DOCUMENT RESUME

ED 030 907 CG 004 000

By-Hendel, Darwin D.; Weiss, David J.

Individual Inconsistency and Reliability of Measurement.

Minnesota Univ., Minneapolis.
Spons Agency-Department of Health, Education, and Welfare, Washington, D.C. Social and Rehabilitation Service.
Report No-WAP-RR-19
Pub Date Oct 68
Note-24p.
EDRS Price MF-$0.25 HC-$1.30
Descriptors-Individual Differences, *Measurement, *Performance Factors, *Reliability, *Statistical Analysis, Statistical Data, Statistical Studies, *Test Reliability

Total circular triad scores (TCT) derived from the pair-comparison Minnesota Importance Questionnaire (MIQ) were used to study the relationship between inconsistency, and both internal consistency reliability and stability. Stability estimates (and Hoyt Coefficients) were computed for each of nine groups (retest internals from immediate retest to 10 months) for the 20 MIQ scales. Stability estimates were also computed for each individual. Results showed that scale stability and individual stability coefficients, as well as internal consistency reliabilities, were higher for low TCT groups. Correlations between individual stability and TCT were from -.24 to -.68. These results indicate that reliability estimates are related to individual differences in response consistency. (Author)

Individual Inconsistency and Reliability of Measurement

Darwin D. Hendel

and

David J. Weiss

Research Report No. 19
October, 1968

Work Adjustment Project
University of Minnesota

Abstract

Individual Inconsistency and Reliability of Measurement

Darwin D. Hendel and David J. Weiss

University of Minnesota

Total circular triad scores (TCT) derived from the pair-comparison Minnesota Importance Questionnaire (MIQ) were used to study the relationship between inconsistency, and both internal consistency reliability and stability. Stability estimates (and Hoyt coefficients) were computed for each of 9 groups (retest intervals from immediate retest to 10 months) for the 20 MIQ scales; stability estimates were also computed for each individual. Results showed that scale stability and individual stability coefficients, as well as internal consistency reliabilities, were higher for low TCT groups. Correlations between individual stability and TCT were from -.24 to -.68. These results indicate that reliability estimates are related to individual differences in response consistency.

Individual Inconsistency and Reliability of Measurement[1]

Darwin D. Hendel and David J. Weiss

University of Minnesota

The concept of reliability of measurement is clearly not as simple and static as standard definitions often imply. Reliability is not an all or none criterion which, if once satisfied, is invariant for a given measuring instrument, for different groups, or for different testing conditions. Reliability may also be examined in relation to a given measure for a given individual, thus implying the relevance of examining specific individual factors contributing to unreliability.

Unreliability, Thorndike's "error variance" (1951), can be seen as being composed of two classes of elements: (1) characteristics of the observer and the environment; and (2) characteristics of the individual. The first group is composed of such factors as poor testing conditions, careless investigators, inaccurate calculations and numerous other factors which are external to the individual being examined. Included in individual characteristics are aspects such as test-taking ability, response sets, response styles and guessing habits.

Reliability of measurement implies more than consistency of response over a time interval. Rather, reliability can be discussed in two different frameworks--test-retest reliability (stability) and internal consistency reliability. Test-retest reliability refers to the stability of measurement across some time interval. Stability depends greatly on the trait being measured, the time interval between administrations, and the

---

[1]This study was supported in part by Research Grant RD-1613-G from the Social and Rehabilitation Service, Department of Health, Education and Welfare, Washington, D. C. The first author was a National Defense Education Act Fellow in Counseling Psychology Research, at the University of Minnesota, during the conduct of this research.

individuals being measured. Internal consistency reliability can be conceived of as replication over items derived from the same domain of response (Ghiselli, 1964). Internal consistency is based on repeatability at one point in time; it implies high intercorrelations among items, high predictability from one response to another. Reliability reflects variation which is systematic; however, it must concurrently be noted that some individual difference variables are also systematic.

