the items using the standard paper and pencil format. For the latter group, the time taken to respond to the items was recorded.

Collecting Response Latencies. In order to collect response latencies, items were presented to subjects on a screen using a slide projector. The slide projector was electrically connected to a clock (timing in milliseconds) and to a two-key control panel. The experimenter controlled the projection of slides. As the slide fell into the projection slot, the clock was started. To indicate a "True" or a "False" response to an item, the subject used his or her index finger on the preferred hand to depress either the right-hand key marked "T" or the left-hand key marked "F" on the control panel. Depressing a key resulted in three simultaneous events. The clock stopped, one of two lights on the clock illuminated to indicate to the experimenter which key was depressed, and the next slide fell into the projection slot. To allow the experimenter time to record the response latency and the response, blank slides were alternated with slides of items in the slide tray. Thus, the experimenter always controlled item projection and each item response was always followed by a brief interval. Between trials, subjects rested their index fingers on a "home" spot on the control panel midway between the "T" and the "F" keys. The first 15 items to which all subjects responded were practice items taken from an earlier form of the PRF (Form AA). Data from these items were not included in any analyses.

Results

Properties of the Testing Material. The means and standard deviations for the PRF scale scores for the first and second testing sessions as well as the KR-20s are reported in Tables 1 and 2, respectively. Scale means, standard deviations and internal consistencies tend to be comparable to those published in the PRF manual (Jackson, 1974). The low mean and negative internal consistency for the Infrequency scale may be the result of little response variability associated with the items. Perhaps the highly supervised nature of the test-taking situation induced subjects to monitor carefully their responses to validity scale items. Test-retest correlations for the 12 PRF scales are reported in Table 3.

The mean reported Grade 13 average was 76.00 per cent, with a standard deviation of 8.03, while the mean score on the Extended Range Vocabulary Test was 26.78 correct responses (out of a possible 40), standard deviation, 6.83.

Properties of the Testing Conditions. All subjects were initially tested in a format allowing the collection

Table 1

Scale means, standard deviations and KR-20s for 12 PRF scales - initial testing session (N=40)

Scale	Mean	Standard Deviation	KR-20
Abasement	6.9	2.9	.65
Achievement	10.3	3.6	.78
Affiliation	9.5	3.9	.82
Aggression	7.2	3.6	.76
Autonomy	6.7	3.7	.78
Change	9.8	3.3	.72
Cognitive Str.	9.2	2.7	.60
Defendence	6.2	3.5	.77
Dominance	8.9	4.1	.85
Endurance	8.8	3.7	.80
Infrequency	.2	.4	-.21
Desirability	11.3	2.8	.64

Table 2

Scale means, standard deviations and KR-20s for 12 PRF scales - retest testing session (N=40)

Scale	Mean	Standard Deviation	KR-20
Abasement	6.9	3.0	.69
Achievement	10.3	4.0	.83
Affiliation	9.1	3.9	.83
Aggression	7.2	3.5	.76
Autonomy	7.1	3.9	.82
Change	9.2	3.6	.77
Cognitive Str.	8.1	3.2	.71
Defendence	6.0	3.7	.82
Dominance	8.2	4.4	.87
Endurance	8.6	4.5	.76
Infrequency	.2	.4	-.20
Desirability	11.4	3.2	.76

Table 3

One week test-retest correlations for 12 PRF scales

(N=40)

Scale	Test-retest Coefficient
Abasement	.78
Achievement	.83
Affiliation	.90
Aggression	.83
Autonomy	.86
Change	.91
Cognitive Structure	.77
Defendence	.75
Dominance	.89
Endurance	.82
Infrequency	.38
Social Desirability	.80

of response latencies. On retest, 20 subjects were retested in the same format so that the reliability of response latencies could be established. The other 20 subjects were retested using the standard paper-and-pencil format. The groups did not differ in terms of the total number of item changes made (t(38) = .90, n.s.). As well, there were no ~~initial~~ differences between the groups in terms of mean latency (t(38) = .81, n.s.).

Mean latency (mean response latency associated with all items across 40 individuals) was 4.7 seconds (standard deviation, 1.3 seconds). For the 20 subjects who were tested twice, mean latency dropped significantly (t(19)= 5.4, p< .0001) from the first testing session (mean 4.9 seconds; standard deviation, 1.4) to the second (mean 4.0 seconds; standard deviation, 1.0) testing session, although the test-retest correlation for mean latency was high (r = .87, p<.001). The test-retest correlations for each of the 20 subjects (based on 192 pairs of latencies) were all positive, ranging from r = .07 to r = .62, with a mean correlation of r = .43.

Mean response latency correlated significantly with the time taken to complete the vocabulary test (r(38) = .45, p<.001) and with the time taken to complete the standard retest on the 12 PRF scales (r(18) = .55, p<.05). Mean response latency did not appear to be significantly correlated with sex (r(38) = .11), age (r(38) = .02) or either verbal ability measure : Grade 13 (*r*(32) = -.27) or vocabulary test score (*r*(38) = -.12).

Response Stability. An unstable item response was defined as the change from a "True" to a "False" response on retest or vice versa. Subjects changed 15.26 per cent of their responses or a mean of 29.30 of 192 item responses with a standard deviation of 11.94 responses. The range of unstable responses was 11 to 84 responses. No items were omitted by any subject.

Model TH. Parameters for the threshold model (see Appendix A) were calculated using the social desirability scale values of the 192 PRF items as item properties (see Table 4) for the 40 subjects. Thresholds and saliences both showed adequate test-retest stability (*r* = .75 and *r* = .85, respectively, $p<.001$) and were only moderately intercorrelated for both test (*r* = .33, $p<.05$) and retest (*r* = .29, $p<.05$).

Initially, the relationship of individual item stability and threshold was examined on a very broad level. The threshold range was defined as the 25 per cent of items around the threshold, 12 1/2 per cent above, 12 1/2 per cent below. Comparing the proportion of items changed in the threshold range (mean .149) to the proportion of items outside the threshold range (mean .154) suggested there was no significant difference. However, since the threshold is a more useful parameter for describing the process of

7

Table 4

Parameters for the threshold model based
on 192 PRF items
(N=40)

		Mean	Standard Deviation
Threshold	-- test	5.46	1.38
	-- retest	5.31	1.53
Salience	-- test	.44	.04
	-- retest	.43	.06
Proportion of True Responses			
	-- test	.51	.01
	-- retest	.53	.01
Intercept	-- test	-.08	.10
	-- retest	-.04	.12
Slope	-- test	.10	.00
	-- retest	.10	.00

responding to items which are relatively neutral in desirability (Rogers, 1970), the analysis was repeated omitting the validity scale items, which tend to have somewhat higher SDSVs. Thus, for the ten content scales comprising 160 items, the proportion of changed items falling in the threshold range was compared to the proportion of changed ıtems falling outside this range. These mean proportions across 40 individuals were .184 (standard deviation, .10) and .163 (standard deviation, .07), respectively. These means were in the correct direction, although they were also not significantly different (_t_(39) = 1.54, n.s.). Further, the 40 subjects were divided into two groups based on a median split on salience. The difference in the proportion of items changed inside and outside of the threshold range was more pronounced for the high (.219 versus .174) than for the low salience group (.149 versus .154), although the differences were not statistically significant.

Model RL. The relationship of latency to stability was first examined by correlating mean latency and total number of unstable responses across individuals. The relationship was small and nonsignificant (_r_(38) = .11) and there was no evidence of a curvilinear relationship. However, the relationship of change and latency for responding to an item at the individual level _did_ suggest that response latency and change were inversely related. The correlations between unstable responses and response latencies for each of the 40 subjects were positive, ranging from .01 to .51, with a mean correlation of .22.

Models SD and PV. Social desirability scale values and p-values for the 192 PRF items were obtained from Helmes, Reed & Jackson (1977). Mean social desirability scale value was 5.20 (standard deviation, 1.84) and mean p-value was .51 (standard deviation, .05). The two sets of item properties correlated $\underline{r}(190) = .75$, $p<.001$. Initially, the relationship of item properties and stability was evaluated at a simple level. Items were divided into two equal-sized groups, those items having relatively extreme social desirability scale values and those items having relatively moderate values. Note that the range of the social desirability scale values of PRF items is not great relative to other psychological inventories. When the mean number of item responses individuals changed to items moderate in desirability (16.5 changes, standard deviation 7.3) and extreme in desirability (12.8 changes, standard deviation 6.0) were compared, significantly more changes were made to moderate than to extreme items ($\underline{t}(39) = 4.03$, $p<.001$). To examine the influence of p-value, items were again divided into two equal-sized groups, those items having relatively extreme p-values (either high or low) and those items having relatively moderate p-values. P-values derived from the independent sample (Helmes, Reed, & Jackson, 1977) were strongly similar to those derived from the present sample (_r_(190) = .90, p<.001). When the mean number of item responses individuals changed to items having moderate p-values (16.6 changes, standard deviation 7.3) and to items having extreme p-values were compared (12.8 changes, standard deviation 5.9), significantly more changes were made to items moderate than extreme in p-value (_t_(39) = 4.23, p< .001).

Model Comparison. The predictive accuracy of the four models was compared as follows. Given the total number of item responses an individual changed on retest, each model was required to predict exactly which items the individual changed. For example, assume an individual changed 25 item responses on retest. Model TH would predict that these items are the 25 items nearest the threshold. Model RL would predict the 25 items with the highest latencies would be changed. Models SD and PV, based on item parameters, would predict the 25 items with the most moderate social desirability scale values or with the most moderate p-values would be changed. The number of correct designations for each model was calculated. For each pair of models, the proportion of identical target items was calculated across subjects. These proportions were as follows: Models PV and SD, .13; Models PV and TH, .14; Models PV and RL, .18; Models TH and SD, .22; Models TH and RL, .18; Models RL and SD, .18.

Table 5

Mean number of correct predictions for each of the four models (N=40)

	Correct Predictions	
	Mean	Standard Deviation
Model PV	5.48	6.27
Model SD	6.00*	5.76
Model RL	7.95**	6.31
Model TH	5.78*	5.92
Chance	5.05	5.57
Mean number of item changes	26.85	10.68

** Significantly different from chance at p< .001.

* Significantly different from chance at p< .05.

