韓國心理學會誌
Korean Journal of Psychology
1981, Vol. 3, No. 2, 85-99

The Reliability of Item Sensitivity and Other Item Indices for Thurstone-Type Attitude Scales

Cha Jae-Ho and Lee Seok-Jae

Department of Psychology
Seoul National University

A son preference attitude scale of Thurstone-type was administered to two groups of college freshmen, one group receiving Form A and the other group receiving Form B of the scale. The two forms differed only in the order of 31 attitude statements. Each group was retested on the same form about two weeks later. For Form A 55 subjects completed both sessions and for Form B 54 did so. The three independent components making up the total sum of squares of an endorsement set (3*n* endorsements given to a particular item by the *n* subjects who endorsed that particular item in one of his three endorsements permitted) were calculated and the resulting three mean squares were correlated, variously, to item scale value, item popularity, item ambiguity, and between testings and between forms. The between-subjects mean square which is defined as item sensitivity showed a statistically significant test-retest reliability in both samples. The interform reliability was also statistically significant for between-subjects mean square but not for within-subject mean square. These results were interpreted as indicating that between-subjects mean square (item sensitivty) measures item characteristics unique to the item uninfluenced by other items in the scale while within-subject mean square measures item characteristics of an item that are influenced by neighboring items (statements) in the scale. It was also found that on the average the between-subjects sum of squares occupies approximately 25% of the total sum of square of an endorsement. Since the remaining portion of the total sum of squares do not measure item characteristics unique to an item and/or measure unique characteristics not useful for selecting items in the second stage of scale construction, it was maintained that item sensitivity is a useful item index which does away with the remaining 75% which contains noise as far as item selection is concerned and that item sensitivity is a more useful and precise substitute for Thurstone's test of irrelevance which basically relies upon the total variance of an endorsement set.

The purpose of the present study was to obtain empirical evidence concerning the reliability of some item indices that may prove to be useful in constructing a Thurstone-type attitude scale. The particular types of reliability studied were test-retest reliability and interform reliability in which scores from two parallel forms differing only in the ordering of attitude statements are intercorrelated. The item (statement) indices studied in the present study are item sensitivity and other two indices derived from the analysis of the endorsemement set. Of these indices, item sensitivity is our primary concern in that this index was proposed as a new item index (Cha, 1973) which could be used as a basis for selecting attitude statements for inclusion into the final form of a Thurstone-type scale. This index was proposed as an alternative to Thurstone's test of irrelevance (Thurstone, 1928; Ferguson, 1952; Edwards, 1957, pp. 98-99) with greater objectivity and free of some of the confoundings the test of irrelevance is subject to.

The sensitivity index is an average of sum of squares of one component of *an endorsement set*. After the preliminary form of an attitude scale is constructed using such information as item (statement) scale values and item ambiguity (Edwards, 1957, pp. 86-92), this preliminary form is administered to a new sample of subjects for whom the scale is intended. It is assumed throughout that each subject is allowed to endorse three and only three statements which he judges to be most representative of his own attitudinal position. With N subjects in the sample, there would be a total of 3N endorsements or a set of 3N scale values associated with the endorsed statements, regardless of the number of statements included in the scale. *An endorsement set is a subset of these 3N scale values and is defined as that set of scale values associated with the endorsements made by the subset of subjects who have in common endorsed a particular attitude statement* (such an attitude statement defining an endorsement set will be referred to as a criterial statement or item). Since each attitude statement in the scale may in turn become the criterial statement, there are attitude statements in the scale. But since the range of attitude dimension covered by an attitude scale is usually much wider than the attitude range shown by any one group of subjects, some statements, particulary those located at either end of the scale, may have no one or very few subjects endorsing them.

The size of an endorsement set in each case is $3n$, where n is the number of subjects who have endorsed a criterial statement. The number of subjects endorsing a criterial statement (n) can be used as an index of item popularity and will be so referred to henceforth in this paper. Since an increases as its scale position approaches the center of the group's attitude distribution, so will the size of an endorsement set. Endorsement sets are not mutually independent but overlap so that the sum of the sizes (3n's) of endorsement sets exceed 3N. It is because each subject contributes three endorsements and therefore he will be represented in three different endorsement sets.

For each endorsement set, one can calculate the total sum of squares based on the deviation of each of 3n scale values from the scale value of criterial statement, K. The total sum of squares of an endorsement set can be shown to be composed of three independent components (Cha, 1973):

$$\sum^{n}\sum^{3}(X-K)^2=\sum^{n}\sum^{3}(X-\bar{X})^2+3\sum^{n}(\bar{X}-M)^2 +3nd^2.$$

In the above expression, X stands for scale value associated with an endorsement, $\bar{X}$ for the mean of a subject's three endorsements, M for the mean of the endorsement set itself, and K for the scale value of the criterial statement. The quantity d is defined as $M-K$. The three components on the right hand side of the expression correspond, respectively, to (1) the within-subject sum of squares, (2) the between-subjects sum of squares, and (3) the "skewness" or deviation sum of squares, and (3) the "skewness" or deviation sum of squares.