Ghiselli (1964), in his discussion of "systematic and unsystematic" variation in test scores, attributes the basis of reliability estimation to individual factors in test scores. Such an approach supports Gulliksen's (1950) reliability model in which only random and unsystematic factors are included in error variance. The traditional model of psychometric reliability, while based on individual differences, estimates individual reliability from group data. In this approach, the "error band" on an individual's score is derived from the "standard error of measurement" based on group data. Such an approach ignores the possibility of the measurement of <u>individual differences</u> in reliability or the identification of <u>individual</u> factors which reflect differential reliability of measurement.

The hypothesis that individuals can be differentiated with respect to factors reflecting reliability of measurement has been suggested by Neff and Cohen (1967). Their data show individual differences in response consistency of single subjects. According to Gulliksen (1964, p. 70), individual differences in response consistency as measured by the circular triads score can reflect the "varying stability of a preference system, or the varying carefulness among subjects...." Both the "stability of a preference system" and "differences in carefulness," as reflected in scores on an instrument, are factors relating to traditional concepts of reliability.

Inconsistency, in addition to its possible relationship to reliability, is important in its own right. Response inconsistency may be a behavioral trait quite independent from the response problems it defines. Pemberton (1966) examined correlates of inconsistency and found biographical descriptions of individuals related to inconsistency scores. Davis (1958) presents evidence for the existence of inconsistency as a stable trait. Based on the assumption that man is rational enough to be capable of a weak ordering of preferences, he concludes that inconsistency cannot be fully explained as a random choice among indifferent objects.

Some evidence concerning the relationship of reliability and inconsistency has been reported. Weksel and Ware (1967), in a study relating test-retest reliability and circular triad scores, found a correlation of -.36, indicating a significant relationship between consistency and stability (high total circular triad scores indicate a tendency toward random response). Jackson (1966) showed a consistent drop in test-retest reliability coefficients as a function of level on an "Infrequency Scale," an indicator of "non-purposeful responding" on his Personality Research Form. Both studies support the hypothesis that individuals can be differentiated in regard to consistency of judgment, and that consistency is related to stability of measurement for these individuals.

The present study is concerned with investigating the generality of these findings and, based on Gulliksen's hypothesis, determining to what extent the total circular triad score (TCT) in pair comparison scaling can differentiate individuals with respect to reliability of measurement. In order to investigate the generality of previous findings, this study used several different groups to determine if results were replicable from group to group or if the findings were group specific. Since the level of inconsistency for differently constituted groups may be different, the relationship between reliability and consistency need not be invariant. To more completely confirm previous findings, this study also examined groups having different time intervals between test and retest sessions to determine the relationship between inconsistency and stability as a function of test-retest time interval. To further study the generality of relationships between inconsistency and reliability, the study considered the following types of reliability measures: 1) scale internal consistency reliability; 2) test-retest scale stability; and 3) individual test-retest profile stability.

Four hypotheses were investigated in the present study. First, if TCT functioned as a moderator variable, it was hypothesized that scale-by-scale stability coefficients for a group lower in TCT would be higher than for a group with higher TCT scores. Second, if consistency of response is related to internal consistency reliability, it was hypothesized that scale internal consistency reliabilities would be higher for groups with lower TCT scores. Third, it was hypothesized that there would be an inverse relationship between TCT scores and test-retest stability for individuals. Fourth, it was hypothesized that the relationship between inconsistency and reliability would be influenced by the nature of the group and the test-retest time intervals.

Method

Instrument. The instrument used in the study was a 190-item form of the Minnesota Importance Questionnaire (MIQ; Weiss, Dawis, England and Lofquist, 1967). This form uses a complete pair-comparison of twenty statements measuring vocational needs. Scale scores used in the analyses were derived by counting, for each of the stimulus variables, the number of times it was chosen over the other nineteen stimuli. The maximum score on any one scale was nineteen, the minimum score, zero. For each individual, the sum of the twenty scale scores was 190 (assuming of course, completed questionnaires for every individual). Inconsistency, as measured by total circular triads (TCT), was computed by Kendall's (1955, p. 125) formula. Low TCT scores reflect logically consistent judgments; high TCT scores indicate intransitive (logically inconsistent) judgments which may be due to a number of individual factors, such as response set, random response, inability to discriminate the stimuli, or carelessness (Gulliksen, 1964).