The number of correct predictions for each model was compared across subjects using a repeated measures analysis of variance with models as the factor. For the purposes of this analysis, only the 160 items from the 10 PRF content scales were used. When models were compared, there was a highly significant effect for models in the predictability of response changes [F(3,117) = 13.07, p< .001]. Using Dunn's Multiple Comparison Test, Model RL was found to be significantly better than Models SD, PV, and TH, none of which differed significantly from one another. A series of t-tests were performed to compare each model to the number of correct predictions made by chance (see Appendix A). Model SD (t(39) = 3.06, p< .05), Model TH (t(39) = 2.05, p< .05) and Model RL (t(39) = 8.11, p< .001) performed significantly better than chance while Model PV (t(39) = 1.65, n.s.) did not (see Table 5). As well, when the predictive accuracy of Models TH and RL was compared to that of Models PV and SD using a t-test, Models RL and TH were found to be significantly better predictors (t(39) = 3.40, p<.002).

Discussion

The results of this study, like those of previous studies (Benton & Stone, 1937; Neprash, 1936; Schofield, 1950), suggest that item responses tend to be stable over a one-week retest interval. Only about 15 per cent of responses were changed on average by the 40 subjects in this study. More importantly, the results of this study suggested that it may be possible to identify those items most likely to be changed. Certain item characteristics as well as item and person characteristics in interaction appear to be related to item instability. The broad analyses conducted to examine the general role of social desirability, p-value, threshold and response latency provide support for this hypothesis. Individuals are more likely to change items which are relatively moderate in social desirability and items which are relatively moderate in p-value. Perhaps moderate items present an ambiguous situation for the respondent (Goldberg, 1963; Payne, 1974). Consider, however, that item characteristics and person characteristics in interaction are also related to the stability of response. Simple analyses suggested that response latencies are inversely related to the stability of item responses on retest. There was also some modest suggestion that items near an individual's threshold, which takes into account individual differences in the tendency to respond desirably, were less stable. These data also indicate that when characeristics of persons and items are simultaneously considered, such characteristics are also related to item response stability.

In addition to simple analyses regarding the hypothesis that each model would be able to predict stable items, stricter analyses required each model to predict exactly which items an individual would change on retest.

When the correct number of predictions made by each model was compared to chance, Models SD, TH and RL were significantly better than chance. Model PV did not predict above chance which items would be changed on retest. Perhaps an actual consideration of the group endorsement probabilities associated with items is not a part of the process of responding consistently. While p-value is certainly related to consistency, it may be that this relationship is moderated by other item characteristics, including content and desirability. On the other hand, the p-values of the 160 PRF items used in this study were all relatively moderate. Originally, items were selected for the PRF if, on pretesting, they showed p-values greater than .2 and less than .8 (Jackson, 1974). It may be that the small range of p-values restricted the relationship of p-value to consistency.

An examination of the proportion of identical target items predicted by pairs of models suggested that models do predict somewhat overlapping distributions of items. No doubt these proportions reflect the empirical relationships of the variables underlying the four models. However, the degree of overlap would appear sufficiently modest to consider the models reasonably empirically distinct. When models were compared in an analysis of variance with repeated measures, a strong, significant main effect for models was obtained, suggesting that some models were indeed better predictors than others. A multiple comparison test indicated Model RL was significantly better than the other models which did not differ from each other. Thus, response latencies appear to be significantly more accurate for predicting precisely which items an individual will change on retest than either SDSV, p-value or proximity to an individual's threshold. It had been predicted that Model RL, like Model TH, would perform significantly better than Models SD and PV because it took into account both aspects of items and persons. Consider how response latency takes into account these aspects and reflects the difficulty of the item for the individual. For example, consider that an individual assesses an item in terms of a particular response property and then compares that item to his or her own position on the same dimension. If the item is far from this position, the decision to endorse or to reject is easy and response latency is short. However, if the item is near his or her position on the dimension, a fine discrimination must be made. To make this fine discrimination, the individual requires several comparisons of the item and position. Hence, response latency is long. Across individuals, it is expected that decisions on extreme items will tend to be easier. Judgements of item characteristics show reliability across individuals and, if a normal distribution of individuals along the dimension is assumed, extreme items will generally be further from individuals and hence, more stable (cf. Cliff, 1977; Cliff, Bradley,

& Girard, 1973; Ebbesen & Allen, 1979; Rogers, 1973b). This may explain why, in general, items extreme in social desirability, and perhaps in endorsement properties, are more stable than moderate items.

The threshold model, which takes individual differences in the tendency to respond desirably into account, was expected to perform as well as Model RL and better than Models SD and PV. Model TH could readily predict above chance which items individuals changed on retest. However, Model TH performed only as well as Models SD and PV and significantly worse than Model RL. Various explanations might be suggested. Model RL has the advantage that individuals can evaluate the items in terms of any item property and the model still remains valid. However, Model TH is based on only one item characteristic, namely, social desirability. While it is true that this characteristic has previously been shown to be quite useful for predicting item responses (Cliff, 1977; Cliff, Bradley, & Girard, 1973; Helmes, 1978), nonetheless, Model TH does take only this one characteristic explicitly into account. It is assumed that items are being evaluated in terms of desirability, but given that the items chosen for this study are relatively neutral in social desirability, perhaps a variety of other item properties become more salient. When predicting individual item response changes using a model based on a single item property, it becomes critical to specify that property correctly if the model is to be reasonably evaluated. Thus, in Studies 2 and 3, Model TH will be evaluated under conditions where the social desirability of items is made especially salient. In Study 2, a group of items ranging greatly in social desirability will be used while in Study 3, social desirability will be made salient using a rating task.

CHAPTER V

STUDY 2 : MODEL COMPARISON EMPHASIZING ITEM CHARACTERISTICS

Method

Subjects. Subjects were 82 (41 males, 41 females) introductory psychology student volunteers selected from a larger group of 349 students. These data were first reported by Jackson (1968) and later by Rogers (1970).

Materials and Procedure. Each subject completed the MMPI (Hathaway & McKinley, 1967) on two occasions, approximately one week apart. Efforts were made to ensure that the two testing sessions were similar. The MMPI is a personality inventory having 550 unique items and 16 repeated items, and measuring various dimensions of psychopathology. It was selected for this study because its items range widely in social desirability scale value and in p-value.

Results

Properties of the Testing Material. The MMPI was scored for the 13 clinical scales (Hathaway & McKinley, 1967) and for Wiggins' (1966) 13 content-based scales. Scale means, standard deviations and KR-20s for both test

Table 6

MMPI clinical scale means, standard deviations and KR-20s

Test condition
(N=82)

Scale	Mean	Standard Deviation	KR-20
L	3.33	1.90	.46
F	6.49	4.53	.76
K	13.60	4.36	.69
Hy	21.99	4.56	.50
D	22.23	6.42	.74
Hs	5.77	3.63	.71
Pd	17.99	5.66	.71
Mf	32.79	5.85	.66
Pa	10.77	3.15	.42
Pt	18.53	8.69	.90
Sc	17.20	8.80	.87
Ma	18.13	4.55	.58
Si	29.65	9.61	.85

Table 7

MMPI clinical scale means, standard deviations and KR-20s

Retest condition
(N=82)

Scale	Mean	Standard Deviation	KR-20
L	3.09	1.97	.52
F	5.63	4.40	.77
K	13.89	4.64	.72
Hy	20.25	4.70	.57
D	21.19	6.15	.74
Hs	4.49	3.35	.73
Pd	16.80	5.17	.67
Mf	32.09	5.83	.66
Pa	9.86	3.25	.49
Pt	16.58	9.28	.92
Sc	15.53	8.88	.88
Ma	17.99	4.34	.55
Si	28.32	10.27	.87

Table 8

Wiggins' MMPI content scale means, standard deviations and KR-20s

Test Condition
(N=82)

Scale	Mean	Standard Deviation	KR-20
Social Adjustment	10.18	5.92	.86
Organic Symptoms	5.49	3.35	.64
Poor Health	5.33	2.97	.59
Psychoticism	9.70	4.73	.75
Hypomania	13.86	3.56	.67
Phobias	7.89	4.01	.75
Poor Morale	10.39	5.65	.88
Religious Fundamentalism	4.29	3.19	.85
Feminine Interests	13.60	4.92	.77
Depression	9.79	5.86	.87
Manifest Hostility	10.13	4.39	.75
Family Problems	5.69	2.93	.70
Authority Conflict	8.44	3.89	.75

Table 9

Wiggins' MMPI content scale means, standard deviations and KR-20s

Retest condition
(N=82)

Scale	Mean	Standard Deviation	KR-20
Social Adjustment	10.25	6.13	.87
Organic Symptoms	4.30	3.22	.70
Poor Health	4.23	2.69	.60
Psychoticism	8.11	4.16	.71
Hypomania	13.95	3.91	.74
Phobias	7.04	4.02	.77
Poor Morale	9.71	6.23	.91
Religious Fundamentalism	4.10	5.08	.85
Feminine Interests	14.05	5.08	.78
Depression	8.72	6.43	.90
Manifest Hostility	10.10	4.80	.80
Family Problems	5.17	2.94	.72
Authority Conflict	8.32	4.61	.84

Table 10

Test-retest correlations for 13 MMPI clinical scales (N=82)

Scale	Test-Retest Correlation
L	.80
F	.91
K	.87
Hy	.86
D	.90
Hs	.84
Pd	.91
Mf	.91
Pa	.78
Pt	.93
Sc	.92
Ma	.82
Si	.91

Table 11

Test-retest correlations for Wiggins' content scales (N=82)

Scale	Test-retest Correlation
Social Adjustment	.92
Organic Symptoms	.84
Poor Health	.76
Psychoticism	.80
Hypomania	.86
Phobias	.85
Poor Morale	.94
Religious Fundamentalism	.95
Feminine Interests	.96
Depression	.94
Manifest Hostility	.90
Family Problems	.91
Authority Conflict	.89

and retest are reported in Tables 6 to 9. Test-retest data for these scales are reported in Tables 10 and 11. These scale data tend to be quite comparable to those published by Hathaway and McKinley (1967) and by Wiggins (1966), suggesting that the present sample exhibits characteristics similar to those exhibited by other samples.