These three compoments of an endorsement set may be examined as to what each signifies. First, the within-subject component represents within-subject variations, and since this variations are largely dependent upon the distances of the scale positions of statements neighboring the criterial statement, any item index based on this component is not a good index of the item characteristics of the criterial statement. An exception would be a situation where attitude statements are equally spaced along the attitude dimension scale-position-wise. In this special case, and differences in the within-subject sum of sqares among endorsement sets will reflect on a characteristic of the criterial statement. One item characteristic which might be indexed by the within-subject sum of squares in this special situation is item ambiguity (Edwards, 1957),. In a more usual situation in which the attitude statements are unevenly distributed along the attitude dimension, the mean square based on the within-subject variations was found unrelated to item ambiguity (Cha & Lee, 1974).

The second component is the between-subjects component and forms the basis of item sensitivity index, which is its mean square. A large between-subiects variation means that the criterial statements tends to attract an attitudinally heterogeneous group of subjects while a smaller between-subjects variation means that the criterial statement receives endorsements from subjects who are highly similar in attitude position. The third component, the "skewness" or deviation component, is expected to be laregly a function of the scale position of the subject sample, an endorsement set of a criterial statement located near the center of the attitude distribution is expected to have about an equal number of endorsements falling on either side of the scale position corresponding to the criterial statement. In a case such as this, the mean of the endorsement set will closely approximate the scale position of the criterial statement. But, for criterial statements located at the right extreme of the attitude distribution will tend to have a negatively "skewed" endorsement set in the sense that more endorsements will be found to the left than to the right of the criterial statement. Similarly, the endorsement sets for the criterial statements located at the left extreme of the attitude distribution will show a positive "skewness." If these expectations are correct, the difference between the set mean and the scale position of the criterial statement ($d=M-K$) will be found to be a negative for criterial statements located at the positive (right) end of the attitude distribution and a positive for criterial statements located toward the negative (left) end of the attitude distribution. In other words, the third component is related to the scale position of the criterial statement. That this is indeed the case was shown by a high correlation coefficient ($r=.95$) obtained between the d score and the scale position of the criterial statement (Cha & Lee, 1974). Since the calculation of this component presupposes the knowledge of the scale position of the criterial statement, and since the scale position of the criterial statement must be known before an endorsement set is established, this component which apparently to indexes only the scale position of the criterial statement does not provide any useful item index.

In sum, it may be said that the the within-subject component is a confounded item index, reflecting an item characteristic or characteristics of the criterial statement as the item characteristics of other statements in the scale and scale characteristics (as opposed to item characteristics), that the between subjects component is a relatively pure index of an item characteristic, reflecting the attitude homogeneity of persons endorsing the criterial statement, and that the "deviation" component is another relatively pure but not very useful item index reflecting mainly the scale position of the criterial statement.

It was proposed previously (Cha, 1974) that the between-subjects sum of squares divided by the number of subjects (n or n-1) may provide a useful index of an as yet undetermined item characteristic of an attitude statement in Thurstone-type attitude scales, that this new item index be called ***item sensitivity***, and that the sensitivity index is a purer and more precise substitute of Thurstone's test of irrelevance (Thurstone, 1928; Thurstone & Chave, 1928).

This last point needs further clarification. Since the shape of the distribution of similarity indices which comprise the basic data in Thurstone' test of irrelevance is directly related to the total variance of the endorsement set as defined in the present paper (See Ferguson, 1952), the test of irrelevance is a confounded test insomuch as the total sum of squares of an endorsement set is shown to be composed of three independent components and among these only the between-subjects sum of squares (which forms the basis of the sensitivity index) appears to offer any pure item index useful for selecting from among statements. In other words, the test of irrelevance is not a test of the unidimensionality of an attitude statement with respect to other attitude statements in a Thurstone type attitude Scale as Thurstone supposed it to be but rather a confounded index of an statement's between-subject sensitivity.

The shape of the distribution of similarity indices so essential in making the test of irrelevance is expected to be closely related to the size of the total variance of an endorsement set because a similarity index is the number of endorsements given to an attitude statement converted into a proportion (the proportion of subjects endorsing any one attitude statement given that they all endorsed at the same time a particular criterial statement). If the shape is an inverted U shape with the magnitude of similarity indices diminishing as their positions move away from the position of criterial statement on either side, it indicates a small total variance and, if the shape is more or less flat over the whole attitude range of the subjects, it indicates a large total variance. In the study cited earlier (Cha, and Lee, 1974), the results of the test of irrelevance (which consisted of the ranking of unidimensionality of attitude statement, an inverted U shaped distribution of similarity indices receiving a high undimensionality rating and a flat shaped distribution receiving a low unidimensionality rating) were positively related to the size of the total variance of an endorsement set with $r=.66$ ($n=13$). There are reasons to believe that the obtained coefficient far underestimate the extent of actual correlation. Thurstone (1928) assumed that similarity indices distributed in an inverted U-shape indicate that the criterial statement is measuring what the remaining statements are measuring (a demonstration of unidimensionality) and that a horizontal distribution of similarity indices means that the criterial statement is measuring something which is irrelevant to what the remaining statements as a whole are measuring (a demonstration of a lack of unidimensionality). The present analysis of endorsement sets implies that ***different shapes of the distribution of similarity indices are amenable to yet another interpretation, namely the total variance of an endorsement set*** (which Thurstone supposed to be indicative of to the unidimensionality of the criterial statement) ***is in part indicative of the attitude heterogeneity of the subset of subjects endorsing a particular (criterial) statement.***

