DOCUMENT RESUME

ED 064 378 TM 001 623

AUTHOR Tyler, Thomas A.
TITLE Test Homogeneity and Response Stability.
INSTITUTION Chicago Univ., Ill. Dept. of Psychology.
SPONS AGENCY National Science Foundation, Washington, D.C.
PUB DATE [68]
NOTE 40p.; Based on a doctoral dissertation submitted to the Department of Psychology, University of Chicago

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Data Analysis; *Hypothesis Testing; *Interaction Process Analysis; *Personality Tests; *Psychological Studies; *Rating Scales; Response Mode; Test Reliability; Test Results

ABSTRACT

It was hypothesized that unstable subject-item interactions on personality scales would be associated with small psychological distances between subjects and items. It was further hypothesized that this relationship would be more demonstrable when the psychological test was more homogeneous. To clarify the rationale of the first hypothesis, an explanation was offered for the commonly observed greater stability of extreme items and extreme subjects. Both of the hypotheses were supported. For the second hypothesis, however, it was demonstrated that a single index of test homogeneity was only partially adequate. The separate contributions of items and of subjects to total test homogeneity had to be considered for the most meaningful interpretation of the results. Implications of the instability of responses associated with small subject-item distances were drawn for those operational definitions of subject variability and subject acquiescence which are based on personality test data. A suggestion is made for controlling acquiescence in personality tests. (Author)

TEST HOMOGENEITY AND RESPONSE STABILITY[1]

Thomas A. Tyler[2]

University of Chicago

Abstract

It was hypothesized that unstable subject-item interactions on personality scales would be associated with small psychological distances between subjects and items. It was further hypothesized that this relationship would be more demonstrable when the psychological test was more homogeneous. To clarify the rationale of the first hypothesis, an explanation was offered for the commonly observed greater stability of extreme items and extreme subjects. Both of the hypotheses were supported. For the second hypothesis, however, it was demonstrated that a single index of test homogeneity was only partially adequate. The separate contributions of items and of subjects to total test homogeneity had to be considered for the most meaningful interpretation of the results. Implications of the instability of responses associated with small subject-item distances were drawn for those operational definitions of subject variability and subject acquiescence which are based on personality test data. A suggestion was made for controlling acquiescence in personality tests.

TEST HOMOGENEITY AND RESPONSE STABILITY[1]

Thomas A. Tyler[2]

University of Chicago

In a series of papers, Glaser (1949, 1950, 1951, 1952) was able to demonstrate a close parallel between psychological measurement and psychophysical measurement. Particularly interesting was his demonstration (1951) that in ability measurement the stability of responses is a joint function of both subject and item parameters. For example, subjects whose raw scores were in the lowest quartile of an ability test were most inconsistent in their ability to solve the easiest group of items; those in the second quartile were most inconsistent in their abi ity to solve the next easiest set, etc. The joint relationship among subject parameters, item parameters, and inconsistency has always been clear in psychophysical measurement where the subject parameter is frequently defined in terms of the maximally unreliable measurement (e.g., the subject's absolute threshold is the point on the stimulus continuum where response instability or inconsistency is maximum).

Glaser's (1949) analysis of personality test data, however, was less satisfying than his analysis of ability data because it was indirect and to some extent ad hoc. For personality test data Glaser hypothesized a number of correlational relationships between total test scores and instability scores, depending upon the adequacy of the test for the range of the group being tested. To illustrate, Glaser hypothesized that a positively skewed score distribution would produce a negative relationship between raw score and the number of changed responses from test to retest. The logic of this statement assumes that a positive skew implies that there are relatively few items

with a high proportion of endorsement (p). Since these high p items would normally have been the items associated with unstable responses for the lowest scoring group (analogous to the instability observed about an absolute threshold) the prediction of a negative correlation follows.

Although the results of Glaser's correlational analysis were positive, Mitra and Fiske (1956) were unsuccessful in attempting to apply the model to personality test data in a more direct test, similar to Glaser's (1951) methodology for ability test data. At that time Mitra and Fiske questioned if the psychophysical model was an appropriate one for personality testing. More recent research (Fiske, 1966) suggests, however, that personality tests vary greatly in degree of unidimensionality or cumulative homogeneity; it could be hypothesized that the data used by Mitra and Fiske were not sufficiently homogeneous for the unidimensional threshold model. A casual inspection of the test content of their study tends to support this interpretation.

In the research reported below a more direct test was made of the applicability of the psychophysical model to dichotomous personality test data with special consideration given to the degree of homogeneity of the test data. The basic prediction of the psychophysical model for this research was that response changes from test to retest are associated with small psychological distances between the person and the item; this prediction is referred to later as the distance-stability hypothesis. The psychological distance between subject and item was defined as the absolute value of the difference between the person scale value and the item scale value; both items and subjects were scaled with a model proposed by Rasch (1960, 1966a, 1966b).

The more interesting prediction of this study was that the basic prediction, just given, would be more accurate or stronger for tests having a higher degree of cumulative homogeneity; this prediction is referred to later as the homogeneity hypothesis. The indices proposed by Fiske (1963, 1966) were used to assess the degree of homogeneity of test data.

An application

This conception of the problem and its importance to personality test research can be made clearer with reference to two consistently reported findings in the literature of personality tests: (1) Extreme items (i.e., items with very high or very low p) tend to be more stable than items with more moderate endorsement levels (Lentz, 1934; Frank, 1936; Mitra & Fiske, 1956; Goldberg, 1963). (2) Extreme subjects (i.e., subjects with very high or very low raw scores) tend to be more stable than subjects whose scores are nearer the median (Lentz, 1934; Frank, 1936; Mitra & Fiske, 1956). The remainder of this section will attempt to illustrate how the application of the psychophysical model to personality tests addresses both of these findings.

Insert Figure 1 About Here

In a typical personality test the distributions of both subjects and items have strong central tendencies. Subject distributions commonly are roughly normal. Items are frequently clustered about midscale with p values between .40 and .60, with few items of more extreme p. The continuum constructed in Figure 1 represents this typical condition, although the number of scaled subjects and scaled items is restricted for clarity. Items are represented as fixed points on the continuum with the extreme (high p) Item 1 to the left; the other extreme (low p) Item 5 is to the right. Moderate p Items 2, 3, and 4 are clustered about midscale. The modal

position for each subject (i.e., his threshold position) is represented by a dashed vertical line; low scoring subjects are to the left. Subjects, however, are considered to be variable over time in their positions on the continuum. The extent of their potential variation has been represented with bell-shaped curves erected about their threshold positions. The height of the curve over the continuum may be interpreted as the probability that the subject is at that point on the continuum at any one instant in time. Two extreme subjects (A and E) and several centrally scaled subjects (B, C and D) have been included to represent the distribution of a typical personality test.