Subjects. The study involved nine different groups with different test-retest time intervals for each of the groups. The group size and test-retest intervals for each of the groups are contained in Table 1. Group 2, for example, was composed of 146 subjects, 65 males and 81 females, with a test-retest time interval of 1 week. Test-retest intervals ranged from an immediate test-retest group to a group having a ten month test-retest time interval.

[Insert Table 1 about here]

Groups 1, 2, 3 and 4 were composed of University of Minnesota students in introductory psychology courses; all classes were represented in these groups, although the groups were predominantly sophomores. Group 7 was composed of students in a night school course in vocational psychology; there was a wide age range and variety of occupational backgrounds in this group. Group 9 was composed of a group of junior and senior college students enrolled in the social work curriculum at the University of Minnesota. Group 6 was composed of 180 high school seniors in four suburban Minneapolis high schools. The subjects in group 5 were high school seniors enrolled in one suburban Minneapolis high school. Students in group 5 were matched with subjects in group 6 on variables such as sex, father's occupation, and grade point average; subjects in group 6 were enrolled in vocational education programs, whereas subjects in group 5 were not. Group 8 was composed of individuals in the Minneapolis New Careers Program, a work-study program for low income adults funded by the Department of Labor. The groups were selected to provide data reflecting various degrees of stability of preference systems with groups 5 and 6 (high school students) and group 8 assumed least stable, and groups 7 and 9 likely to be most stable.

Analysis. In order to investigate the relationship between TCT and reliability, the groups were divided into subgroups on the basis of number of circular triads (Kendall, 1955, p. 125) on the first administration of the MIQ. Subgroup sizes and range of TCT values can be found in Table 2. Because of the initial small number of subjects in groups 1, 5, 7, 8 and 9, these groups were divided into two subgroups, low TCT and high TCT. In group 1, for instance, there were 21 subjects in each subgroup; the ranges of TCT were 15-50 and 55-133 for the low and high TCT groups respectively. The four larger groups (2, 3, 4 and 6), were divided into approximately equal thirds for the low, middle and high TCT subgroups.

[Insert Table 2 about here]

In examining reliability on a group basis, both test-retest and internal consistency reliabilities were computed for each of the 20 MIQ scales. Test-retest scale stabilities were computed for each of the total groups and their respective subgroups by correlating scores on each of the 20 MIQ scales at the first administration of the questionnaire with those obtained at the retest session. Ranges and median scale stability coefficients (across the 20 MIQ scales) were computed for each of these TCT subgroups. Internal consistency reliability coefficients for TCT subgroups for each of the 20 scales on the first administration of the MIQ were computed by Hoyt's (1941) formula. Ranges and median scale internal consistency reliabilities were computed for each TCT subgroup.

In order to test the significance of differences in test-retest scale correlations between low and high TCT subgroups, test-retest correlations were transformed to z's and tested for differences between groups on each of the 20 scales (Hays, 1966, p. 531).

In examining stability on an individual basis, product-moment stability coefficients (0 correlation) were computed for each individual across the 20 MIQ scales (Cattell, 1952, p. 503). Product-moment correlations were appropriate for these data since the MIQ is completely ipsative; hence no level differences were possible between first and second administrations. In order to test the relationship of TCT and individual stability, a median test was used on the distribution of individual test-retest correlations between TCT subgroups. Median individual stabilities were found for each of the nine groups; individuals in each of the TCT subgroups were then classified as having low or high stability coefficients based on the total group median. Chi-square values were computed for six of the groups (2, 3, 4, 5, 6 and 8); because of the small numbers of subjects in groups 1, 7 and 9, Fisher's exact probability test was used as a test of the hypothesis. In order to obtain a more concise estimate of the predictive relationship between inconsistency and individual stability, product-moment correlations were computed between individual test-retest reliabilities and the number of circular triads on the first administration of the MIQ. This procedure was used to provide further explication of the results which were obtained in the median test analysis.