It is perhaps important to draw a distinction between item characteristic and item index which is a mathematically defined quantity. An item index may be found to be associated with a given item characteristics but it is also possible that an item index is not related to any item characteristics but to a host of other factors outside the item itself. If an item index does tap one or more item characteristics, the relative magnitude of the index is expected to remain relatively unchanged over different testing sessions despite changes in the situation. But if the index is related to factors other than item characteristics, slight changes in the situation, e.g., a different group of subjects, a different set of items, a different ordering of items in the scale, are expected to affect the magnitude of the index.

The principal purpose of the present study was to present evidence concerning the reliability of the sensitivity and other related indices. The specific questions for which answers were sought were: (1) how reliable are these indices in terms of test-retest reliability; (2) are the magnitudes of these item indices affected by a change in the ordering of items within the scale (interform reliability); (3) what proportion of the total sum of squares of an endorsement set does the between-subjects sum of squares take up which forms the basis for the sensitivity index; and finally (4) how are these indices derived from an endorsement set related to various other item indices.

Method

Subjects

The subjects were 109 freshmen students attending two introductory psychology classes at Seoul National University, Seoul. Fifty-five of them came from one class and the remainder from the other class. The first group received Form A of an attitude scale and the second group Form B of the identical scale.

Attitude Scales

Two different forms of a Thurstone-type attitude scale were constructed by reshuffling the order of attitude statements (items) in the original scale. The original form will be referred to as Form A and the new form with reordered items as Form B. The two forms were identical except different ordering of the items. The attitude scale consisted of 31 attitude statements expressive of son preference attitude of varying intensity. Each form of this attitude scale was a 2-page booklet of 20cm×27cm in size, all 31 attitude statements appearing on the second page in double columns. For each attitude statement, its scale position and ambiguity were already known. The construction of this attitude scale (Form A) is described elsewhere (Cha, Kong, & Lee, 1973).

Precedure

Form A of the attitude scale was given to one class of 67 subjects, and Form B to another of 69 subjects. The experimenter introduced himself to the class, gave a brief description of the purose of the study and the task. Then the experimenter passed out the Thurstone-type attitude scale. Each of these classes were tested again with the identical form about two weeks *from* after the day of the first testing (Form A 14 days later and Form B 16 days later). In the retesting sessions, 69 subjects were present to take Form A and 61 subjects to take Form B. The first testings *occurred* were adminstered in May 1981, but Form A was administered 20 days ahead of Form B. Throughout the four testing sessions, it was the junior writer who administered the scale. The study was a 2 (2 forms)×2 (2 testings) design.

Results

The data used in the final analysis consisted of responses obtained from 55 subjects on Form A and 54 subjects on Form B. These figures represent the number of subjects remaining

Table 1. *The Within-subject(w), the Between-subjects(b), and the Deviation(d) Sums of Squares for Each Attitude Statement, by Scale Form and Test Session.*