From this geometric model the response is considered to be determined by the ordered relationship between the subject's momentary position on the continuum and the item's continuum position . If the subject's position is above (to the right of) the item's position an endorsement is indicated. Otherwise the item is rejected.

An explanation for the apparent stability of extreme subjects can be offered from this model assuming that the true variability is equal for all subjects. Under this assumption it can be shown that variability scores, defined as the number of changed responses from test to retest, is a function of the distribution of item points on the psychological continuum. Extreme Subject A has only one item (Item 1) within range of his typical variation over time. Thus, the ordered relationships between Subject A and the several items of the test, with the exception of item 1, are unlikely to change over time. Subject A would be expected to change only one response, at most, from a test to a retest on this five item test.

Subject C, on the other hand, has three items (Items 2, 3 and 4) within range of his potential variation. If this subject is at the low end of his range of variation during the first administration of the test and at the high end during the second administration, three response changes would be indicated. Thus, the number of subject response changes is predicted to be positively related to the number of items that are scaled at or close to the subject's position on the continuum. Since there typically are few extreme items in a personality test, extreme subjects have few items within their usual range of variation; it follows that extreme subjects would typically change few of their responses from test to retest and would appear to be more stable than would the other test subjects.

The explanation of the stability of extreme items is complementary to the one just given for subjects. In Figure 1, responses to extreme items 1 and 5 would be relatively stable since in each case only one subject (A and E respectively) could reasonably change positions with respect to these items. Item 3, however, has three subjects close to its position on the continuum, and potentially all three of these subjects could change their response to this item. Thus, the stability of an item is predicted to be negatively related to the number of subjects that are scaled at or close to the item's position on the continuum.

The Role of Homogeneity

From the discussion above, it would appear that the crucial variable in determining response consistency-inconsistency for a single subject-item interaction is the distance between the subject and the item on the continuum: small distances would be associated with inconsistent responses; large distances would be associated with stable responses. This conclusion follows from the hypothesized psychological continuum.

To apply the model, that is to use distance for prediction purposes, the subject-item distance must be estimated from an analysis of response data. The adequacy of any unidimensional scaling analysis for the reproduction of the psychological distance between the person and item is a complex function of several factors, including the dimensionality of item content, the dispersion of the subjects in some unknown dimensions, and the dispersion of the items in some unknown dimensions. Any computed estimate of location of an item or of a person on the continuum incorporates a potential for error--which can be estimated in some cases in terms of a standard error function.

What is important for the use of the distance concept in the interpretation of response consistency-inconsistency is that the error in estimating subject and item positions should be small relative to the total distance between the person and item. The high inter-person and inter-item correlations of the cumulatively homogenous test tend to be achieved by an increased dispersion of both items and persons along the continuum. With a homogenous test each subject has a few items that are close to his position, and many more items at increasing distances from his position. Likewise, each item has a few subjects close to its position, and many more subjects at increasing distances from the item position. From either an item or a subject point of view, then, there is a basis for differential prediction of inconsistency: one can predict which items are likely to be responded to inconsistently by any subject, or one can predict which subjects are going to respond inconsistently to a given item. Thus, the more a test approaches the homogeneity model, the more predictive computed distance will be for response consistency-inconsistency.

There are, additionally, two situations in which the distance concept will have some predictive value, even though a test (as a whole) may not be particularly homogeneous. The first case involves the test in which the items are clustered in one region of the continuum (mid scale usually). With such a test the stability of subjects can be predicted: those subjects most distant from the item cluster will be most stable. The greater the dispersion of subjects on the continuum, the better this prediction will be. However, the stability of items cannot be predicted. In the extreme case, if all items were at the same location, there would be no differences in distances from the items relative to the subjects and hence no basis for differential prediction.

The second situation in which the concept of distance will have some predictive value may seem to be an empty set since it involves the test in which subjects are clustered on the continuum. Since tests typically try to discriminate between people, finding the extreme case of all subjects at the same point is implausible. Arguing from the extreme case, however, suggests the the concept of distance would still predict which items would be most stable, i.e., those items most distant from the person cluster. Since all the subjects are at the same point on the continuum there would be no basis for differential predictions of subject stability. However, greater dispersion of items on the continuum produces better prediction of differential item stability, even if subject dispersion is low, and a test is not particularly homogeneous.

In summary then, the degree of test homogeneity is related to the prediction of consistency-inconsistency in three ways: (1) better overall prediction will be obtained from tests which are more homogeneous, that is,

with tests which have a good dispersion of both persons and items, (2) increasing dispersion of items on a test will contribute to better prediction of item stability, and (3) increasing dispersion of persons on a test will contribute to better prediction of person stability.

Finally, the model developed above relates only to response inconsistentcy which could be attributed to the subject's variability. Clearly, some changes in response from test to retest could be attributed to factors that have nothing to do with the psychological distance between the subject and the item. For example, the subject may simply misread the item on one occasion, or he may make an error in marking the answer sheet. This type of response inconsistency is beyond the scope of this study.

Insert Table 1 About Here

Method

Materials

Twenty-two dichotomously scored personality tests were administered on two occasions. These tests (and their respective identifying letters in Table 1) were: Thurstone's (1950) Dominance and Reflective Scales (A and B); twenty items each from Welsh's (1956) A and R scales (C and D); Jackson's (1966) Dominance, Inpulsivity, Need Achievement, and Social Acceptance Scales (E through H); Buss and Durkee's (1957) Hostility-Assault Scale (I); the Direct, Initiate, Lead and Persuade subscales from Fiske's Interpersonal Relations Inventory (J through M); the High Energy Level, Hurting Another, Completing a Task, Obtaining Power and Mastery, Ulterior Motives, Superior Ability, Social and Personal Responsibility, and Self-Direction subscales from Butt's Motivation for Dominating Behavior Scale (N through U); and Butt's Reported Strength of Dominant Acts (V). (For scales J through V see Butt and Fiske, 1968.)