Response Stability. Subjects changed 11.8 per cent of their responses or a mean of 66.96 responses (standard deviation, 17.93) of 566 responses. Males changed a mean of 68.93 responses (standard deviation, 18.73) while females changed a mean of 65.00 responses (standard deviation, 17.48). This mean difference was not significant (t(80) = .99, n.s.). The range of unstable responses was 24 to 124 responses. A number of other consistency indices were calculated for these data. The TR index, based on a count of identical responses to the repeated MMPI items, was calculated for both the test and retest conditions, as was the number of missing responses. Finally, across-session profile stability estimates were calculated for individuals using both the clinical scale data and Wiggins' content scales. Means, standard deviations and intercorrelations are reported in Table 12.

Model TH. Parameters for the threshold model were calculated using the social desirability scale values of the 566 MMPI items as item properties (see Table 13). As previously demonstrated by Rogers (1970), thresholds and

Table 12

Means, standard deviations and intercorrelations of MMPI consistency indices (N=82)

	Mean	Standard Deviation
TR index		
- test	14.32	1.08
- retest	14.62	.87
Missing responses		
- test	4.62	9.52
- retest	4.11	12.43
Profile stability*		
- clinical scales	2.12	.33
- Wiggins' scales	1.66	.35

* based on an r to z transformation

Intercorrelations

	1	2	3	4	5	6
1. TR - test	1.00					
2. TR - retest	.38	1.00				
3. MR - test	.04	.01	1.00			
4. MR - retest	-.08	-.06	.83	1.00		
5. PS - clinical	.03	-.06	-.26	-.26	1.00	
6. PS - Wiggins	.19	.17	-.18	-.20	.39	1.00

Significant correlation: .21

Table 13

Parameters for the threshold model based on 566 MMPI items (N=82)

		Mean	Standard Deviation
Threshold	-- test	5.22	.58
	-- retest	5.10	.59
Salience	-- test	.51	.18
	-- retest	.55	.19
Proportion of True Responses			
	-- test	.43	.06
	-- retest	.43	.06
Intercept	-- test	-.16	.24
	-- retest	-.20	.24
Slope	-- test	.12	.04
	-- retest	.14	.04

saliences showed adequate test-retest stability (r = .84 and r = .96, respectively, p<.001) and were only moderately intercorrelated for test (r = .26, p<.05) and retest (r = .32, p<.01).

Models PV and SD. The range of p-values for the 566 MMPI items was 1.00 to 0.00. The mean p-value was .43, with a standard deviation of .07. These p-values were based on the responses of the 82 subjects in Study 2. The range of social desirability scale values for the 566 MMPI items was 8.73 to 1.11 (on a 9-point scale). The mean social desirability scale value was 4.64, with a standard deviation of 2.35 (Messick & Jackson, 1961). The correlation of p-values and social desirability scale values was r = .73, p<.001.

Model Comparison. The predictive accuracy of the three models was compared as before. Given the total number of item responses an individual changed on retest, each model was required to predict exactly which items the individual would change. Model SD and PV predicted these items would be most moderate in social desirability and p-value, respectively. Model TH predicted items falling closest to the individual's threshold would be most likely to be changed. The number of correct predictions was compared across subjects using a one way analysis of variance with repeated measures.

The mean number of correct predictions for each of the three models is reported in Table 14 and compared to chance. A series of *t*-tests suggested that Model PV (*t*(81) = 7.79, p< .001) was significantly better at predicting changes than chance as were Models SD (*t*(81) = 2.95, p<.01) and TH (*t*(81) = 3.47, p<.001). An analysis of variance with repeated measures comparing the three models suggested that there were clearly significant model differences [*F*(2,162) = 10.23, p< .001]. When models were compared using Dunn's Multiple Comparison Test, Model PV was found to be significantly better than Models SD and TH, which did not differ from one another.

Model TH was also evaluated for individuals placed in a high and a low salience group, based on a median split. Model TH was not a better predictor for the high than for the low salience group. Means of 9.4 and 10.5 correct predictions were made for the high and low salience groups, respectively. When these means were divided by the number of changes made by the two groups, Model TH predicted 14.9 per cent of response changes correctly for each group.

Discussion

The results of this study once again indicated that individuals' item responses tend to be stable across a one week interval. Individuals tended to change less than 12 per cent of their responses.

Table 14

Mean number of correct predictions for each of the three models and compared to chance (N=82)

Model	Correct Predictions	
	Mean	Standard Deviation
Model SD	10.15*	6.47
Model PV	11.77**	6.64
Model TH	9.96**	6.29
Chance	8.91	4.61
Mean item changes	66.96	17.93

** Significantly different from chance at $p < .001$.
* Significantly different from chance at $p < .01$.

A variety of within- and across-session indices of person reliability were calculated. The modest intercorrelations among these indices support previous findings (e.g., Berdie, 1969; Fiske, 1957a, 1957b; Holden et al., 1981) and hypotheses that such indices do not reflect a single, stable attribute of individuals' responding. For example, the TR index calculated at Time 1 only correlated significantly with Profile Stability based on Wiggins' scales and with the TR index calculated at Time 2. One possible contributor to these small correlations, however, may be the low variability associated with all but the missing response index. Alternatively, these consistency indices may reflect different facets of the person consistency construct.

In this study, an evaluation of the three models suggested that all models were significantly better than change at predicting exactly which items would be changed on retest. Surprisingly, Model PV predicted which items would be changed better than either Model SD or Model TH, both related to the social desirability of items. The express purpose of Study 2 was to allow evaluation of models, especially Model TH, for predicting unstable items under conditions where the social desirability of items was very salient. Indeed, a straightforward comparison of the mean salience for respondents in Study 2 (.51, standard deviation .18) and in Study 1 (.44, standard deviation .04) would suggest that social desirability was a stronger determinant of responses in Study 2. Yet, Models SD and TH did not predict unstable items better than Model PV. The data collected in this study do not provide support for the hypothesis that Model TH will be a better predictor than models based on single item properties when the property underlying the threshold model is clearly related to one's responses.

Model PV was the best predictor of unstable items in this study. As said previously, the relationship of p-value to the consistency of items may be moderated by a variety of other item properties. That is, the relationship of p-value to response stability may be a function of the fact that willingness to endorse an item may reflect a variety of item characteristics. The decision to endorse may be made in terms of content, desirability or some irrelevant item characteristic, such as item length, negative wording, etc. Although there may be no homogeneous response determinant underlying p-value, particular items' p-values could reflect the same determinant across individuals. When the range of p-values is not restricted (as in Study 1), p-values may indeed be significantly related to item stability.

The use of the threshold model for responding to predict across-session item consistency for items varying in terms of certain item properties demands a consideration of individual differences in the importance of item characteristics. Neither Study 1 nor Study 2 explicitly considered the role of individuals' perceptions of the item characteristics. Empirical studies have indicated that items with extreme properties are very stable across individuals. These items may be so stable because the item property is so salient that all individuals tend to respond in terms of it. However, for items which are relatively moderate in, for example, social desirability, considerable individual differences are apparent in the influence of the item property on response selection. The salience parameter, in the threshold theory for responding, actually reflects the strength of the item property in determining the response. The threshold model for responding is hypothesized to be a more accurate model for predicting the stability of individual item responses for individuals with high saliences than for individuals with low saliences. Consider two individuals with the same threshold but one having a very low salience and the other having a very high salience parameter. The individual with the low salience appears to have his or her responses somewhat determined by the item property explicitly considered by the threshold model and somewhat determined by other properties. The individual with the higher salience appears to have his or her responses largely determined by the item property explicitly considered by the threshold model. The first individual might be expected to have no more difficulty responding to items inside the threshold range (based on a single item property) than outside it because responses are being influenced by a variety of factors. Thus, little difference in stability is expected inside or outside the threshold range. On the other hand, the individual with the higher salience is expected to experience relatively more difficulty responding to items near the threshold because responses are indeed being influenced by the item property explicitly considered by the threshold model.

The primary purpose of Study 3 was to examine the role of the salience of item properties. It was expected that the predictive accuracy of the threshold model would be better for individuals with higher saliences. This hypothesis was tested by experimentally manipulating saliences and by collecting individuals' ratings of the social desirability of items and using these ratings to estimate threshold.

CHAPTER VI

STUDY 3 : MODEL COMPARISON EMPHASIZING DECISION DIFFICULTY

Method

Subjects. Subjects were 90 student volunteers at a major Ontario university. Subjects were recruited through summer school classes, and through ads placed in campus newspapers and posted on campus bulletin boards. The average age of the 30 males in this study was 22.1 years (standard deviation of 4.7 years) and of the 60 females was 23.4 years (standard deviation of 5.4 years). In exchange for their participation, subjects received a computerized personality profile and a one dollar Canadian coin.

Materials. The twenty content scales and two validity scales (Infrequency and Desirability) yielding 352 items were adopted from Jackson's (1974) PRF.

Procedure. Subjects were asked to complete the following three tasks. First, six PRF content scales (Abasement, Achievement, Affiliation, Aggression, Autonomy and Change) comprising 96 items were administered using a PDP-12 computer (Computer Task). Items were presented one after the other on the computer video terminal. Subjects responded to the items using a separate control panel so that response latencies could be collected. The control panel had two keys, the right one marked "T" to indicate a True response and the left one marked "F" to indicate a False response. The control panel also had a "Home" or a resting spot to which they were requested to return their index finger between responses. The second task involved rating the social desirability of the same 96 items on a nine-point scale (Rating Task). The third task involved completing the entire PRF in paper and pencil format (PRF). After completing these three tasks, all subjects were asked to complete a fourth task, which involved making a series of global personality judgements and consistency ratings for an unrelated study. Instructions for the three experimental tasks are listed in Appendix B.