Scale Form		Form A						Form B						
item No.	order	1st testing			2nd testing			order	1st testing			2nd testing		
		SS_w	SS_b	$3Nd^2$	SS_w	SS_b	$3Nd^2$		SS_w	SS_b	$3Nd^2$	SS_w	SS_b	$3Nd^2$
1	1	*—	—	—	2.24	0.47	2.54	10	—	—	—	—	—	—
2	27	0.97	0.00	0.76	1.74	0.05	1.47	28	1.86	0.00	3.57	1.15	0.06	1.69
3	7	48.73	17.05	51.18	49.00	26.06	45.10	20	41.38	10.63	47.19	62.45	19.06	76.23
4	18	0.41	0.02	0.38	0.37	0.02	0.15	11	—	—	—	—	—	—
5	11	3.07	2.60	0.30	12.94	7.30	1.19	29	14.4	1.72	6.34	9.92	1.48	2.72
6	12	16.35	10.16	0.65	11.02	6.30	0.01	23	7.27	4.18	1.00	6.54	3.07	0.78
7	19	29.36	16.77	9.12	40.38	35.61	12.10	8	28.87	15.90	7.36	16.41	14.04	1.21
8	5	32.96	16.79	0.14	24.01	15.82	0.00	22	16.92	14.01	0.27	33.36	21.91	6.16
9	8	27.35	8.46	2.06	13.24	7.82	0.81	27	27.2	18.48	2.12	37.03	11.35	6.59
10	16	15.84	15.07	2.88	20.69	6.62	4.56	19	31.46	22.22	1.56	26.14	11.41	11.06
11	25	6.17	5.89	1.02	5.98	1.71	0.24	13	14.08	4.7	5.94	5.22	1.36	0.96
12	10	40.1	19.33	22.85	46.99	19.46	24.20	17	28.00	19.09	8.64	57.12	18.43	33.67
13	2	—	—	—	8.25	1.42	6.39	28	1.83	5.29	0.54	7.74	3.35	0.86
14	9	—	—	—	—	—	—	21	0.48	0.65	1.01	—	—	—
15	20	7.72	6.43	3.76	8.56	3.43	5.39	3	26.42	14.33	16.48	33.11	9.37	20.58
16	6	20.43	5.85	17.01	27.03	6.06	27.29	18	7.52	5.88	2.34	5.54	3.51	5.65
17	28	—	—	—	—	—	—	26	—	—	—	—	—	—
18	22	—	—	—	—	—	—	25	0.37	2.34	3.69	3.07	0.42	4.13
19	24	11.78	5.51	15.57	7.84	2.81	6.05	14	11.97	11.60	8.67	—	—	—
20	14							5		—	—	—	—	—
21	30							12	10.41	2.7	11.76	15.47	0.30	15.68

* Sums of squares were not computed since the total number of endorsements received was fewer than 2.

after eliminating those who failed to be present in both the first and the second testings and those who failed to follow instructions adequately. In what follows, overall fidings on various components of endorsement sets will be presented first, with paricular attention to the proportions of three independent sums of squares, followed by evidence on the relationship between item scale position and three mean squares of the critical item (statement), evidence on the possible relationship between item popularity on one hand and the three sums of squares and their mean squares on the other, evidence on the relationship between item ambiguity and the three mean squares, evidence on the test-retest reliability of the three mean squares, and finally evidence on interform reliability of the same three mean squares.

(1) *Overall results on the three sums of squares and the proportions of the three sums of squares within endorsement sets.*

Table 1 presents the sums of squares of the three components of an endorsement for each item, separately for Form A and Form B and for each testing session. Only the results for 21 items corresponding to the left-most scale positions (i.e., less son-preferring positions)

Table 2. *Proportions of Within-subject, Between-subjects, and Deviation Sums of Squares in the Total Sum of Squares in Each Endorsement Set, Shown Separately for Form A and Form B.*

Scale Form		Form A					Form B					
item No.	order	No. of endorsements	SS_w/SS_t	SS_b/SS_t	$3Nd^2/SS_t$	SS_t	order	No. of endorsements	SS_w/SS_t	SS_b/SS_t	$3Nd^2/SS_t$	SS_t
1	1	3	0.43	0.09	0.48	2.63	10	1	—	—	—	—
2	27	7	0.55	0.01	0.45	2.49	28	5	0.36	0.00	0.63	4.17
3	7	55	0.41	0.18	0.41	118.56	20	35	0.40	0.12	0.48	128.47
4	18	5	0.57	0.03	0.40	0.68	11	1	—	—	—	—
5	11	18	0.58	0.36	0.05	13.72	29	12	0.66	0.09	0.25	18.29
6	12	29	0.62	0.37	0.01	22.25	23	21	0.61	0.32	0.08	0.11
7	19	43	0.49	0.37	0.15	71.67	8	31	0.60	0.36	0.10	41.90
8	5	35	0.64	0.36	0.00	44.86	22	26	0.54	0.39	0.07	46.32
9	8	31	0.68	0.27	0.05	29.87	27	35	0.63	0.29	0.08	51.39
10	16	20	0.56	0.33	0.11	32.83	19	34	0.55	0.31	0.12	51.93
11	25	11	0.58	0.36	0.06	10.51	13	19	0.60	0.19	0.21	16.13
12	10	32	0.47	0.21	0.26	91.72	17	43	0.52	0.23	0.26	82.48
13	2	5	0.51	0.09	0.40	8.03	28	5	0.48	0.43	0.09	10.04
14	9	0	*—	—	—	—	21	2	0.29	0.39	0.32	0.84
15	20	8	0.46	0.28	0.26	17.65	3	27	0.49	0.20	0.31	60.15
16	6	14	0.45	0.12	0.43	51.84	18	5	0.43	0.31	0.26	15.22
17	28	2	—	—	—	—	26	0	—	—	—	—
18	22	0	—	—	—	—	25	5	0.25	0.20	0.56	7.01
19	24	9	0.40	0.17	0.44	24.78	14	5	0.37	0.36	0.27	16.12
20	14	0	—	—	—	—	5	1	—	—	—	—
21	30	1	—	—	—	—	12	7	0.46	0.05	0.49	28.16
average			0.50	0.25	0.25	34.01	average		0.50	0.23	0.27	34.71