The length of each of these scales is reported in Table 1. Complete item listing for scales A through D are given by Turner (1967); item listings for scales E through V are given by Butt (1968). The present study is based on a different analysis of part of the data of the studies by Butt and by Turner. (See also Butt and Fiske, 1968, and Turner and Fiske, 1968.)

Subjects

Test data were collected from two different groups of introductory psychology students at the University of Illinois at Chicago Circle. The first group, consisting of 40 males and 40 females, took scales A through D. The second groups of subjects, consisting of 76 males and 61 females, took scales E through V. Test-retest time in both cases was four weeks.

Assessing Cumulative Homogeneity

The degree of homogeneity of each test was assessed from first administration data. The following indices, suggested by Fiske (1963, 1966), were computed:

s^2_e/s^2_t The proportion of remainder variance (i.e., that part of the total test variance not associated with item and person means). The proportion of remainder variance should be as low as possible for high homogeneity. Fiske (1966) suggests that the complement of this index (i.e., $1 - s^2_e/s^2_t$) may be the best single index of test homogeneity.

r_{ii} An index of test internal consistency derived from KR-20 reliability estimates (r_{tt}) with the application of the Spearman-Brown formula inversely projecting to a one-item test (Guilford, 1954, Formula 14.18). Informally, r_{ii} may be viewed as an index of the degree to which items agree in their ordering of people, or as an average inter-item correlation. The most useful interpretation for this study is the degree to which the items taken together "spread" people out on the continuum. A substantial value of r_{ii} is a necessary but not sufficient condition for high test homogeneity.

r_{pp} The dual of r_{ii} for persons. Parallel to the index above, r_{pp} may be viewed informally as an index of the degree to which persons agree in their ordering of items, as an average inter-person correlation, or, more usefully in the present study, as an index of the degree to which persons "spread" the items out on the continuum. A substantial value of r_{pp} is a necessary but not sufficient condition for high test homogeneity.

The complement of the proportion of remainder variance ($1 - s^2_e/s^2_t$) is used in this study as a general index of test homogeneity. The two differential indices (r_{ii} and r_{pp}) are included to reflect the relative contribution of items and persons to total test homogeneity.

Scaling Persons and Items

Persons and items were scaled according to a model proposed by Rasch (1960, 1966a, 1966b) for dichotomously scored ability tests. A non-technical discussion of the model is provided by Wright (1968). With suitable translation of terms, such as person ability to person trait-strength and item difficulty to item trait-demand, the model can be applied to dichotomous personality tests.

The model assumes that each person and each item can be characterized by a construct parameter (call these parameters P and I, respectively). Using the trait of dominance as an example, if Person 1 is twice as dominant as Person 2, then $P_1 = 2P_2$. Likewise, if Item 1 demands twice as much dominance for a keyed response as Item 2, then $I_1 = 2I_2$. The probability that Person 1 endorses Item 1 is assumed to be the same as the probability that Person 2 endorses Item 2. The probability of an endorsement in any

subject-item interaction is assumed to be a function of P/I , such that $f(P_1/I_1) = f(P_2/I_2)$. Arbitrarily, Rasch chooses the simple function $f(P/I) = P/(I + P)$ which has the desirable features that the probability of endorsement is always between zero and one, whatever the values of I and P, and when I = P the probability of an endorsement is appropriately .50, and finally, the function is mathematically tractable.

From these assumptions it is possible to derive paradigms for the solution of I and P from test data. A computer program developed by Wright and Panchapakesan (1969) using an iterative maximum likelihood solution was used in this study. The logistic transforms of the solutions were used as scale values rather than the computed estimates of I and P.

Insert Table 2 About Here

Results

Preliminary Analyses

Homogeneity Indices

The observed values of the homogeneity indices from the first administration, together with N (the sample size) and n (the number of items) are presented in columns 2 through 6 of Table 1. The intercorrelations between the indices on the first administration and the stability of the indices over the two administrations are presented in Table 2. All coefficients in Table 2 are rank-order correlations. It should be noted that r_{pp} tends to be more highly correlated with $1 - s^2_e/s^2_t$ than is r_{ii}. That is, the differences between tests with respect to the general index of homogeneity is more a function of item spread than of person spread with the present sample of scales and subjects. It should also be stated that there are

consistent trends in the homogeneity coefficients from test to retest: s^2_e/s^2_t tends to go down (median .67 to .64); r_{pp} tends to decrease (median .13 to .12); while r_{ii} tends to increase (median .16 to .19). Similar trends of second test indices have been observed before (Fiske, 1966). The two indices r_{pp} and r_{ii} are correlated more evenly with $1 - s^2_e/s^2_t$ for the retest data, .70 and .59 respectively.

The interpretation of r_{pp} as an index of the "spread" of items on the continuum is supported by an observed correlation of .98 between r_{pp} and the variance of the means of items for the first administration and .97 for the second administration. Likewise, the interpretation of r_{ii} as an index of the "spread" of persons on the continuum is supported by an observed correlation of .93 between $r_{i'}$ and the variance of the means of persons for the first administration and .94 on the second administration.

Insert Figure 2 Above Here

Rasch Scaling

Illustrative scale values are presented in Figure 2 (data from Scale M). This particular scale is slightly atypical in that the item scale values fall into three rather distinct clusters, but Scale M was chosen for its pertinence to a later discussion. Note that while each arrow on the right generally refers to the continuum location of a single item, each arrow on the left refers to a given raw score group and all subjects in it. Subjects who respond in the keyed direction to no items, or all items, and items that are endorsed by no subjects, or all subjects, cannot be scaled in the model and are omitted from the analysis. The scaling N of column 7 and the item n of column 12 in Table 1 reflect these omissions.

The stability (r) of the scale values for items, computed for six tests selected to provide a subjectively stratified sample, ranged from .87 to .99 with a mean of .94. The stability of scale values for subjects is essentially the same as test-retest estimates of score reliability due to a norming artifact of the scaling procedure. Test-retest r for the 22 scales of this study ranged from .48 to .89 with a mean of .73.