Subjects were assigned randomly within sex to one of three groups so that the order of the three tasks could be varied. Each group was made up of 20 females and 10 males. In Group 1, subjects completed the Rating Task first, then the Computer Task and then the PRF. This manipulation was to induce a strategy of responding in terms of social desirability. In Group 2, subjects completed the Computer Task, the Rating Task and then the PRF. This manipulation was to induce a change in response strategy from the Computer Task to the PRF. In Group 3, subjects completed the Computer Task, the PRF and then the Rating Task. Group 3 was the control group to whom no particular strategy for responding was suggested. In each group, subjects were tested individually. The testing room contained a table on which the video terminal and control panel were placed, a desk and two chairs. The PDP-12 computer was located in an adjoining room so that the noise generated by the computer would not distract the subject.

Results

Properties of the Testing Material. The means, standard deviations and KR-20s for the PRF scales are reported in Tables 15 and 16 for the Computer Task and the PRF, respectively. Scale data are quite comparable to those reported in the PRF manual (Jackson, 1974). Test-retest correlations, also reported in Table 15, for the six PRF scales which subjects completed twice tend to be very high.

The mean latency for responding to an item across all 90 subjects was 2.98 seconds (standard deviation, 1.83). Individual subjects' mean latencies varied from 1.35 to 11.62 seconds. Males took a mean of 3.46 seconds to respond to items while females took a mean of 2.73 seconds to respond to an item ($\underline{t}(89) = 2.09, p < .04$). There was no relationship between age and mean latency ($\underline{r}(88) = -.10$, n.s.).

Response Stability. An unstable item response was again defined as the change from a "True" response to a "False" response on retest or vice versa. The average

Table 15

Means, standard deviations, KR-20s and test-retest correlations for the six computer administered PRF scales (N=90)

Scale	Mean	Standard Deviation	KR-20	Test-retest Correlation
Abasement	7.30	2.75	.59	.88
Achievement	11.54	2.83	.68	.92
Affiliation	8.92	4.01	.83	.97
Aggression	7.22	2.88	.62	.91
Autonomy	7.41	3.33	.70	.93
Change	10.27	3.17	.72	.96

Table 16

Means, standard deviations and KR-20s for 22 PRF scales (N=90)

Scale	Mean	Standard Deviation	KR-20
Abasement	7.10	3.07	.68
Achievement	11.46	3.01	.72
Affiliation	9.01	4.18	.85
Agression	7.09	3.16	.68
Autonomy	7.20	3.52	.74
Change	10.09	3.14	.70
Cognitive Str.	9.36	3.31	.75
Defendence	6.56	2.87	.63
Dominance	9.67	4.27	.86
Endurance	10.26	3.04	.70
Exhibition	6.93	4.52	.88
Harmavoidance	8.84	4.32	.86
Impulsivity	6.11	3.57	.77
Nurturance	10.98	2.93	.70
Order	8.04	4.99	.90
Play	8.64	3.32	.72
Sentience	10.19	2.91	.68
Social Recog.	8.73	3.53	.79
Succorance	7.13	3.42	.74
Understanding	10.04	3.16	.72
Infrequency	.38	.68	.22
Desirability	11.21	3.12	.73

subject changed 8.60 per cent of his or her responses or a mean of 8.26 of 96 item responses with a standard deviation of 2.71 responses. The range of unstable responses was 1 to 31.

<u>Sources</u> <u>of</u> <u>Group</u> <u>Differences</u>. The effectiveness of the response strategy manipulation was evaluated by examining a variety of hypotheses that derive from the manipulation. Since Group 3 was the control group, it was predicted to have the lowest mean social desirability scale score on the PRF. While Group 3 did have the lowest mean scale score of the three groups on desirability, this difference was not significant [<u>F</u>(2,87) = .65, n.s.]. Group 2 was predicted to make the most item changes, since this group was expected to experience a change in response strategy after completing the Rating Task. Group 1 was expected to make the least item changes because a desirable responding strategy had been introduced before they completed either of the two other tasks. Group 3, which experienced neither a strategy suggestion nor a strategy change, was expected to fall between Groups 1 and 2. As predicted, Group 2 made the most item changes (9.6 changes, standard deviation 4.3) while Group 1 made the least (7.1 changes, standard deviation 4.1) and Group 3 fell between the two (8.0 changes, standard deviation 5.5). However, the differences were not significant at the .05 level [<u>F</u>(2,87 = 2.20, p< .12]. In a 3 by 2 analysis of variance comparing the mean number of changes for each group by sex,

Table 17

Percentage of changes made in the desirable direction

Desirable changes relative to group SDSV

	Mean	Standard Deviation
Group 1	.37	.22
Group 2	.47	.14
Group 3	.53	.26

Desirable changes relative to individual SDSV

	Mean	Standard Deviation
Group 1	.45	.27
Group 2	.50	.15
Group 3	.60	.26

a significant main effect for sex was obtained [F(1,84) = 10.80, p< .001]. Males changed an average of 10.4 responses while females changed an average of 7.1 responses. There was no significant effect for group nor for a group by sex interaction.

An analysis of variance comparing the groups on the number of item changes made in the desirable direction suggested that the groups were significantly different (see Table 17). When desirable changes were calculated relative to group desirability scale values, Group 1 made the smallest percentage of changes in the desirable direction and Group 3 made the largest percentage of changes in this direction [F(2,87) = 4.04, p< .01]. When desirable changes were calculated relative to individuals' own ratings of desrability, again Group 1 made the smallest percentage of desirable changes while Group 3 made the largest percentage of desirable changes [F(2,87) = 2.93, p< .06].

Model RL. The relationship of latency to stability was again examined by correlating mean latency and total number of item changes for all 90 subjects. This relationship was small and nonsignificant (r = -.002). The correlation of change and latency for responding to an item at the individual level did suggest that response latency and change were inversely related. The correlations between unstable responses and response latencies for each of the 90 subjects were generally positive, ranging from

-.06 to .53, with an average correlation of *r* = .15.

Model TH. In this study, parameters for the threshold model could be calculated in various ways. Thresholds and saliences were calculated using both group SDSV and individual SDSV. Further, thresholds and saliences could be calculated for both individuals' test and retest PRF item responses. The means and standard deviations of the threshold parameter are reported in Table 18. A series of one way analyses of variance comparing groups' mean thresholds for each condition suggested that groups did not differ significantly on threshold calculated under any one of the four conditions. Thus, data were collapsed across groups to examine the relationships among different threshold estimates. Intercorrelations among thresholds tended to be moderately high (see Table 19). The means and standard deviations of the salience parameter are reported in Table 20. A series of one way analyses of variance indicated that the groups did not differ significantly on salience calculated under any one of the four conditions. However, a three-way analysis of variance, with repeated measures on salience, with groups, sessions and type of social desirability scale values as grouping variables, yielded a significant main effect for type of SDSV. No other main effects were significant (see Table 21). The interaction between session and SDSV type was also significant.

Table 18

Comparison of thresholds
for the three experimental groups (N=90)

	Thresholds based on 96 items			
	Session 1	Session 1	Session 2	Session 2
	Group SDSV	Individual SDSV	Group SDSV	Individual SDSV
Group 1	5.36 (.83)	5.40 (.85)	5.18 (1.01)	5.33 (.81)
Group 2	5.40 (1.11)	5.21 (.91)	5.26 (1.25)	5.15 (1.03)
Group 3	5.31 (.83)	5.13 (.92)	5.06 (1.01)	4.82 (.74)
	------	------	------	------
	5.36 (.92)	5.25 (.89)	5.17 (1.09)	5.10 (.88)

2 OF/DE 2

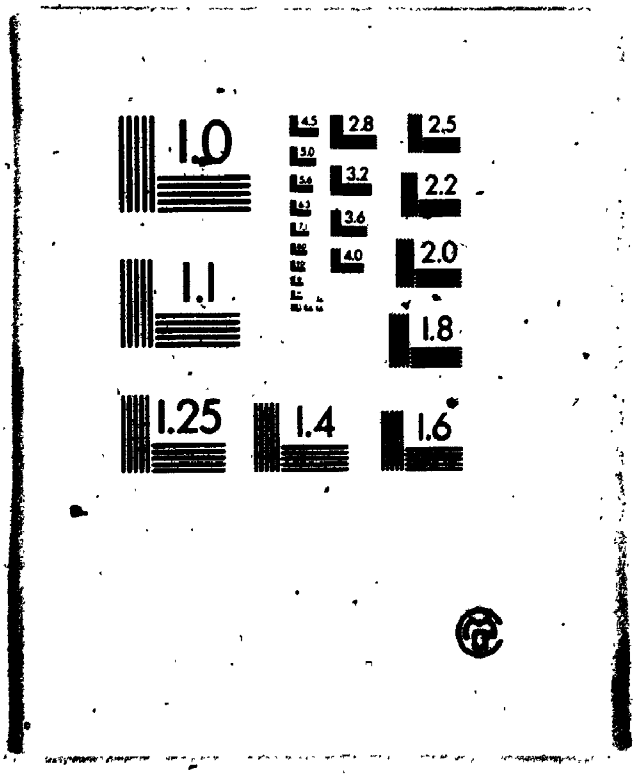

Table 19

Intercorrelations among different threshold estimates
(N=90)

Thresholds	1	2	3	4
Session 1 Group SDSV	1.00			
Session 1 Individual SDSV	.61	1.00		
Session 2 Group SDSV	.80	.56	1.00	
Session 2 Individual SDSV	.55	.76	.67	1.00

Table 20

Comparison of saliences
for the three experimental groups (N=90)

	Saliences based on 96 items			
	Session 1	Session 1	Session 2	Session 2
	Group SDSV	Individual SDSV	Group SDSV	Individual SDSV
Group 1	.37 (.25)	.60 (.21)	.35 (.24)	.60 (.21)
Group 2	.36 (.20)	.59 (.23)	.34 (.23)	.60 (.24)
Group 3	.33 (.22)	.50 (.23)	.32 (.26)	.52 (.27)
	----- .35 (.22)	----- .57 (.23)	----- .34 (.24)	----- .57 (.24)

Table 21

Analysis of variance with repeated measures comparing the three experimental groups on salience with type of desirability scale values and session as factors

Source	Sum of Squares	DF	Mean Square	F	p
Group	.29	2	.14	.85	.43
Error	14.70	87	.17		
Session	.00	1	.00	.05	.82
S x G	.00	2	.00	.40	.67
Error	.39	87	.00		
Desirability	4.66	1	4.65	113.06	.00
D x G	.06	2	.03	.78	.46
Error	3.58	87	.04		
S x D	.01	1	.01	6.96	.01
S x D x G	.00	2	.00	.14	.87
Error	.14	87	.00		

Data were collapsed across the three groups to compare the relationship among different salience estimates. Intercorrelations among saliences tended to be high when saliences were based on the same type of SDSV and moderate otherwise (see Table 22).