* Sums of squares were not calculated on either testing session, because the number of endorsements received did not exeed 2.

are presented because these are the positions receiving most endorsements from the subjects. To the right of the 21 item (counting from the non-son-preference end), there were no items which received more than two endorsements on any one testing. Table 2 presents the proportions of the three sums of squares in the total sum of squares, separately for Form A and Form B but averaged through the two testing sessions. The table also presents the number of endorsements given to each item (an index of item popularity). Since the items located near the center of the subjects' attitude distribution are more "popular" in the sense that they receive more endorsements, one would expect their total sums of squares to be larger in magnitude than those of other items. But a casual inspection of Table 2 for the total sums of squares shows that if there is any such tendency, it was by no means a pronounced one. The last rows in the table show that the proportions remain fairly stable across scale forms and that the within-subject component occupies about 50% of the total sum of squares, the between-subjects component about 25%, and the "skewness" or deviation component about another 25%. Thus, we have here an indication that only about 25% (corresponding to the between-subjects sum of squares) of the total sum of squares in an endorsement set is providing useful information about the criterial item, useful for selecting items in that only this portion of the total sum of squares provide a pure item index.

Within each component, there is a fairly large inter-item variability. If one confines his attention to results from form A, it may be noted that the within-subject sum of squares ranges from 40% to 68%, the between-subjects component from 1% to 37%, and the "skewness" or deviation component from 1% to 48%, depending on the item. Furthermore, one notices that there is a systematic relationship between the proportion of a component and the scale value of an item. The proportions for the scale position component are related to item scale value in an U-shaped function (Fig. 1) while those for the between-subjects component are related to item scale value in an inverted U-shape function. The proportions for the within-subjects component shows a similar inverted U-shape relationship with item scale value (Figs. 1a, 1b, and 1c). This means that although the absolute size of the sum of squares for the between-subjects component is not necessarily larger in more popular items, the proportion fo that component in the total sum of squares of an endorsement set is larger, the more popular an item is. Findings show that the proportions within a component are *positively* related to

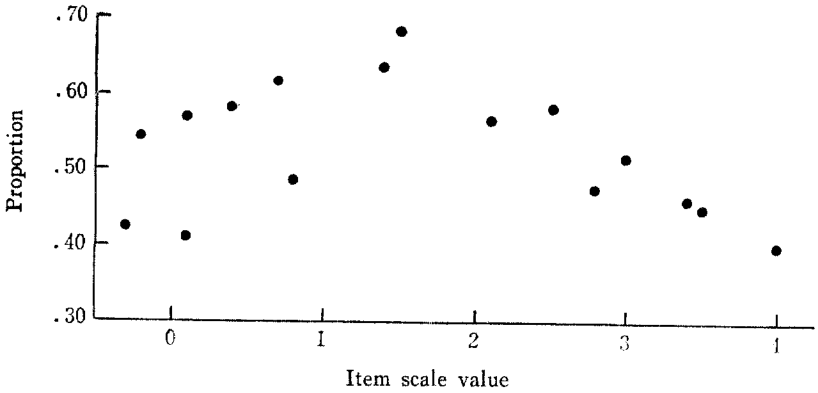

Figure. 1a. Proportion of Within-subject Sum of Squares for Each Item.

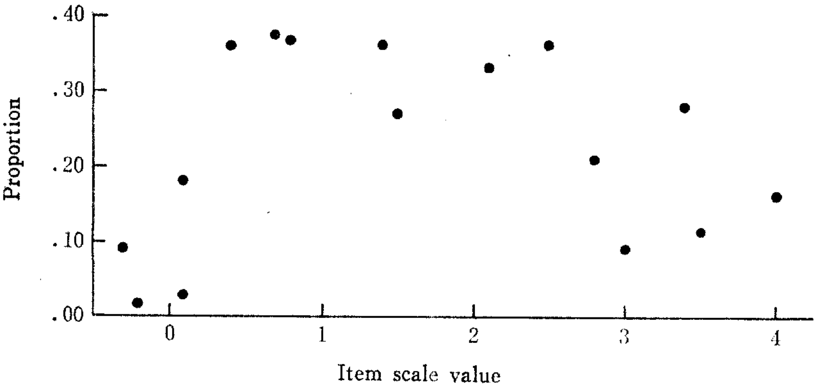

Figure. 1b. Proportion of Between-Subjects Sum of Squares for Each Item.