Tests of the Distance-Stability Hypothesis

Subject Instability

In the first analysis from a subject orientation a comparison was made between the number of subjects whose response inconsistency was associated with items near their position on the continuum, and hence supportive of the distance model, and the number of subjects whose response inconsistency was associated with items more distant from their position on the continuum, and hence contrary to this hypothesis. For each subject on each of the 22 scales, distances were computed to each of the n items of the scale by taking the absolute value of the difference between the subject scale value and the item scale value. These distances were divided into two distributions for each subject; one distribution of the distances associated with the subject's stable response patterns over the two administrations, and one distribution of the distances associated with the subject's response changes (in either direction) over the two administrations. Medians were computed for both of these distributions.

Since smaller distances should be associated with unstable response patterns, the median for the change distribution should be smaller than that for the stable distribution; subtracting the median for the change distribution from the median for the stable distribution for each subject should produce a positive sign to conform to the distance prediction.

The results of this analysis are presented in Table 1. Column 8 contains the reduced sample size (non-scalable subjects and subjects with no changed responses are omitted). Column 9 indicates the frequency of subjects for whom the prediction was sustained, column 10 the contrary cases. Not tabled are the zero differences (ties) which occured only in scales B (9 ties), D (1 tie), and S (4 ties). The critical ratio for the binomial test of the sign distribution (normal approximation corrected for continuity but not for ties) is given in column 11. All of the results are in the correct direction; three tests were not significant, two were significant at the .05 level, four at the .01 level, and 13 are significant at the .001 level.

For each scale, inspection of the distributions of signs at each possible score level indicates that the results are consistent at all score levels and that the obtained results cannot be attributed to differential effects for extreme or modal scoring subjects.

Although the above analysis demonstrates the tendency of subjects to change their responses to "near" items more commonly than to "far" items a second analysis was performed to provide an estimate of the strength of the effect. Subjects that are "close to" the most items should be those with the largest number of changed response patterns. An index of average distance for each subject to all items was formed by computing the mean of the absolute differences between the scale value of each subject and the n item scale values for each of the 22 tests. On each test a product-moment correlation over subjects was computed between this index of average distance and the number of changed responses. The sign of the correlation was changed to indicate average closeness rather than average distance. Over the 22 scales these correlations ranged from -.17 to .46 with a mean of .25. Seven of these correla-

tions (including the single negative value) were not significantly different from zero, two were significant at the .05 level, and 13 were significant at the .01 level. If the squared correlation coefficient were to be interpreted as the proportion of variance in the number of changed responses that can be explained by the index of subject distance to the items, a mean $\underline{r}^2$ of .08 is obtained. Although the results of this analysis were significant and positive in support of the distance stability hypothesis the low value of the correlations suggests little predictive value in terms of which subjects are going to be inconsistent. An analysis from an item point of view in the next section suggests that item stability is much more predictable.

Item Instability

In the first analysis from an item orientation, a comparison was made between the number of items for which the basic distance prediction was supported and the number of contrary items, parallel to the first analysis for subjects above. A distribution for change response patterns and a distribution for stable response patterns was formed for each item; means were computed for each distribution. The mean of the change distribution was subtracted from the mean of the stable distribution; the sign of this difference should be positive to conform to the prediction for items.

The results of this analysis are presented in Table 1. Column 12 gives the number of items in the test (reduced by one non-scalable item in the case of Scale Q). The frequency of items for which the prediction was sustained is presented in column 13. No ties were observed. The test ratio in column 14 is: (frequency positive - frequency negative)/$\sqrt{\underline{n}}$. The significance state-

ment for the distribution of signs in column 14 is based on an exact binomial test. Seven of the outcomes (including the single negative instance) were not significant, seven were significant at the .05 level, two at the .01 level, and six were significant at the .001 level.

There were 344 items used in these analyses and none common to two scales. Of these items, 302 conformed to the prediction, and 42 did not. With a two tailed normal approximation to the binomial, this difference in distribution of signed differences is significant at the .001 level.

Although the above analysis demonstrates that "near" subjects more commonly change their responses to an item than "far" subjects do, a second analysis was performed to provide an estimate of the strength of the effect. Items that are "close to" the most subjects should be those with the largest number of changed response patterns. An index of average distance from each item to all subjects was formed by computing the mean of the absolute differences between the scale value of the item and each of the N subject scale values. A product-moment correlation was computed between this index of distance and the number of response changes to each item. The sign of the correlations ranged from -.08 to .92 with a mean of .60. Ten of these correlations (including the single negative value) were not significantly different from zero, three were significant at the .05 level, and nine were significant at the .01 level. The squared coefficient, however, averaged substantially higher for items (mean r^2 = .43) than for subjects (mean r^2 = .08).

Over all Comparisons

Subject-Item Distance with Change-Stable Correlation

Considering each test separately, there are N times n subject-item interactions. Each of these interactions has an associated distance value and an associated dichotomous change-stable index. Computed biserial correlations between these variables ranged over tests from .06 to .32 with a mean of .23. These coefficients are all in the correct direction and indicate, again, that small subject-item distance on the continuum tend to be associated with unstable subject-item interactions.

Algebraic Distance and Response Stability.

Finally, the relationship between algebraic distance and response stability was investigated. The total N times n algebraic distances for each test were grouped by simply dividing the total range into eight intervals; the two intervals at each end of the distribution were combined since there were few observations in these intervals. If small distance is associated with response instability, then the interval containing the zero distance interactions should have the greatest proportion of changed responses; the adjacent interval closest to the zero distance should have the next greatest proportion of changed responses, etc. The process can be continued until each of the arbitrary intervals is ranked in terms of expected proportion of changed responses.

This process implies a graph relating algebraic distance to proportion of changed responses which would be roughly normal in shape; the proportion of changed responses would be maximum in the interval containing the zero distance and monotonically decreasing in either direction from this point. Most of the graphs followed this general form. When the observed proportion

of changed responses for the several intervals were computed and ranked, the rank-order correlations between the observed and expected ranks (a rough goodness-of-fit test) ranged from .32 to 1.00 (mean .82). Eleven of the rank--order coefficients were .94 or higher (significant at the .02 level).

Even though the interval containing the zero distance was not always the interval with the greatest proportion of changed responses, the proportion of changed responses in the interval containing zero distance ranged from .18 to .39 (mean .31). Smaller frequencies in the extreme intervals make a scale-by-scale computation of the proportion of changed responses in these intervals difficult to interpret; however, the extreme intervals over the 22 scales included a total of 4,677 person-item interactions, of which 551 were unstable. Thus, the over-all proportion of changed responses in the extreme distance ranges was .12.