Results

Scale analysis. The range and median of scale test-retest correlations for the TCT subgroups and total groups are shown in Table 3. In group 2, for example, scale stability correlations ranged from .62 to .91 for the total group, and .70-.98, .61-.90, .45-.91 for the low, middle, and high TCT subgroups respectively. For this same group, the median correlation was .81 for the total group and .87, .82 .75 for the low TCT, middle TCT, and high TCT subgroups respectively. In quite similar fashion, the ranges and medians are listed for the groups in which the breakdown was into two subgroups only--low TCT and high TCT. For eight of the nine groups, median reliability coefficients were highest for the low TCT subgroup, with ranges of coefficients also exhibiting a similar pattern. These data show that traditional scale-by-scale test-retest reliability coefficients were generally higher for the low TCT group than for the high TCT groups, thus supporting the first hypothesis.

[Insert Table 3 about here]

In examining the significance of the differences in scale-by-scale test-retest reliability between low TCT and high TCT subgroups, statistically significant differences were obtained for many of the scales. Results of the significance tests for the 20 MIQ scales for each of the nine groups are given in Table 4. Group 7, for instance, yielded no significant differences (in either direction) for any of the 20 MIQ scales; for group 4, significant differences in the expected direction were obtained for 14 of the scales. In three of the smaller groups (1, 8 and 9), a few of the differences were not in the predicted direction. Considering the total results, however, the data tend to show that low TCT subgroups had many significantly higher scale-by-scale test-retest correlations than did high TCT subgroups.

[Insert Table 4 about here]

Results of the internal consistency analysis, as shown in Table 5, yielded results similar to those obtained in the scale stability analysis. In group 3, for example, the median coefficient for the total group was .80;

medians for the low TCT, middle TCT, and high TCT groups were .85, .83 and .75 respectively. For eight of the nine groups, the low TCT subgroup had the highest median Hoyt coefficient. For all groups, the highest single scale reliability coefficient was for the low TCT subgroup. The data in Table 5, therefore, support the second hypothesis, that groups low in TCT would have higher scale-by-scale internal consistency reliabilities than groups higher in TCT.

[Insert Table 5 about here]

Individual analysis. Results obtained from an analysis of the relationship between individual stability and inconsistency (as measured by TCT) support the previous analyses. These data, contained in Table 6, also provide further support for the hypothesis that individual differences variables are related to stability of measurement. In group 4, as an example, the median individual stability coefficients were .87 for the total group and .91, .86, and .81 for the low TCT, middle TCT and high TCT subgroups respectively. The p-value of .001 obtained from the median test calculation for this group supports a rejection of the null hypothesis of no significant differences in the distribution of subgroup stability correlations. The p-values for all the larger groups (2, 3, 4, 5, 6, 8), were significant far beyond the .001 level. Results obtained by using Fisher's exact probability test in groups 1, 7 and 9 were significant only for group 9. Yet for all groups the high TCT subgroup had the lowest median stability correlation, and for eight of the nine groups, the range of stability correlations was smallest for the low TCT subgroup.

[Insert Table 6 about here]

Product-moment correlations between TCT at time 1 and individual stability coefficients are shown in Table 7. These correlations were all negative, ranging from -.24 for group 9 to -.68 for group 7. The product-moment correlations were significant at the .01 level for seven of the nine groups. These product-moment correlations further confirm the third hypothesis that there is an inverse predictive relationship between TCT scores and test-retest stability for individuals. Considering stability on an individual basis thus provides similar results as when reliability is considered on the basis of group data. That is, inconsistency tends to be negatively correlated with reliability; individuals low in TCT are likely to have higher test-retest profile stability correlations than are individuals scoring high on the TCT variable.