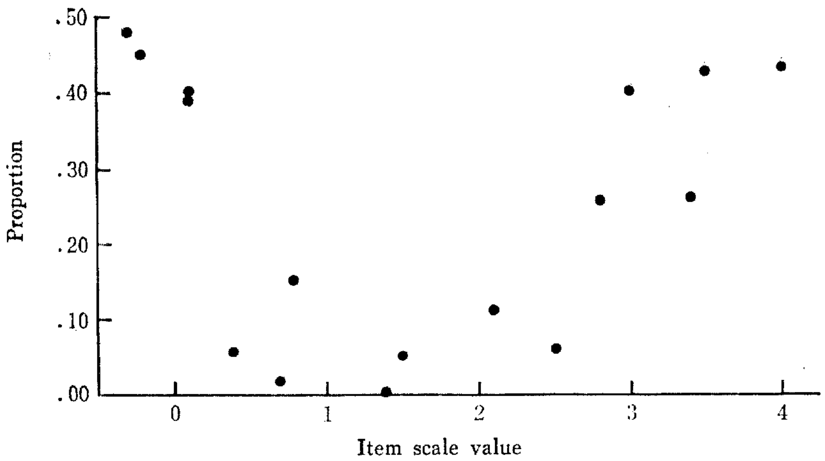

Figure. 1c. Proportion of Deviation Sum of Squares for Each Item.

item popularity in at least two cases, the within-subject component and the between-subjects component. These positive relationships may have been caused by the *negative* relationship that exists between the proportion of the deviation ("skewness") component and the popularity of the criterial item. However, these relationships are difficult to interpret because different proportions within an endorsement set are not mutually independent.

(2) *The relationship between item scale value and the three mean squares*

Since proportions of different components within an endorsement set are interrelated, they do not render themselves to simple interpretations. The sums of squares within an endorsement set are mutually independent, but they have another shortcoming, namely that they are positively related to item popularity. One obvious way to overcome this problem is to use instead the mean squares, that is, sums of squares divided by respective degree of freedom. The degree of freedom for the total sum of squares in this case is $3n-1$, and the degrees of freedom for the within-subject, the between-subjects, and the deviation or "skewness" components are $2n-1$, $n-1$, and 1, respectively. Since means squares are quantities which control for item popularity, their magnitudes can be compared easily between items. Even though mean squares control for item popularity (the number of endorsements received by the item), it is still possible that the mean square of a component show a relationship to item popularity. Either and U-shape or an inverted U-shape relationship of a mean square and item scale value would suggest a relationship between the mean square and item popularity. Mean square valuse of different components are presented in Table 3. Analysis showed that the within-subject mean square is positively related to item scale value ($r=.76$, $df=16$, $p<.01$, two-tail test) in the data obtained from Form A but not related to item scale value ($r=.25$, $df=15$, $p>.05$, two-tail test) in the data from Form B. The latter result is consistent with the finding from the earlier study (Cha & Lee, 1974) in showing no relationship between the within-subject mean square and item scale value. Again, the results on the relationship between the between-subject mean square (the sensitivity index) and item scale value are conflicting between the scale forms. With Form A, there was a significant positive relationship between the two variables ($r=.62$, $df=14$, $p<.05$,

Table 3. *Within-subject, Between-subjects, and Deviation Mean Squares in Each Endorsement Set, Shown Separately for Form A and Form B and Pooled over Two Testings.*

Scale Form		Form A				Form B				
Item No.	Order	No. of endorsements	MS_w	MS_b	MS_d	Order	No. of endorsements	MS_w	MS_b	MS_d
1	1	3	0.38	0.24	1.27	10	1			
2	27	7	0.22	0.01	1.12	28	5	0.38	0.03	2.63
3	7	55	0.91	0.81	48.14	20	35	1.53	0.89	61.71
4	18	5	0.11	0.02	0.27	11	1			
5	11	18	0.43	0.58	0.75	29	12	1.18	0.34	4.53
6	12	29	0.49	0.63	0.33	23	21	0.37	0.40	0.89
7	19	43	0.83	1.24	10.61	8	31	0.74	1.04	4.24
8	5	35	0.83	1.00	0.07	22	26	0.98	1.49	3.22
9	8	31	0.66	0.59	1.44	27	35	0.94	0.93	4.36
10	16	20	0.96	0.87	3.72	19	34	0.87	1.04	6.31
11	25	11	0.66	0.78	0.72	13	19	0.51	0.32	3.45
12	10	32	1.41	1.30	23.53	17	43	0.99	0.95	21.16
13	2	5	0.59	0.24	3.20	28	5	1.08	3.49	0.7
14	9	0				21	2	0.08	0.33	0.51
15	20	8	1.16	1.64	4.58	3	27	1.15	0.96	18.53
16	6	14	1.83	1.00	22.15	18	5	1.68	3.23	4.00
17	28	2				26	0			
18	22	0				25	5	0.55	0.80	3.91
19	24	9	1.32	1.26	10.81	14	5	0.86	1.94	4.34
20	14	0				5	1			
21	30	1				12	7	2.29	0.53	13.72
Sum		328					320			

two-tail test), but with Form B, the correlation was positive but not large enough to be statistically significant ($r=.40$, $df=15$, $p>.05$).

These results are inconsistent with earlier findings (Cha & Lee, 1974) which showed the correlation to be close to nil. Since Form A is the identical form used in the earlier study, it means that the same scale form yielded a correlational evidence diametrically opposed to each other in the two studies.