Tests of the Homogeneity Hypothesis

Although the results of the analysis by items and by subjects, as well as the general analysis as presented above support the basic distance-stability prediction of the study, the outcomes over the 22 scales, as expected, are not uniform. According to the second prediction of the study, the better or stronger outcomes should be associated with those scales with a higher degree of cumulative homogeneity.

To test this prediction the test ratios of columns 11 and 14 of Table 1 were taken as indices of the degree of support for the basic prediction relating unstable subject-item interactions to the distance between subject and item on the continuum. The complement of the proportion of error variance

(i.e., $1 - s_e^2/s_t^2$) was correlated (rank-order) with the two outcome test ratios. Additionally, the rank-order correlations between r_{ii} and r_{pp} and the two test ratios were also computed. These correlations are presented in Table 3.

Insert Table 3 About Here

The correlations between the quality of prediction and the general index of homogeneity for both subject and item orientations are positive, as predicted. The subject oriented correlation is substantial, but not significant with a two-tailed test. The item oriented coefficient is near zero.

The correlations with the two differential indices of homogeneity (r_{ii} and r_{pp}) form an interesting pattern. The index r_{pp} is significantly correlated with the outcome of the subject oriented prediction, while r_{ii} is slightly negatively correlated with the subject oriented prediction. This pattern is reversed for the item oriented prediction. This result, which has important implications for the practice and theory of psychological testing, will be discussed in more detail later.

Table 3 uses the data from the outcomes of the frequency comparisons of Table 1; a parallel analysis was also performed with the outcomes of the correlational analyses. The results of these analyses were essentially the same as presented in Table 3: i.e., consistent positive correlations with the general index of homogeneity and the outcome index, and a reversal between the two differential indices depending upon whether the analysis was from the subject or the item orientation.

Discussion

Subject-item Distance

First, it can be stated that the concept of psychological distance between person and item (operationalized here as the absolute difference between the two scale values) has explanatory power for the stability of subject-item interactions. The tests of the basic prediction in this study generally were significant statistically, although of little predictive value in the case of the subject orientation.

The concept of distance would appear to have greater predictive value from an item orientation. For example, in one scale about 85 per cent of the variance in the number of changed responses per item could be explained by the average distance on the continuum between the item and the N subjects. Not too surprising was the finding that average item-to-subjects distance correlated .99 with the traditional index of item variation ($\sqrt{pq}$) since the two indices are artifactually related. Although distance is essentially no more predictive than $\sqrt{pq}$, distance has more conceptual appeal in terms of explaining the phenomena of response inconsistency in personality tests.

An argument can be made that the stability of extreme items can be explained in terms of a statistical artifact of the proportion of endorsement. For example, if only 10 out of 100 subjects endorse an item on the first occasion, and a like proportion endorse it on the second occasion, then, at most, 20 percent of the subjects can change their response to the item. On the other hand, if $p = .50$ on both occasions, 100 percent of the subjects could change their responses over two administrations. Thus, response insta-

bility might seem to be simply a function of item popularity. Prediction based on a "ceiling" effect may have some validity, but in a strict sense it is not a pure prediction model since prior knowledge is required of second administration test statistics. Additional evidence can be offered to suggest the distance-stability model is adequate for prediction of stability without recourse to second - test data.

Directly analogous to Glaser's plots of response instability for different ability groups over items of increasing difficulty, the subjects from Scale M (Figure 2) were divided into thirds; the lowest scoring third (T_1, N = 55) had raw scores from 0 through 6, the middle third (T_2, N = 45) from 7 through 10, and the highest scoring third (T_3, N = 37) from 11 through 18. Items were divided into three sets of decreasing p. The cutting points for both items and subjects were made on the basis of an ad hoc inspection of Figure 2.

Insert Figure 3 About Here

Plotting the proportion of changed responses against the ranked scale value for item sets demonstrates that the peak of inconsistency over occasions for each group of subjects corresponds to the appropriate item set. Thus, if a personality scale is sufficiently homogeneous, the contribution to response instability of the interaction between subject and item parameters can be clearly displayed. As a general rule, however, graphs constructed for the other scales of this study were not as symmetrical or "pretty" as Figure 3, even with the advantage of ad hoc grouping, but they too supported the hypothesis.

The superfluousness of the "ceiling" argument can also be revealed with a separate consideration of extreme items, although most of the items of the present study were not in the range typically defined as extreme (i.e., most had endorsement values between .10 and .90). A restricted analysis was performed using the most popular and least popular items of the study and the comparable extreme subjects for each test. Combining over tests, the mean proportion of the changed responses for those subjects scoring in the lowest 10 per cent when responding to the 2 or 3 (depending upon test length) most popular items, was .29. A parallel analysis for the highest scoring group on the least endorsed items indicated a proportion of .30 changed responses. For comparison, the subjects in the score interval containing the median changed .29 of their responses to items with p values around .50.

In contrast, the subjects at the high end of the continuum changed only .12 of their responses to items at the low end of the continuum, a numerically identical proportion of changed responses was observed for subjects at the low end of the continuum, responding to items at the high end.

In an even more restricted analysis, items that conformed to the typical definition of extreme were selected--a total of 13 items. The one subject raw score group nearest each of these 13 items was then considered for a total of only 21 subject-item interactions. Of these 21 interactions, however, 6 changed responses were observed for a mean proportion of changed responses of about .29.

In every analysis of this study, subjects typically changed about 30 percent of their responses to those items near their position on the continuum; this statement holds regardless of the raw score of the subject. It would seem that the stability of extreme scoring subjects, or the stability of extreme items is not essentially different from that of other subjects or items. Extreme subjects are apparently equally variable in their responses to extreme items in their own range on a scale as modal scoring subjects are to those items in their range. The greater frequency of response changes observed for modal scoring subjects can be attributed to the greater frequency of items with endorsements of .40 to .60 on typical personality tests. A parallel statement can be made for items.

Cumulative Homogeneity

The second prediction of the study that the basic relationship between psychological distance and response stability would be more strongly confirmed with more homogeneous tests was only weakly supported when the general index of test homogeneity ($1-s_e^2/s_t^2$) was considered (see Table 3). An examination of the reasons why this result was not stronger will also reveal some of the virtues of the cumulative homogeneity model for personality test research.