Models PV and SD. The range of p-values for the 96 PRF items used in this study was .78 to .09. The mean was .51, with a standard deviation of .03. These p-values were based on data reported by Helmes, Reed & Jackson (1977). The range of social desirability scale values for these items was 7.54 to 2.48. The mean SDSV was 5.21, with a standard deviation of 1.48. The correlation of p-values and social desirability was $\underline{r}$ = .69, p<.001.

Model Comparison. Unstable items were to be predicted by six models: Model RL, Model SD-G (based on group SDSV), Model SD-I (based on individuals' SDSV), Model PV, Model TH-G (based on group SDSV) and Model TH-I (based on individuals' SDSV). Each model was again required to predict exactly which items an individual changed, given the total number of items the individual changed. The means and standard deviations tor each of these models for each experimental group is given in Table 23. A multivariate analysis of variance was undertaken to compare the three experimental groups. Both Hotellings' and Wilks' tests of significance indicated that there were no significant differences in predictability among the three

Table 22

Intercorrelations among different salience estimates
(N=90)

Saliences	1	2	3	4
Session 1 Group SDSV	1.00			
Session 1 Individual SDSV	.56	1.00		
Session 2 Group SDSV	.93	.60	1.00	
Session 2 Individual SDSV	.56	.96	.64	1.00

Table 23

Means and standard deviations for correct predictions made by each model for each experimental group

	Models					
	Model PV	Model SD-G	Model SD-I	Model RL	Model TH-G	Model TH-I
Group 1	.60 (.72)	.70 (.99)	.70 (.95)	1.63 (1.38)	.70 (1.06)	.77 (1.01)
Group 2	1.23 (1.33)	1.17 (1.42)	1.73 (1.66)	2.07 (2.05)	1.53 (1.54)	1.30 (1.49)
Group 3	.90 (1.81)	1.00 (1.91)	1.20 (2.86)	1.57 (2.11)	.97 (1.97)	1.33 (1.60)
	.91 (1.37)	.96 (1.48)	1.21* (2.08)	1.76** (1.87)	1.07 (1.59)	1.13 (1.40)

** Significantly different from chance at p< .001.

* Significantly different from chance at p< .02.

groups [F(12,162) = 1.07, n.s.] and [F(12,164) = 1.08, n.s.], respectively. Univariate analyses suggested that the groups did not differ significantly on any model (see Table 24).

A 3 by 6 analysis of variance with repeated measures was performed, using groups and models as factors (see Table 25). Again, there was no significant main effect for groups although there was a highly significant effect for models. There was no significant effect for a model by group interaction. Dunn's Multiple Comparison Test suggested that Model RL was significantly better than all other models, which did not differ from one another significantly. Since the predictability of changes may be a function of the number of changes, an analysis of covariance was performed (see Table 26), covarying the number of unstable items out of the correct predictions made by each model. Again, there was no significant main effect for group. The covariate was highly significant as was the main effect for models. There was no significant model by group interaction.

Univariate tests comparing models to chance (collapsing across groups) did indicate that some models predicted better than chance while others did not. A series of t-tests comparing each model to the number of correct predictions expected by chance (.93 changes, standard deviation 1.28) indicated that only Models RL and

Table 24

Univariate analyses comparing the three experimental groups on each model

Model	F	Significance of F with 2 and 87 degrees of freedom
Model RL	.63	.54
Model SD-I	2.03	.14
Model SD-G	.76	.47
Model PV	1.62	.20
Model TH-G	2.20	.12
Model TH-I	1.57	.21

Table 25

Analysis of variance with repeated measures comparing three experimental groups on six models

Source	Sum of Squares	DF	Mean Square	F	p
Group	38.71	2	19.36	1.68	.19
Error	1002.77	87	11.53		
Model	42.26	5	8.45	9.59	.00
M x G	8.04	10	.80	.91	.52
Error	383.19	435	.88		

Table 26

Analysis of covariance with repeated measures comparing three experimental groups on six models covarying out the total number of response changes (N=90)

Source	Sum of Squares	DF	Mean Square	F	p
Group	.59	2	.29	.14	.87
Covariate	826.67	1	826.67	403.70	.00
Error	176.10	86	2.05		
Model	42.26	5	8.45	9.56	.00
M x G	8.04	10	.80	.91	
Error	383.19	435	.88		

SD-I predicted item changes significantly above chance. These data are also reported in Table 23. The prediction that models based on person and item characteristics in interaction would perform better than models based on item characteristics alone was evaluated using a t-test. Models RL and TH-G predicted significantly more item changes than Models PV and SD-G (t(89) = 4.44, p<.001). Similarly, Models RL and SD-I predicted significantly more item changes than Models PV and SD-I (t(89) = 3.84, p<.001).

Discussion

The results of this study support those of the previous two studies in terms of response stability. Subjects changed only approximately eight per cent of their responses. No doubt the low rate of response change was in part a function of the extremely short test-retest interval. The high test-retest scale correlations associated with the PRF scale scores also suggest that responding was extremely stable.

The purpose of Study 3 was to evaluate the different models for predicting item response changes when the social desirability scale values of items were made salient. Specifically, the usefulness of the threshold model for responding for predicting exactly which items an individual would change on retest was to be examined under conditions where the item property considered by the threshold model was experimentally made salient. The first issue was whether the social desirability manipulation was effective. The data do not clearly indicate that the manipulation had the hypothesized effect. Although, as predicted, Group 1 consistently had the highest salience in all conditions and Group 3 consistently had the lowest saliences, these differences were not significant. For social desirability scores as well, Groups 1 and 3 were expected to have the highest and lowest scale scores, respectively. Mean differences were in the correct direction but were not significant. Group 2 made the most item changes, as predicted, and Group 1 made the least item changes but again mean differences were not significant. These data resemble Bentler's (1964) findings: making the SDSVs of items salient for Group 1 did not significantly reduce the variability of retest changes for that group. Perhaps, as Bentler argues, the SDSVs are generally so neutral that increasing their salience still does not make desirability a significant influence on responding. Only when considering the number of response changes in the desirable direction were significant differences among groups evident. As predicted, Group 1 made the fewest desirable changes, presumably because they had already been responding in terms of desirability. Group 3, the control group, showed the largest percentage of changes in the desirable direction, a result in keeping with Windle's (1954) findings. Group 2 made more desirable changes than

Group 1 but fewer than Group 3. Group 2 might have been predicted to make the most desirable changes, since a strategy of responding desirably was suggested specifically to them. Alternatively, rating the desirability of items after having responded to them may have served to make desirability more salient and, hence, simply increased stable responding. Thus, it is not necessarily clear just how the rating task influenced the response strategy of Group 2.

In Study 3, the second approach to making the desirability of items salient was actually to use individuals' own ratings of social desirability to calculate the threshold model parameters. The advantage of using individuals' ratings of the social desirability of items is that the ratings should reflect the individuals' perception of the desirability of the item. In a sense, any model based on individuals' desirability judgements can be considered to take into account an interaction between items and persons. Such a model would predict that individuals will change items on retest which they themselves judge as moderate in desirability and not those per se which the group judges as moderate.

The evidence in Study 3 indicates that individuals' responses were more strongly and significantly related to their individual SDSV ratings than to group SDSV ratings. That is, saliences based on individual SDSV were significantly higher than saliences based on group SDSV for all three groups in this study.

Model Comparison. The results of this study suggest that it may be possible to predict above chance level precisely which items an individual will change on retest. Both Models RL and SD-I predicted unstable items above chance level. Models PV, SD-G, TH-G, and TH-I were unable to predict above chance. Further, Model RL was significantly better than all other models. It had been predicted that models which take into account an interaction between persons and items would perform better than models taking into account only item parameters. This prediction was easily upheld when Model RL and TH-G or TH-I were compared to Model PV and SD-G or SD-I. This hypothesis was also supported by the finding that Model RL was overall the best model. The data did not support rejecting the null hypothesis in the case of Model TH. In the current evaluation, Model TH, based either on group or individual SDSVs, did not perform better than models based on single item parameters. The superiority of the threshold model was not supported even though there was some evidence, albeit weak, that the salience of the desirability of items had been experimentally increased for some groups. Further, the superiority of the threshold model was not supported even when threshold parameters were based explicitly on individuals' perceptions of item desirability.

Given the varying degrees of success for the models, an interesting avenue for future research might be to examine the usefulness of these models for individuals making more or fewer response changes. Results of the analysis of covariance, in which the number of response changes was the covariate, yielded a highly significant effect for this covariate. Perhaps, different response processes are operating when individuals make relatively few or relatively many item changes. Hence, models might show differential usefulness for such subgroups.

CHAPTER VII

GENERAL DISCUSSION

To recapitulate, the overall purpose of this research was to evaluate four models for predicting person reliability, which was defined in terms of across-session item stability. A review of the relevant literature showed that reliable indices of both within- and across-session person reliability could be constructed. Empirical evidence on the generalizability of these indices was interpreted as support for a multidimensional view of person reliability. An attempt to clarify the nature of the person reliability construct involved adopting a simple definition of person reliability, namely, the across-session consistency of individuals' item responses. Obtaining such a more refined index of person reliability has both practical implications for the assessment of person reliability and theoretical implications for understanding the process of consistent responding. Previous research has shown that the stability of individuals' responses to items is a function of the test-taking situation as well as the characteristics of the items and the persons themselves. Formulations which have taken into account aspects of both items and persons simultaneously have generally been better at predicting which items individuals would change on retest than formulations which consider item or person characteristics alone.