As for the scale position mean square, it failed to correlate significantly with item scale value in both forms ($r=-.08$, $df=14$ for Form A and $r=.21$, $df=15$ for Form B). It must be noted in passing that the scale position mean square 3 nd^2 is identical to its sum of squares. Since it was known beforehand that the quantity d^2 will be related to item scale value in a U-shape function, that is, curvilinearly, no significant correlation was expected using a Pearson r. But, if d rather than d^2 is used, this quantity will be negatively related to item scale value.

As expected, the "skewness" deviation score d was found negatively correlated to item scale value in both Form A ($r=-.43$, $df=14$, $p<.05$, one-tail test) and Form B ($r=.71$, $df=15$, $p<0.005$, one tail test). These last results corroborate an eariler finding (Cha & Lee, 1974) which showed that the two variables were correlated to each other with $r=.95$ (the reversed sign is apparently due to the error made in connection with calculating d in the earlier study). Above results are summarized in Table 4.

Table 4. *Pearsonian Correlation Coefficients Sshowing the Relationship between Item Scale Value on One Hand and Each of the Tthree Mean Squares and d, Separately for Form A and Form B.*

Components Sample	MS_w	MS_b	MS_d	d
Form A($df=14$)	.76**	.62*	−.08	$-.43^+$
Form B($df=15$)	.25	.40	−.21	$-.71^{++}$

* $p<.05$ two-tail test ** $p<.01$ two-tail test
$^+$ $p<.05$ one-tail test $^{++}$ $p<.01$ one-tail test

(3) *The relationship between item popularity and the three mean squares*

Item popularity of different items is shown in Fig. 2. Two facts are noteworthy about the distribution of endorsements shown in the figure. First, endorsements are approximately normally distributed within the subjects' attitude range, the items near the center of the range receiving the largest numbers of endorsements. Second, within this limits, there are individual variations, some items receiving more endorsements than others though similar in terms of their scale position while some other items receiving less than what would be normal for their scale positions.

As a first step the number of endorsements received by an item was correlated with each of the within-subject mean square, between-subjects mean square, scale position mean square, and d. None of the correlations was significant except he one involving the scale position mean square which showed a statistically significant positive correlation in Form A($r=.65$, $df=14$, $p<.01$, two-tail test) and an insignigicant but sizable correlation in Form B ($r=.46$, $df=15$, $p>.05$, two-tail test). However, these statistically significant and near significant positive correlations can be attributed to the factor n which enters into the scale position mean square ($=3\ nd^2$). In each item, the number of endorsements received (item popularity) is equal to n, the number of subjects endorsing the item.

The number of endorsement is a measure of overall item popularity which does not make distinction between the two components of item popularity mentioned earlier. Of the two components, the subjects' attitude distribution or simply the attitude component is of less interest for item selection purposes because it is a function of subjects' attitude and item scale value. The individual item component, though more interesting for the above purposes, is difficult to isolate. One rough way of testing for possible relationship between the individual item component of item popularity and the mean squares is to identify those items that have received a disproportionately large number of endorsements and those which have received less than their normal share of endorsements. From Fig. 2, it appears that Items 3, 6, and 11 seem to belong to the first group and Items 10, 12, and 15 to the second group. Ignoring the scale position component, the within-subject mean square values for the first group of items were, for Form A only with the two testing sessions averaged, .91, .49, and .66. The mean square values for the second group of items were .96, 1.14, 1.16, respectively. It would appear then that the high popularity items are characterized by a smaller within-subject variation as compared with the low popularity items. This implies that subjects endorsing very popular items tend to have a narrower latitude of acceptance (Sherif, Sherif, & Nebergall, 1965) or that popular items tend to