In order for a test to be homogeneous, it is necessary to have not only a dispersion of person scores (which is essential in any measure of individual differences), but it is also necessary to have a dispersion of item endorsements. In terms of the indices of this study substantial values of both r_{ii} and r_{pp} are necessary for homogeneity; in one sense it might be said that both r_{ii} and r_{pp} contribute to general or total homogeneity. Because the basic distance-stability prediction was tested from two different reference positions (first from a subject orientation, then from an item orientation), it was not essential that the general index of test homogeneity be high in order to support the prediction in any of the separate analyses.

To illustrate, consider the method in which the index was computed for the subject orientation. After the scale values were obtained for subjects and items each subject was analyzed individually. The prediction was that a subject was more likely to change his responses to items near his position on the continuum than to more distant items. In principle, the subject's raw score (or scale position) is irrelevant to the prediction; if every subject had earned exactly the same raw score (hence no dispersion of persons, $r_{ii} = 0$), the prediction should still be supported. Thus, the

value of r_{ii} (to the extent that r_{ii} is independent of r_{pp}) is irrelevant to the support of the basic distance-stability prediction, when the orientation is from the frequency comparison of supporting or non-supporting subjects.

On the other hand, if the items are not reasonably dispersed on the continuum (i.e., if r_{pp} is low), the distance from any subject to the several items of the item cluster will tend to be nearly equal and a restriction of range effect tends to conceal the basic relationship. Thus, when the distance-stability test is from the subject orientation, the effect appears strongest in those scales with a good dispersion of items, i.e., high r_{pp}. The dispersion of subjects, r_{ii}, is essentially unrelated, and the general index is related only to the extent that it reflects the contribution of r_{pp}.

The relationship of the basic prediction through the item orientation, line 2 of Table 3, is directly reversed from that of the discussion above. The dispersion of items, r_{pp}, is essentially unrelated to the support of the prediction and the general index is related only to the extent that it reflects the contribution of the dispersion of persons, r_{ii}. (The contribution being low in these data as Table 2 indicates). The dispersion of persons is the essential aspect of demonstrating the distance-stability effect from an item orientation as computed here.

Thus, at least for the study of response instability, two indices of test adequacy are required. The traditional index of internal consistency (r_{tt} or KR-20 for dichotomous data) does not reflect the dispersion of items, or the extent to which persons agree in their ordering of items, and is not a complete description of test quality.

It is most important to state that the relationship between distance and stability can be more readily demonstrated as the homogeneity of the test increases. (Paradoxically, the total contribution of small distances to unreliable measurement should be much greater in heterogeneous tests as items for heterogeneous tests tend to be selected on the basis of p values around .50.)

Implications

The finding that response instability in personality measurement is a joint function of subject and item score related parameters has implications in several areas of psychological measurement.

Subject Variability

Several investigators have tried to study individual differences in variability using the number of changed responses on psychological tests as an operational measure of variability (see Fiske and Rice, 1955, for a review). Since the number of changed responses for a subject on a psychological test has been demonstrated to be at least partially a function of the number of items that are near his position on the construct, and since most psychological tests tend to have distributions of items that are clustered about mid-scale, it is clear that at least part of the inter-test correlation of response changes can be attributed to the extent that the subject is scaled in the same area of the several tests. For example, if two measures are highly related, it is likely that a subject who is centrally scaled on one will be centrally scaled on the other and have relatively unstable response patterns to both instruments. The more highly related the measures, the more likely instability scores will correlate over tests.

Controlling Acquiescence in Practice

Several personality scales have attempted to control acquiescence by keying half of the items True, and the other half False. Cronbach (1942) has stated that acquiescence becomes operative when the subject is in doubt as to the right answer for him. One situation in which the subject may be uncertain about his response is when the item is too close to his position on the continuum. That is, acquiescence may be a set that is associated with small subject-item distances. To illustrate how this effect can be a factor, consider an extreme case in which the alternate keying results in all of the items on the lower half of the continuum being keyed True. In this situation a subject in the upper regions of the continuum has no True-keyed items near his position on the continuum and hence, would not exhibit acquiescence from this source (he may still acquiesce because he doesn't understand the question, etc.). The opportunity to acquiesce would be inflated for lower scoring subjects. Thus, as a first approximation for controlling acquiescence in personality scales, item keying should be systematically reversed in each region of the continuum.

Subject Acquiescence

Acquiescence, as a variable of individual differences, has been measured by counting the number of Yes-Yes (Agree-Agree, etc.) response patterns to subsequent administrations of an item and its grammatical reversal. Samelson (1964) and Rorer (1966) have emphasized the importance of equivalent psychological scaling of the item and its reversal for the identification of a true acquiescence response. Even if the item and its reversal are equivalently scaled, however, normal subject

variability may result in an acquiescent response pattern if the item and its reversal are both scaled near the subject's position on the continuum. Thus, subjects located on the continuum near item clusters will appear to be more acquiescent simply on the basis of normal variability. Therefore, acquiescence scores based on the reversal paradigm may be, in part, related to a joint function of the distributions of item endorsements and of the subjects' scale positions.

Homogeneous versus Heterogeneous Tests

Small subject-item distances produce unstable, and hence, unreliable measurement. Most existing personality scales are heterogeneous and tend to crowd a large number of items into the center of the psychological continuum (with p values between .4 and .6). The effect of this practice is to create a large number of unstable, unreliable measurements for subjects with modal scores. Extreme scoring subjects are measured with a greater degree of reliability, but with a decreasing amount of differentiation, and hence, with decreasing validity. Of course, it can be shown that even if test items tend to correlate weakly, this strategy produces maximum validity (since there are typically more modal subjects). If one subscribes to this logic, it is better to measure the ten subjects at midscale with higher validity and the two extreme subjects with lower validity since the total validity over the scale is then largest, as Cronbach and Warrington (1952) suggest for ability tests.

For personality research, however, the surrender of uniform validity and reliability over subjects to increase total validity is a false economy. In many cases it is the extreme subjects in the sample that provide the clearest indications for the support or rejection of an experimental hypothesis or a personality theory and good practice would require that

the extreme subjects be measured as adequately as modal subjects.