The four models for predicting across-session item consistency evaluated by this research were as follows. Model PV, based on the p-values of items, was derived from the empirical evidence in the literature that extreme item endorsement probabilities are positively related to item stability. Model SD, based on the social desirability scale values of items, was derived from the empirical finding that extreme social desirability is positively related to item stability. Models TH and RL were based on Jackson's threshold model for responding and response latencies, respectively. These models were expected to take into account an interaction between item parameters and the person in determining item stability. Model TH predicted an inverse relationship between item stability and nearness to an individuals' threshold for endorsing items in terms of some item property. Model RL predicted an inverse relationship between response latency for an item and item stability.

Three studies were conducted to evaluate the general hypothesis that the two models taking individual differences into account (i.e., Models TH and RL) would be better predictors of across-session item consistency than the two models based on item properties alone. Specifically, models were compared on how well they could predict exactly which items an individual would change on retest. Thus, the focus in this study was on the proportion of changed items correctly predicted by the models. One could, of course, examine those items incorrectly predicted by the models -- either those predicted to change but which remained stable or those predicted to remain stable but which changed. The evaluation of such data would certainly have interesting implications. However, since the present research was an initial evaluation of the proposed models, the definition of across-session item consistency that was adopted was consistent with definitions found in the literature.

In Study 1, the models were evaluated using the test and retest responses of 40 introductory psychology students to 192 items selected from the Personality Research Form (Jackson, 1974). Models SD, TH, and RL predicted changed item responses significantly above chance level. Further analyses indicated that Model RL was significantly better than the other three models, which did not differ from one another. A post hoc analysis was conducted, in which the predictive power of Model TH was compared for individuals divided into a high and low salience group. Model TH was a modestly stronger predictor of item stability for the high salience group, which tended to respond to items in terms of desirability. Thus, in Studies 2 and 3, the models for predicting item response stability were to be compared under circumstances favoring the assumption that the social desirability of items was salient.

In Study 2, Models SD, PV, and TH were compared using the 566 MMPI items, which were selected because they varied greatly in terms of social desirability. The results of this study, based on the test-retest responses of 82 college students, did not support the hypothesis that Model TH would be a better model than those based on item characteristics. While all models could readily predict above chance level which items would be changed by individuals on retest, Model PV was found to be significantly better than either Model SD or Model TH.

In Study 3, involving 90 subjects, models were compared under conditions designed to make the social desirability of the relatively neutral PRF items salient. One strategy used in Study 3 was to manipulate experimentally the salience of desirability using a rating task. Empirical evidence that this rating task made desirability salient is weak. Predicted mean differences were in the right direction but generally failed to reach statistical significance. The second strategy for increasing the salience of desirability was actually to use individuals' own judgements of items' desirability. A significantly stronger relationship was found between individuals' responses and their own individually judged desirability than between responses and desirability scale scale values based on the judgements of an independent group.

Thus, in Study 3, six models were compared across three experimental groups: the four models outlined previously plus a desirabilty and a threshold model based on individuals' desirability judgements. While there were no significant group differences to support the efficacy of the rating task, there was a strong significant main effect for models. As in Study 1, Model RL was the model best able to predict exactly which items an individual would change on retest. In Study 3, the only other model able to predict above chance which items would be changed was Model SD-I, which predicted that items judged by individuals themselves to be relatively moderate in social desirability would be more likely to be changed on retest than items judged to be extreme. It was argued that this model takes into account an interaction between persons and items. That is, the basis of this model is individuals' judgements of items' desirability. Unlike the models based on group desirability (which predict individuals all will change the items most moderate according to group desirability judgements), Model SD-I does not make invariant predictions. Rather, it predicted that individuals change items which they personally view as moderate; hence, this model takes into account individual differences in the perception of desirability.

There are a variety of implications to be derived from these three studies. First, these studies argue that even though item responses tend to be quite stable, predicting precisely which items individuals change on retest is possible. Second, when the different models were compared in terms of their predictive accuracy, the response latency model, Model RL, was clearly the best model. This finding generalized across both studies evaluating Model RL. Although the testing conditions employed in these studies were by necessity not standard testing conditions, two sources of empirical evidence would support the hypothesis that Model RL will show further generalizability. First, in Study 1, considerable similarities emerged between the retest responses of groups tested in the latency format and the standard format. Second, the predictive superiority of Model RL was a highly significant phenomenon in both studies.

Consider the predictive accuracy of Model RL in light of the response process. Model RL simply predicted that long latencies or decision times would be related to response instability on the assumption that latency was a function of decision difficulty. What might contribute to decision difficulty and, hence, to instability? Decision difficulty may be related to item characteristics which increase decoding time, such as item length, difficulty ratings and controversiality (Hanley, 1962; Rogers, 1973b). As well, response latency may be a result of subjects' poor item conceptualization, resulting in the use of inappropriate response categories, such as adding new elements or responding in terms of recent experiences (Kuncel, 1973). Most interesting, however, may be the conceptualization of response latency as a function of the difficulty of a comparison of the item to the self on the basis of their relative positions on some underlying attribute (Ebbesen & Allen, 1979; Kuiper, 1981; Rogers, 1974). Model RL simply predicts that when this comparison involves a relatively long time, the response is likely to be unstable because of discrimination difficulty. Thus, rather than emphasizing the relationship between inappropriate responding and response instability, Model RL is useful for predicting response instability within appropriate categories. Not only does Model RL subsume Model TH, but it includes other decision bases as well. Response instability may be increased when the decision involves making a fine discrimination regarding the relative positions of the self and item on the underlying continuum. Or, as others have suggested (Kuiper, 1981; Rogers, Kuiper, & Kirker, 1977; Johnson, 1981), response instability may be increased when the conceptualization of the self, to which the item is compared, is unstable or poorly crystallized. Indeed, individual differences in personality differentiation or crystallization within domains may yield significant influences on the decision difficulty and response latency associated with particular item responses. Future research might seek to differentiate among the influences of item characteristics, of item conceptualization, as related to verbal ability, and, of domain articulation on response stability or difficulty as measured by response latency.

There are practical as well as theoretical implications for Model RL's success in identifying unstable items. In the age of the microcomputer and computerized testing, practical applications for such information include use in evaluating the interpretation of individual "critical items", constructing tailored tests and indicating a degree of domain articulation.

Model RL's predictive accuracy was conceptualized as a tunction of the model's ability to consider persons and items in interaction. Similarly, this ability was also expected to make Model TH useful for predicting unstable items as well. The results of these three studies, however, suggest that Model TH, whether using group or individuals' SDSVs, is not significantly better than models based on SDSV alone. This finding bears mention. First, note that each model considered in this research was required to meet a very stringent test, namely, predicting individual item changes. Predicting specific item responses in the personality domain is a very difficult task (Cliff, 1977; DeBoeck, 1981; Helmes, 1978). Likewise, predicting which items an individual will change on retest is also difficult. All four models can easily predict in general which items will be changed, as was illustrated in Study 1. What the models have considerably more difficulty predicting is exactly which items an individual will change, rather than which items individuals generally change.

Second, consider the performance of the threshold model in light of the relative neutrality of the PRF items. The PRF items have both a content and a desirability component but they were explicitly selected to be more likely to elicit responding in terms of content than desirability. The predictive accuracy of the threshold model is surely a function of social desirability being a determinant of responding. To the extent that individuals respond to items using a strategy independent of desirability (e.g., content, acquiescence, preference for extreme response categories, etc.), the predictive accuracy of the threshold model is weakened.

This problem was further addressed in Study 2, in which Model TH was examined using MMPI items. When using items ranging in SDSV, Model TH was quite able to predict unstable items above chance level. However, Model TH was still not the strongest predictor of unstable items. Surely, MMPI items have both a content and a desirability component. Even though the SDSVs associated with the MMPI tend to be more extreme than those associated with PRF items, individuals may still have been responding in terms of content.

The salience of the desirability of the PRF items was again addressed in Study 3. The desirability of items was to be made salient by two strategies. Evidence for the success of the rating task for making SDSV salient was weak. Bentler (1964) also found that having subjects rate the desirability of PRF items did not yield desirable responding as indicating by making item changes in the desirable direction. Perhaps desirable responding could be induced by using unselected items from the original item pool (cf., Morf & Jackson, 1972) or by using a different task altogether. For example, desirability might be made salieht by presenting items to subjects for an inordinately short period of time (e.g., one second) or by having subjects fill out the PRF in the desirable direction, that is, fake good.

The second strategy for making desirability salient in Study 3 appeared to be relatively successful, given the significantly increased salience scores as well as the relative superiority of Model SD-I, which was based on subjects' own ratings of item desirability. The efficacy of Model SD-I over models based on group judgements of desirability may be explained by conceptualizing Model SD-I as a model concerned with person-item interactions. However, Model TH-I would still have been hypothesiszed as a better model.

The general finding across the three studies that the threshold model was not as useful as predicted yields the following suggestions. Although this model could be evaluated using other items or means of inducing desirable responding, perhaps desirability is not the most appropriate item characteristic on which to focus. Future research might examine the prediction of response instability using the threshold model and content scale values rather than desirability scale values.

The positive findings with Model RL might also be extended in future research to improve item selection for structured tests and to understand further cognitive processes related to responding to personality items. Items with short latencies might be selected for structured tests to help maximize test-retest stability. As well, specific item characteristics which appear to contribute to response stability (e.g., high content scale values or short items) could be outlined. In terms of the response process, the research trend examining aspects of responding to personality items may parallel some research in cognitive psychology on decision processes. Clearly, empirical findings on the relationship among decision correctness, reaction times and stimulus characteristics can be extended to the personality domain in terms of item response stability, response latencies and item-characteristics. Further, researchers in personality have extrapolated certain decision models developed within cognitive psychology to self-referent decisions (Ebbesen & Allen, 1979; Kuiper, 1981; Rogers, 1974). The findings associated with Model RL in the present research support the emphasis of such authors on the role of relevant individual differences in making personality decisions.