[Insert Table 7 about here]

The fourth hypothesis in this study was concerned with interactions of type of group, test-retest time interval and the relationship between inconsistency and reliability. Inconsistency appears to be related to internal consistency reliability in the same fashion for all the groups in this study, regardless of type of individual (see Table 5). In all cases the low TCT subgroup had higher reliabilities than did the high TCT subgroup. The tendency was least marked for group 8 (New Careers) which was also the group with the highest proportion of females. The scale stability analyses showed no apparent trend for retest time interval to be related to the relationship between reliability and consistency; total group reliabilities as well as TCT subgroup reliabilities tended to decrease uniformly with increasing retest interval (see Table 3). For group 7 (night school students), however, the predicted relationships did not occur between consistency and scale stability. These results may have been due to any or a combination of three factors unique to group 7: 1) it was the smallest group; 2) it had the largest proportion of males; and 3) it was the only regularly employed group. Since both the stability and consistency of vocational needs as measured by the MIQ would be expected to be confounded by employment status, the negative findings for group 7 do not necessarily disconfirm the hypothesis. When the stability data were examined on an <u>individual</u> basis, group 7 showed the highest ($r = -.68$)

correlation between TCT and stability. The correlations between TCT and stability (Table 7) suggest that either 1) the predictive relationship does not hold up for relatively long time intervals (9 or 10 months); or 2) that sex moderates this relationship (since groups 8 and 9 had both the highest proportion of females and longest time intervals). These hypotheses must be qualified, however, because of the small groups used for the 9 and 10 month analyses.

Conclusions

The differentiation of individuals with respect to factors reflecting reliability of measurement has been previously noted by Neff and Cohen (1967). The results of the present study support this hypothesis, demonstrating that response consistency, as measured by TCT scores, is related to reliability, regardless of the type of reliability being considered. In terms of Thorndike's (1951) formulation, consistency of response, as measured by circular triads, can be appropriately seen as a factor characteristic of individuals. Results of the correlation analysis between time 1 TCT and individual profile stability replicate the results obtained by Jackson (1966) and Weksel and Ware (1967), thus confirming the importance of examination of specific individual factors contributing to reliability. These data also support Gulliksen's (1964) hypothesis that TCT scores reflect the stability of an individual's preference system, and can therefore be considered as an index of individual reliability. The use of nine different groups in the present study suggests that the relationship is quite general in that similar results were obtained for different groups and for a variety of test-retest time intervals, although sex of subjects and/or time interval appear to interact with the relationship between consistency and reliability.

The inverse relationship between inconsistency and individual test-retest profile stability points out the relevance of consideration of individual factors in such a manner that reliability of measurement can be increased. Consideration of specific individual difference variables contributing to reliability, instead of estimating reliability completely from group data, allows a more complete and understandable examination of reliability. The use of inconsistency is but one of numerous factors which may be studied in an effort to determine the precise meaning of unreliability.

Furthermore, the present study shows that individual response consistency can act as a moderator variable within the traditional reliability model. The fact that the significant test-retest stability estimates between TCT subgroups did not appear for the same scales on all groups, indicates that TCT scores identify an important source of unreliability related to individual differences variables. By further examination, it may be found that reliability and inconsistency are related for specific domains of questionnaire stimuli. This suggests that different variables in pair comparison scaling are differentially related to number of circular triads. If random response were the only factor causing high TCT scores, it would be expected that all stimuli would be equally affected. It can be further hypothesized that circular triad scores may represent a composite of sub-scores related to differential scalability of stimuli in a given set, as well as a component reflecting random response. Inability to make fine discriminations between stimuli, lack of understanding, and carelessness, are three possible subfactors.

In general, these results show that: 1) there are individual differences in response consistency in pair comparisons scaling; 2) response consistency moderates traditional reliability estimates, with the more consistent groups having the highest reliability (both internal consistency and stability) and the least consistent groups the lowest reliability; and 3) individuals with consistent responses have more stable preferences systems than those of low consistency. Thus, it would appear that traditional models of reliability, in which reliability estimates for an individual are estimated from group data, could yield more accurate estimates if individual differences variables, such as response consistency, were taken into consideration in the estimation of reliability.