There is no question but that the heterogeneous test is the best predictor of a multidimensional criteria for practical applications. For basic research in personality measurement, and particularly for measurement in the service of personality theory, the cumulative homogeneity model is the superior model. The researcher can have greater confidence in his measures and can begin to exert control over some of the extraneous factors that confound the measure of personality constructs. Clearly, as the results of this study indicate, any investigation of the stability of measurement, or any study that employs the test-retest paradigm, must explicitly consider the issue of cumulative homogeneity.

Footnotes

[1] This report is based on a doctoral dissertation submitted to the Department of Psychology, University of Chicago. The support and advice of Donald W. Fiske is gratefully acknowledged. The contributions of D. Susan Butt, Paul Greene, Nargis Panchapakesan, Laura Rice, William Rucker, Jack Sawyer, Castellano Turner, Benjamin Wright, and Robert Wyer are sincerely appreciated. This research was supported in part by a Cooperative Fellowship granted by the National Science Foundation and in part by Grant No. GS-1060 from the National Science Foundation.

[2] Now at the Counseling and Testing Center, Southern Illinois University.

References

Buss, A. H., and Durkee, A. An inventory for assessing different kinds of hostility. Journal of Consulting Psychology, 1957, 21, 343-349.

Butt, D. S. A comparison of measurement strategies in developing scales for dominance. Unpublished doctoral dissertation, University of Chicago, 1968.

Butt, D. S., and Fiske, D. W. Comparison of strategies in developing scales for dominance. Psychological Bulletin, 1968, 70, 505-519.

Cronbach, L. J. Studies of acquiescence as a factor in the true-false test. Journal of Educational Psychology, 1942, 33, 401-415.

Cronbach, L. J., and Warrington, W. G. Efficiency of multiple-choice tests as a function of spread of item difficulties. Psychometrika, 1952. 17, 127-147.

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

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

Fiske, D. W., and 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.

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

Glaser, R. Multiple operation measurement. Psychological Review, 1940, 57, 241-253.

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.

Guilford, J. P. Psychometric methods, New York: McGraw-Hill, 1954.

Jackson, D. N. A modern strategy for personality assessment: the personality research form. Department of Psychology, University of Western Ontario, London, Canada, Research Bulletin No. 30, October, 1966.

Lentz, T. F., Jr. The reliability of opinionnaire technique studied intensively by the retest method. Journal of Social Psychology, 1934, 5, 338-364.

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

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

Rasch, G. An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology. 1966, 19, Part 1, 49-47. (a)

Rasch, G. An individualistic approach to item analysis. In Readings in mathematical and social sciences. Edited by Lazarsfeld and Henry. Chicago, Science Research Associates, Inc., 1966, 89-107. (b)

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

Samelson, F. Agreement set and anticontent attitudes in the F-scale: A reinterpretation. Journal of Abnormal and Social Psychology, 1964, 68, 338-342.

Thurstone, L. L. Examiner manual for the Thurstone Temperament Schedule. Chicago: Science Research Associates, 1950.

Turner, C. B. Stability and homogencity of test responses as related to appropriateness of response processes. Unpublished doctoral dissertation, University of Chicago, 1966.

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

Welsh, G. S. Factor Dimensions A and R. In G. S. Welsh and W. G. Dahlstrom (Eds.), Basic Readings on the MMPI in Psychology and Medicine. Minneapolis: University of Minnesota Press, 1956, Pp. 264-281.

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

Wright, B. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1968, 3, 85-101.

TABLE 1

TEST INDICES AND RATIOS FROM SUBJECT AND ITEM ORIENTED PREDICTIONS

	Homogeneity Analysis Results					Scaling	Results of Subject Analysis				Results of Item Analysis			
Scale	N	n	s_e^2/s_t^2	r_{pp}	r_{ii}	N	N	+	-	Test Ratio	n	+	-	Test Ratio
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	80	20	.672	.108	.222	78	73	57	16	4.68***	20	18	2	3.58***
B	80	20	.810	.086	.067	80	77	42	26	1.82	20	9	11	-0.45
C	80	20	.745	.065	.164	78	77	51	26	2.74**	20	20	0	4.47***
D	80	20	.797	.115	.052	80	77	46	30	1.72	20	14	6	1.79
E	137	20	.693	.107	.200	135	133	84	49	2.95**	20	19	1	4.02***
F	137	20	.766	.082	.129	136	134	90	44	3.89***	20	18	2	3.58***
G	137	10	.659	.188	.130	133	116	76	40	3.25**	10	10	0	3.16**
H	137	10	.589	.268	.159	135	116	87	29	5.29***	10	9	1	2.53*
I	137	10	.674	.098	.187	128	115	71	44	2.42*	10	8	2	1.90
J	137	21	.711	.124	.165	133	133	92	41	4.34***	21	21	0	4.58***
K	137	15	.691	.167	.135	132	125	97	28	6.08***	15	13	2	2.84*
L	137	9	.560	.261	.210	131	111	77	34	3.99***	9	8	1	2.33*
M	137	18	.666	.107	.229	128	127	88	39	4.26***	18	18	0	4.24***
N	137	10	.592	.175	.243	122	111	73	38	3.23**	10	9	1	2.53*
O	137	10	.695	.055	.191	96	86	51	35	1.62	10	7	3	1.26
P	137	10	.572	.354	.066	135	117	84	33	4.62***	10	9	1	2.53*
Q	137	10	.588	.255	.176	126	114	84	30	4.96***	9	9	0	3.00**
R	137	10	.692	.104	.159	107	96	71	25	4.59***	10	8	2	1.90
S	137	10	.638	.242	.102	136	120	85	31	4.55***	10	9	1	2.53*
T	137	10	.671	.141	.157	129	119	74	45	2.57*	10	9	1	2.53*
U	137	10	.758	.136	.038	130	118	87	31	5.06***	10	7	3	1.26
V	137	10	.594	.305	.099	135	113	79	34	4.14***	10	8	2	1.90

*p < .05

**p < .01

***p < .001

Table 2

STABILITY COEFFICIENTS AND FIRST ADMINISTRATION RANK INTERCORRELATIONS OF THE HOMOGENEITY COEFFICIENTS (N = 22 tests)

	1	2	3
1. $1 - s^2_e/s^2_t$	(.90)*		
2. r_{pp}	.80	(.85)	
3. r_{ii}	.29	-.20	(.86)