Conclusions

In summary, then, the following conclusions may be derived from the present research.

1) Person reliability, defined in terms of across-session item consistency, tends to be high; that is, individuals' item responses tend to show considerable test-retest stability.

2) Empirical support exists for the usefulness of the response latency model for predicting exactly which items an individual will change on retest. This finding has practical applications for item selection and person reliability index construction and theoretical implications for the process of responding.

3) Empirical support for the usefulness of three other models, the threshold model, the desirability model and the p-value model, for predicting unstable items was only modest. The threshold model might be re-evaluated by using potent techniques for making desirability salient or by using content scale values for calculating thresholds.

REFERENCES

Ace, M. Identifying sources of inconsistency. *Proceedings, 77th Annual Convention, American Psychological Association*, 1969, 125 - 126.

Bain, R. Stability of questionnaire response. *American Journal of Sociology*, 1931, *37*, 445 - 453.

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

Bentler, P. M. *Response variability: Fact or artifact?* Unpublished doctoral dissertation, Stanford University, 1964.

Benton, A. L., & Stone, I. R. Consistency of response to personality inventory items as a function of interval between test and retest. *Journal of Social Psychology*, 1937, *8*, 143 - 146.

Berdie, R. F. Intra - individual variability and predictability. *Educational and Psychological Measurement*, 1961, *21*, 663 - 676.

Berdie, R. F. Consistency and generalizability of intra - individual variability. *Journal of Applied Psychology*, 1969, *53*, 35 - 41. (a)

Berdie, R. F. Intra - individual temporal variability and predictability. *Educational and Psychological Measurement*, 1969, *29*, 235 - 257. (b)

Brebner, J. M. T., & Welford, A. T. Introduction: An historical background sketch. In A. T. Welford (Ed.), *Reaction times*. Toronto: Academic Press, 1980.

Chance, J. E. Prediction of changes in a personality inventory on retesting. *Psychological Reports*, 1955, *1*, 383 - 387.

Cliff, N. Further study of cognitive processing models for inventoty 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.

De Boeck, P. Individual differences in the validity of a cognitive processing model for responses to personality inventories. *Applied Psychological Measurement*, 1981, *5*, 481 - 492.

Dunn, T. G., Lushene, R. E., & O'Neil, Jr., H. F. Complete automation of the MMPI and a study of its response latencies. *Journal of Consulting and Clinical Psychology*, 1972, *39*, 381 - 387.

Ebbesen, E. B., & Allen, R. B. Cognitive processes in implicit personality trait inferences. *Journal of Personality and Social Psychology*, 1979, *37*, 471 - 488.

Eisenberg, P. Individual interpretation of psychosomatic inventory items. *Journal of General Psychology*, 1941, *25*, 19 - 40.

Eisenberg, P., & Wesman, A. G. Consistency in response and logical interpretation of psychoneurotic inventory items. *Journal of Educational Psychology*, 1941, *32*, 321 - 338.

Embretson, S. Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 1983, *93*, 179 - 197.

Fiske, D. W. The constraints on intra - individual variability in test responses. *Educational and Psychological Measurement*, 1957, *17*, 317 - 337. (a)

Fiske, D. W. An intensive study of variability scores. *Educational and Psychological Measurement*, 1957, *17*, 453 - 465. (b)

Fiske, D. W. Items and persons: Formal duels and psychological differences. *Multivariate Behavioral Research*, 1968, *3*, 393 - 401.

Fiske, D. W., & Rice, L. Intra - individual response variability. *Psychological Bulletin*, 1955, *52*, 217 - 250.

Frank, B. Stability of questionnaire responses. *Journal of Abnormal and Social Psychology*, 1936, *30*, 320 - 324.

French, J. W. *Extended Range Vocabulary Test*. Princeton, New Jersey: Educational Testing Service, 1962.

Ghiselli, E. E. Differentiation of individuals in terms of their predictability. *Journal of Applied Psychology*, 1956, *40*, 374 - 377.

Ghiselli, E. E. Moderating effects and differential reliability and validity. Journal of Applied Psychology, 1963, 47, 81 - 86.

Glaser, R. A methodological analysis of the inconsistency of response to test items. Educational and Psychological Measurement, 1949, 9, 727 - 739.

Glaser, R. The application of the concepts of multiple operation measurement to the response patterns on psychological tests. Educational and Psychological Measurement, 1951, 11, 372 - 382.

Glaser, R. The reliability of inconsistency. Educational and Psychological Measurement, 1952, 12, 60 - 64.

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

Goldberg, L. R. The reliability of reliability: The generality and correlates of intra - individual consistency in responses to structured personality inventories. Applied Psychological Measurement, 1978, 2, 269 - 291.

Goldberg, L. R., & Jones, R. R. The reliability of reliability: The generality and correlates of intra - individual consistency in responses to structured personality inventories. Oregon Research Institute Monograph, 1969, 9, No. 2.

Goldberg, L. R., & Rust, R. M. Intra - individual variability in the MMPI-CPI common item pool. British Journal of Social and Clinical Psychology, 1964, 3, 145 - 147.

Greene, R. L. An empirically derived MMPI Carelessness Scale. Journal of Clinical Psychology, 1978, 34, 407 - 410.

Greene, R. L. Response consistency on the MMPI: The TR Index. Journal of Personality Assessment, 1979, 43, 69 - 71.

Hambleton, R. K., & van der Linden, W. J. Advances in item response theory and applications: An introduction. Applied Psychological Measurement, 1982, 6, 373 - 378.

Hanley, C. The "difficulty" of a personality inventory item. Educational and Psychological Measurement, 1962, 22, 577 - 584.

Hathaway, S. R., & McKinley, J. C. The Minnesota Multiphasic Personality Inventory Manual (revised), New York: Psychological Corporation, 1967.

Helmes, E. A multidimentional approach to personality inventory responding. Unpublished doctoral dissertation, University of Western Ontario, 1978.

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

Hendel, D. D., & Weiss, D. J. Individual inconsistency and the reliability of measurement. Educational and Psychological Measurement, 1970, 30, 579 - 594.

Holden, R. R., Helmes, E., Fekken, G. C., & Jackson, D. N. The multidimensionality of person reliability: Implications for interpreting individual item responses. Paper presented at the Annual Convention of the Canadian Psychological Association, Toronto, June, 1981.

Horn, D. Intraindividual variability in the study of personality. Journal of Clinical Psychology, 1950, 6, 43 - 47.

Jackson, D. N. A threshold model for stylistic responding. Paper presented at the American Psychological Association, San Francisco, September, 1968.

Jackson, D. N. A sequential system for personality scale development. In C. D. Spielberger (Ed.), Current topics in clinical and community psychology. (Vol 2). 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 (revised edition). Port Huron, Michigan: Research Psychologists Press, 1974.

Jackson D. N. The appraisal of person reliability. Paper presented at the Multivariate Society Meeting, State College, Pa., November, 1976.

Jackson, D. N. Jackson Vocational Interest Survey Manual, London, Canada: Research Psychologists Press, 1977.

Jackson, D. N. *Threshold model for personality assessment.* Paper presented at the International Conference on Personality Measurement, Bielefeld, West Germany, 1982.

Johnson, J. A. The "self - disclosure" and "self - presentation" views of item response dynamics and personality scale validity. *Journal of Personality and Social Psychology*, 1981, *40*, 761 - 769.

Kuiper, N. A. Convergent evidence for the self as a prototype: The "inverted - U RT effect" for self and other judgments. *Personality and Social Psychology Bulletin*, 1981, *7*, 438 - 443.

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.

Layton, W. L. The variability of individuals' scores upon successive testings on the Minnesota Multiphasic Personality Inventory. *Educational and Psychological Measurement*, 1954, *14*, 634 - 640.

Lentz, T. F. Reliability of the opinionaire technique studied intensively by the retest method. *Journal of Social Psychology*, 1934, *5*, 338 - 364.

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

Lumsden, J. Tests are perfectly reliable. *British Journal of Mathematical and Statistical Psychology*, 1978, *31*, 19 - 26.

Mauger, P. A. *The test - retest reliability of persons: An empirical investigation utilizing the MMPI and the PRF.* Unpublished doctoral dissertation, University of Minnesota, 1972.

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

Mills, W. W. *MMPI profile pattern and scale reliability throughout four years of college attendance.* Unpublished doctoral dissertation, University of Minnesota, 1954.

Mitra, S. K., & Fiske, D. W. Intra - individual variability as related to test score and item. *Educational and Psychological Measurement*, 1956, *16*, 3 - 12.

Morf, M. E., & Jackson, D. N. An analysis of two response styles: True responding and item endorsement. *Educational and Psychological Measurement*, 1972, *32*, 329 - 353.

Neprash, J. A. The reliability of questions in the Thurstone Personality Inventory. *Journal of Social Psychology*, 1936, *7*, 239 - 244.

Payne, F. D. Relationships between response stability and item endorsement, social desirability and ambiguity in the MMPI and CPI. *Multivariate Behavioral Research*, 1974, *9*, 127 - 148.

Pepper, L. J. *The MMPI: Initial test predictors of retest changes.* Unpublished doctoral dissertation, University of North Carolina, 1964.

Pintner, R., & Forlano, G. Four retests of a personality inventory. *Journal of Educational Psychology*, 1938, *29*, 93 - 100.

Raine, W. J., & Hills, J. D. Measuring intra - individual variability within one testing. *Journal of Abnormal and Social Psychology*, 1959, *58*, 264 - 266.

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

Rogers, T. B. *Evaluation of a threshold theory for personality assessment.* Unpublished doctoral dissertation, University of Western Ontario, 1970.

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. (a)

Rogers, T. B. Toward a definition of the difficulty of a personality item. Psychological Reports, 1973, 33, 159 - 166. (b)

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

Rogers, T. B. Experimental evidence for the similarity of personality and attitude item responding. Acta Psychologica, 1978, 42, 21 - 28.

Rogers, T. B., Kuiper, N. A., & Kirker, W. S. Self-reference and the encoding of personal information. Journal of Personality and Social Psychology, 1977, 35, 677 - 688.