Table 1

Group size and test-retest time interval

Group	Number of Individuals Total	 Male	 Female	Time Interval
1	42	19	23	Immediate test retest
2	146	65	81	1 week
3	157	70	87	2 weeks
4	283	115	168	6 weeks
5	73	31	42	4 months
6	180	69	111	6 months
7	27	19	8	7 months
8	53	8	45	9 months
9	38	7	31	10 months

Table 2

Number of individuals and range of total circular triad (TCT) scores for subgroups based on TT scores

Group	Low TCT		Middle TCT		High TCT	
	N	Range	N	Range	N	Range
1	21	15-50	..	...	21	55-133
2	49	3-32	49	33-57	48	58-252
3	50	11-33	53	34-65	54	66-234
4	94	4-37	94	38-63	95	64-199
5	36	8-59	..	...	37	61-250
6	61	3-46	59	47-87	60	88-286
7	13	12-32	..	...	14	35-211
8	26	18-68	..	...	27	76-262
9	19	4-33	..	...	19	36-141

Table 3

Range and median of scale test-retest stability correlations, for total group and TCT subgroups

	Range of Scale Stability Correlations								Median Scale Stability Correlation			
	Total Group		Low TCT		Middle TCT		High TCT					
Group	Lo	Hi	Lo	Hi	Lo	Hi	Lo	Hi	Total Group	Low TCT	Middle TCT	High TCT
1	.72	.93	.64	.95	..	..	.69	.94	.89	.91	..	.84
2	.62	.91	.70	.98	.61	.90	.45	.91	.81	.87	.82	.75
3	.66	.89	.73	.94	.49	.94	.54	.84	.79	.86	.78	.71
4	.65	.83	.62	.92	.64	.82	.52	.80	.75	.82	.75	.66
5	.35	.77	.26	.89	..	..	.25	.72	.63	.72	..	.50
6	.40	.68	.37	.72	.17	.75	.17	.64	.48	.55	.49	.35
7	.40	.83	.29	.91	..	..	.25	.92	.68	.67	..	.69
8	.19	.77	.03	.70	..	..	.18	.81	.49	.52	..	.44
9	.46	.79	.35	.80	..	..	.15	.80	.53	.56	..	.50

Table 4

Significance of differences in test-retest scale stability correlations between low and high TCT groups

	Group and Number of Subjects								
	1	2	3	4	5	6	7	8	9
Scale	N=42	N=146	N=157	N=283	N=73	N=180	N=27	N=53	N=38
1		**							*
2	-	*			**				
3			**	**				-	
4				*	*	**		*	
5			**	**		*			
6	*	**		**	*				
7		*	*	**		**	No significant differences for any scale		
8	-	*	*	*	*				-
9		**		**					**
10	**	**			*			*	
11		**	*	**					
12	*					*			
13		**	*	**	*				*
14	*	*	**	**	**	**			**
15	-	**	**	**	**	**			*
16		**		**	**				
17		**	**	**		*			
18			*						
19			*			**		*	-
20				*		*		*	

*Significant at .05 level.

**Significant at .01 level.

-Significant at .05 level (Not in predicted direction)

Table 5

Range and median of Hoyt internal consistency reliability coefficients, for total group and TCT subgroups

	Range of Coefficients								Median Coefficient			
	Total Group		Low TCT		Middle TCT		High TCT					
Group	Lo	Hi	Lo	Hi	Lo	Hi	Lo	Hi	Total Group	Low TCT	Middle TCT	High TCT
1	.42	.92	.41	.94	..	..	.65	.92	.81	.82	..	.78
2	.58	.94	.62	.97	.56	.93	.38	.95	.81	.83	.82	.75
3	.55	.95	.68	.96	.49	.95	.52	.89	.80	.85	.83	.75
4	.63	.95	.58	.96	.61	.95	.59	.94	.81	.86	.82	.77
5	.62	.93	.48	.95	..	..	.53	.91	.79	.84	..	.75
6	.59	.92	.67	.96	.61	.92	.34	.80	.80	.84	.82	.71
7	.48	.95	.33	.97	..	..	.46	.93	.81	.79	..	.73
8	.40	.89	.49	.93	..	..	.26	.86	.77	.77	..	.73
9	.30	.90	.15	.93	..	..	.11	.89	.78	.82	..	.72