* Rank-order stability coefficients in diagonal

Table 3

RANK-ORDER CORRELATIONS BETWEEN QUALITY OF PREDICTION FOR SUBJECT AND ITEM ORIENTATIONS AND THE DEGREE OF TEST HOMOGENEITY (N = 22 tests)

Prediction	$1 - s^2_e/s^2_t$	Homogeneity Indices r_{pp}	r_{ii}
Subject Orientation	.40	.54*	-.09
Item Orientation	.11	-.03	.41

* $p < .05$

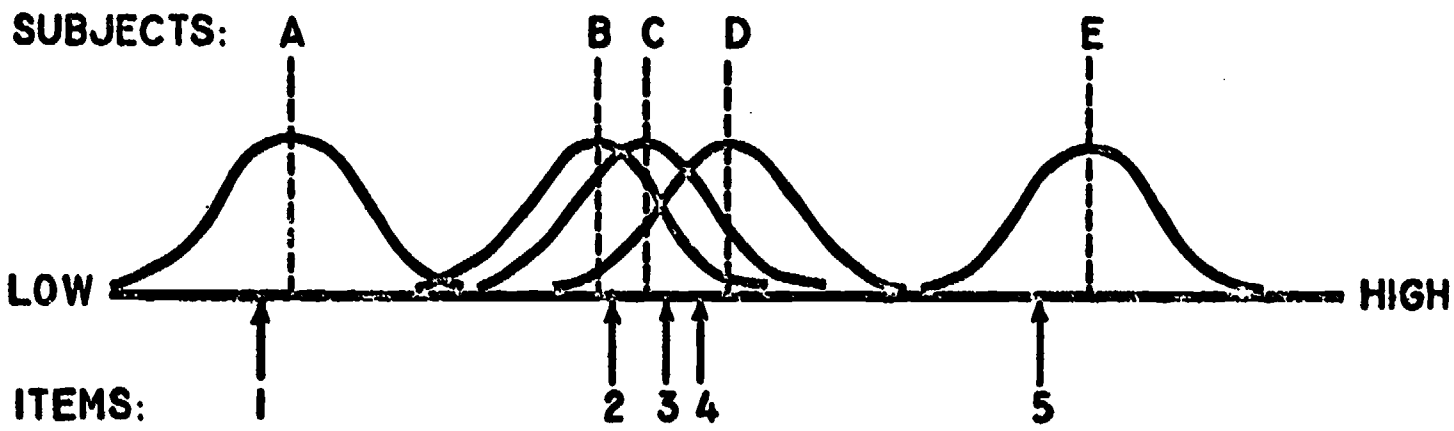

Figure 1: Representation of person and item distributions on a psychological continuum for a typical personality test

PERSONS			CONTINUUM	ITEMS		
Scale Value	Freq.	Raw Score	N = 137 n = 18	Item No.	Prop. End.	Scale Value
*	3	18		13	.25	1.07
3.10	3	17	+3	1	.26	1.02
2.32	1	16		7	.27	.98
1.82	3	15	+2	8	.28	.88
1.43	6	14		2	.29	.84
1.10	7	13	+1	14	.30	.79
.81	3	12		16	.40	.24
.53	11	11		17	.41	.20
.27	16	10	0	3	.43	.08
.01	16	9		9	.46	- .07
- .25	7	8		4	.47	- .11
- .51	6	7	-1	10	.48	- .15
- .79	10	6		5	.48	- .15
-1.09	8	5	-2	18	.62	- .90
-1.43	10	4		11	.64	- .98
-1.83	10	3	-3	15	.66	-1.07
-2.34	8	2		6	.67	-1.17
-3.13	3	1		12	.73	-1.51
*	6	0				

* Cannot Be Computed

Figure 2: Rasch log scale estimates and continuum relationships of persons and items for scale M

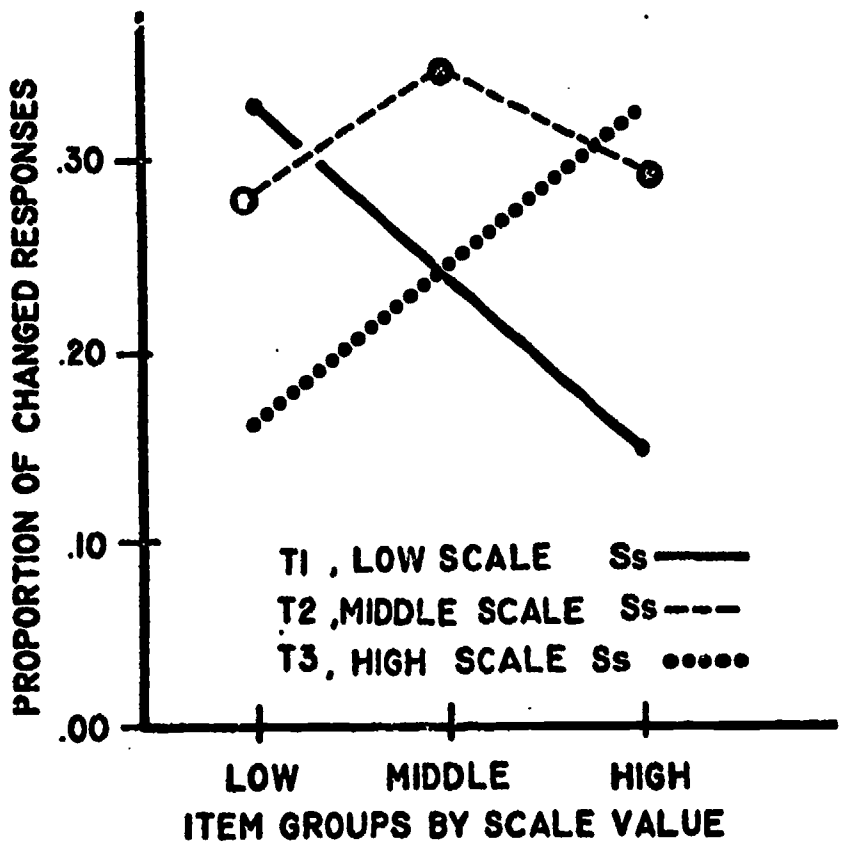

Figure 3: Response instability and subject-item interaction on scale M