Rogers, T. B., Kuiper, N. A., & Rogers, P. J. Symbolic distance and congruity effects for paired comparisons judgements of degree of self-reference. Journal of Research in Personality, 1979, 13, 433 - 449.

Schneiderman, W. A. A personality dimension of consistency versus variability without the use of self-reports or ratings. Journal of Personality and Social Psychology, 1980, 39, 158 - 164.

Schofield, W. Changes in responses to the Minnesota Multiphasic Personality Inventory following certain therapies. Psychological Monographs, 1950, 64, (5, Whole No. 311).

Schubert, D. S. P. Increase of personality response consistency by prior response. Journal of Clinical Psychology, 1975, 31, 651 - 658.

Schubert, D. S. P., & Fiske, D. W. Increase of item response consistency by prior item response. Educational and Psychological Measurement, 1973, 33, 113 - 121.

Smith, M. A note on stability in questionnaire response. American Journal of Sociology, 1933, 38, 713 - 720.

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

Underwood, B., & Moore, B. S. Sources of behavioral consistency. Journal of Personality and Social Psychology, 1981, 40, 780 - 785.

Weksel, W., & Ware, E. E. The reliability and consistency of complex personality judgments. Multivariate Behavioral Research, 1967, 2, 537 - 541.

Whitely, S. E. Individual inconsistency: Implications for test reliability and behavioral predictability. Applied Psychological Measurement, 1978, 2, 571 - 579.

Wiggins, J. S. Substantive dimensions of self-report in the MMPI item pool. Psychological Monographs, 1966, 80, (22, Whole No. 630).

Windle, C. Test-retest effect on personality questionnaires. Educational and Psychological Measurement, 1954, 14, 617 - 633.

Windle, C. Further studies of test-retest effect on personality questionnaires. Educational and Psychological Measurement, 1955, 15, 246 - 253.

Appendix A

Calculations of Threshold Model Parameters and of Chance

Threshold: The threshold was calculated by determining the point at which an individual had endorsed some critical proportion of a subset of items rank-ordered by social desirability scale value. In particular, the proportion of true responses was calculated for the 80 items lowest in desirability. If this proportion met the .5 criterion, the mean social desirability scale value of the items was taken to define the threshold. If the proportion of true responding was less than .5, the item lowest in social desirability scale value was dropped and the non-included item next highest in social desirability was added to the subset. The proportion of true responses was calculated over successive subsets of 80 items until the .5 criterion was met. The threshold was defined by the mean social desirability scale value of the group of 80 items which met the criterion. Both an ascending and a descending threshold (beginning with the 80 items highest in social desirability and proceeding in intervals descending in social desirability) were computed. The final threshold used in this research was the mean of the ascending and descending thresholds.

Salience: The salience parameter was defined as the biserial correlation between the items' social desirability scale value and an individual's pattern of endorsement, where a true response was designated as "1" and a false response as "0".

Chance: The number of correct predictions expected by chance was calculated as follows. Note that each model made predictions given the number of changed items. Thus, the number of changed items expected by chance for a subset of items equal in size to the total number of changed items was simply calculated using the overall proportion of change for the entire item set. For example, consider an individual changed 10 of 100 items. In any subset of 10 items, a proportion of .1 correct predictions (i.e., one correct item prediction) might be expected by chance.

Appendix B

Instructions for the Computer Task

In this study, you will be asked to respond to a series of statements which a person might use to describe himself or herself. You will read each statement and decide whether or not the statement describes you. Then you will indicate your response to the statement using the response apparatus placed on the table in front of you.

Begin by placing the index finger of your preferred hand on the "orange dot" on the response apparatus. Statements will appear on the screen in front of you, one at a time. Read the statement. If you agree with the statement or decide that it does describe you, answer TRUE by pressing the key marked "T". If you disagree with a statement or decide that it does not describe you, answer FALSE by pressing the key marked "F". Once you have answered, return your index finger to the "orange dot". The next statement will automatically appear on the screen.

Answer every statement either TRUE or FALSE, even if you are not completely sure of your answer.

Desirability Judgements

Instructions

In the question booklet you will find 96 statements that people might use to describe themselves. Each statement reflects certain tendencies, preferences or traits of people.

You are to judge whether a "True" response to a statement would reflect a desirable or an undesirable characteristic of people. For example, consider the three items below.

	EXTREMELY UNDESIRABLE				NEITHER DESIRABLE NOR UNDESIRABLE				EXTREMELY DESIRABLE
A. It doesn't affect me one way or another to see a child spanked.	1	[✓ 2]	3	4	5	6	7	8	9
B. I have a great curiosity about many things.	1	2	3	4	5	6	7	[✓ 8]	9
C. I like to feel sculptured objects.	1	2	3	4	[✓ 5]	6	7	8	9

This judge thought that a "True" response to statement A, "It doesn't affect me one way or another to see a child spanked", would reflect an undesirable characteristic in a person and circled 2. On the other hand, the judge thought that a "True" response to the statement "I have a great curiosity about many things" would reflect a desirable characteristic. Thus, 8 was circled. For statement C, "I like to feel sculptured objects", a "True" response was considered to reflect neither a desirable nor an undesirable characteristic.

In a similar manner, you are asked to judge whether a "True" response to the statements in the booklet would reflect a desirable or an undesirable characteristic of people. Remember that you are judging the desirability of these characteristics in other people, not in yourself. The scale for judging the desirability of each statement ranges from a value of 1, for extremely undesirable, to a value of 9, for extremely desirable. Please try to use all values and be sure to judge every statement.

	Extremely Undesirable 1	2	3	4	Neutral 5	6	7	8	Extremely Desirable 9
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	8	9
20.	1	2	3	4	5	6	7	8	9

	Extremely Undesirable				Neutral				Extremely Desirable
	1	2	3	4	5	6	7	8	9
21.	1	2	3	4	5	6	7	8	9
22.	1	2	3	4	5	6	7	8	9
23.	1	2	3	4	5	6	7	8	9
24.	1	2	3	4	5	6	7	8	9
25.	1	2	3	4	5	6	7	8	9
26.	1	2	3	4	5	6	7	8	9
27.	1	2	3	4	5	6	7	8	9
28.	1	2	3	4	5	6	7	8	9
29.	1	2	3	4	5	6	7	8	9
30.	1	2	3	4	5	6	7	8	9
31.	1	2	3	4	5	6	7	8	9
32.	1	2	3	4	5	6	7	8	9
33.	1	2	3	4	5	6	7	8	9
34.	1	2	3	4	5	6	7	8	9
35.	1	2	3	4	5	6	7	8	9
36.	1	2	3	4	5	6	7	8	9
37.	1	2	3	4	5	6	7	8	9
38.	1	2	3	4	5	6	7	8	9
39.	1	2	3	4	5	6	7	8	9
40.	1	2	3	4	5	6	7	8	9

	Extremely Undesirable				Neutral				Extremely Desirable
	1	2	3	4	5	6	7	8	9
41.	1	2	3	4	5	6	7	8	9
42.	1	2	3	4	5	6	7	8	9
43.	1	2	3	4	5	6	7	8	9
44.	1	2	3	4	5	6	7	8	9
45.	1	2	3	4	5	6	7	8	9
46.	1	2	3	4	5	6	7	8	9
47.	1	2	3	4	5	6	7	8	9
48.	1	2	3	4	5	6	7	8	9
49.	1	2	3	4	5	6	7	8	9
50.	1	2	3	4	5	6	7	8	9
51.	1	2	3	4	5	6	7	8	9
52.	1	2	3	4	5	6	7	8	9
53.	1	2	3	4	5	6	7	8	9
54.	1	2	3	4	5	6	7	8	9
55.	1	2	3	4	5	6	7	8	9
56.	1	2	3	4	5	6	7	8	9
57.	1	2	3	4	5	6	7	8	9
58.	1	2	3	4	5	6	7	8	9
59.	1	2	3	4	5	6	7	8	9
60.	1	2	3	4	5	6	7	8	9

	Extremely Undesirable				Neutral				Extremely Desirable
	1	2	3	4	5	6	7	8	9
61.	1	2	3	4	5	6	7	8	9
62.	1	2	3	4	5	6	7	8	9
63.	1	2	3	4	5	6	7	8	9
64.	1	2	3	4	5	6	7	8	9
65.	1	2	3	4	5	6	7	8	9
66.	1	2	3	4	5	6	7	8	9
67.	1	2	3	4	5	6	7	8	9
68.	1	2	3	4	5	6	7	8	9
69.	1	2	3	4	5	6	7	8	9
70.	1	2	3	4	5	6	7	8	9
71.	1	2	3	4	5	6	7	8	9
72.	1	2	3	4	5	6	7	8	9
73.	1	2	3	4	5	6	7	8	9
74.	1	2	3	4	5	6	7	8	9
75.	1	2	3	4	5	6	7	8	9
76.	1	2	3	4	5	6	7	8	9
77.	1	2	3	4	5	6	7	8	9
78.	1	2	3	4	5	6	7	8	9
79.	1	2	3	4	5	6	7	8	9
80.	1	2	3	4	5	6	7	8	9

	Extremely Undesirable				Neutral				Extremely Desirable
	1	2	3	4	5	6	7	8	9
81.	1	2	3	4	5	6	7	8	9
82.	1	2	3	4	5	6	7	8	9
83.	1	2	3	4	5	6	7	8	9
84.	1	2	3	4	5	6	7	8	9
85.	1	2	3	4	5	6	7	8	9
86.	1	2	3	4	5	6	7	8	9
87.	1	2	3	4	5	6	7	8	9
88.	1	2	3	4	5	6	7	8	9
89.	1	2	3	4	5	6	7	8	9
90.	1	2	3	4	5	6	7	8	9
91.	1	2	3	4	5	6	7	8	9
92.	1	2	3	4	5	6	7	8	9
93.	1	2	3	4	5	6	7	8	9
94.	1	2	3	4	5	6	7	8	9
95.	1	2	3	4	5	6	7	8	9
96.	1	2	3	4	5	6	7	8	9

END

20 10 3 18 4

FIN