Table 6

Range and median of stability coefficients for individual

test-retest profile correlations, by TCT sub-group

	Range of Individual Stability Coefficients								Median Stability Coefficients				
	Total Group		Low TCT		Middle TCT		High TCT						
Group	Lo	Hi	Lo	Hi	Lo	Hi	Lo	Hi	Total Group	Low TCT	Middle TCT	High TCT	Level of Significance[a]
1	.71	.97	.85	.97	..	..	.71	.97	.95	.94	..	.92	Not significant
2	.43	.98	.82	.98	.51	.96	.43	.96	.90	.93	.91	.86	.001
3	.11	.98	.81	.98	.59	.96	.11	.94	.89	.93	.90	.84	.001
4	.19	.97	.52	.97	.41	.96	.19	.96	.87	.91	.86	.81	.001
5	-.04	.97	.59	.97	..	..	-.04	.89	.78	.85	..	.68	.001
6	-.44	.95	-.30	.95	-.44	.89	-.44	.82	.71	.78	.72	.52	.001
7	.48	.93	.67	.91	..	..	.48	.93	.83	.83	..	.81	Not significant
8	.29	.94	.29	.94	..	..	.52	.93	.76	.78	..	.71	.001
9	.58	.97	.66	.97	..	..	.58	.93	.87	.88	..	.84	.05

[a]The median test was used for groups 2-6 and 8; Fisher's exact probability test was used for the other three groups.

Table 7

Product-moment correlations between time 1

TCT and individual stability coefficients

Group	N	Time Interval	Pearson Correlation	Level of Significance
1	42	Immediate test retest	-.57	$p < .01$
2	146	1 week	-.47	$p < .01$
3	157	2 weeks	-.56	$p < .01$
4	283	6 weeks	-.61	$p < .01$
5	73	4 months	-.50	$p < .01$
6	180	6 months	-.45	$p < .01$
7	27	7 months	-.68	$p < .01$
8	53	9 months	-.25	Not significant
9	38	10 months	-.24	Not significant

References

Cattell, R. B. The three basic factor-analytic research designs--their interrelations and derivatives. Psychological bulletin, 1952, 49, 499-520.

Davis, J. M. The transitivity of preferences. Behavioral science, 1958, 3, 26-33.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw-Hill, 1964.

Gulliksen, H. Intercultural studies of attitudes. In Fredericksen, N and Gulliksen, H. (Eds.). Contributions to mathematical psychology. New York: Holt, Rinehart and Winston, 1964.

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

Hays. W. L. Statistics for psychologists. New York: Holt, Rinehart and Winston, 1963.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Jackson, D. N. A modern strategy for personality assessment: The Personality Research Form. Department of Psychology, University of Western Ontario, Research Bulletin No. 30, October, 1966.

Kendall, M. G. Rank correlation methods. New York: Hafner, 1955.

Neff, W. S. and Cohen, J. A method for the analysis of the structure and internal consistency of Q-sort arrays. Psychological bulletin, 1967, 68, 361-368.

Pemberton, W. A. Consistency of preference as an aspect of personality. Paper read at the 1966 Eastern Psychological Association Convention, New York City.

Thorndike, R. L. Reliability. In Lindquist F. (ed). Educational measurement. American Council on Education, 1951.

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

Weiss, D. J., Dawis, R. V., England, G. W., and Lofquist, L. H. Instrumentation for the Theory of Work Adjustment. Minnesota studies in vocational rehabilitation. 1966, 21.