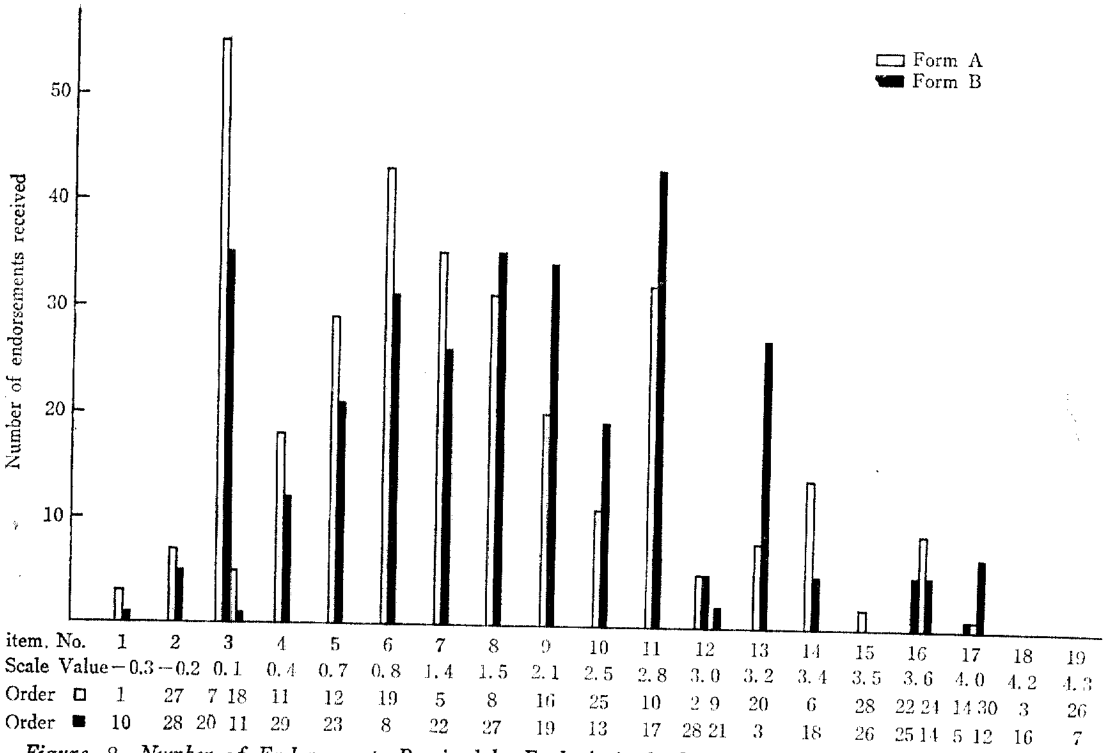

Figure. 2. Number of Endorsements Received by Each Attitude Statement (Item), Shown Separately for Form A and Form B.

attract people who have a narrower latitude of acceptance.

The between-subjects mean square values for the first group of items (the high popularity items), again confining our attention to data from Form A with the two testing sessions pooled, were .81, .63, and .78, respectively and corresponding figures for the second group of items (the low popularity items) were .87, 1.30, and 1.64, respecitively. The contrast between the two groups of items seems to suggest that the high popularity items are associated with high item sensitivity (small between-subjects mean square) while the low popularity items are associated with low item sensitivity. Thus, on the surface, it would appear that the high popularity items attract subjects who are homogeneous with respect to attitude position whereas the low popularity items attract subjects who are more dissimilar in attitude.

But, it is quite possible that the observed difference between high popularity and low popularity items in the size of both within-subject mean square and between-subjects mean square may simply reflect the positive correlation observed earlier between these mean squares on one hand and item scale value on the other (cf. Table 4). This scepticism is based on the fact that the low popularity items tended to have higher item scale values. More ideal situations will be to make comparisons between high and low popularity items with their scale positions held constant. But opportunities for making tests under such ideal situations are not easily available, and an inspection of available data (Table 3) at places where such tests could be made failed to produce any consistent trends.

(4) *The relationship between item ambiguity and the three mean squares*

For each item on the scale, item ambiguity was known through an earlier study (Cha, Kong, & Lee, 1973). Item ambiguity is defined as the quartile deviation of scale position ratings of an item by a group of judges. None of the mean square values were significantly correlated with item ambiguity, and this was true in both Form A and Form B. An earlier study (Cha & Lee, 1973) had shown that within-subject mean square was unrelated to item ambiguity.

(5) *Test-retest reliability of the three mean squares*

Table 5. *Pearson Correlation Coefficients Showing the Test-retest Reliability of Each of the Three Mean Squares and Number of Endorsements, Separately for Form A and Form B.*

Sample	MS_w	MS_b	MS_d	n
Form A(df=14)	.73**	.53*	.95**	.96**
Form B(df=15)	.36	.47*	.91**	.95**

* $p<.05$ one-tail tests
** $p<.01$

The test-retest reliability coefficients of three mean squares based on the three components of the total sum of squares of an endorsement set are presented in Table 5. It may be noticed that the reliability coefficients (Pearson r's) are generally higher in Form A than in Form B. Among the three mean squares, the deviation or "skewness" component shows the highest reliability. The reliability coefficients for the other two components are not large but substantial and statistically sigificant, the only exception being within-subject mean square under Form B, whose coefficient fell short of statistical significanee. These significant test-retest reliabilities mean that three mean squares do reflect certain aspects of item characteristics (and certain scale characteristics). These evidences are particularly significant for the within-subject mean square and the between-subjects mean square (the item sensitivity index) because they are thought to be measuring in part latitude of acceptance and endorsers' attitude heterogeneity, respectively.

Test-retest reliability coefficient was calculated for number of endorsements as well. As expected, there was a high significant correlations between the first and the second test on this measure (see the last column in Table 5). Since the attitude distribution of the subject sample remains essentially the same through two testings, the items located near the center of the attitude range will always receive the largest number of endorsements compared to other items farther removed from the center of the distribution. This correlation then largely reflects the scale position of items rather than their other characteristics.

(6) *Interform reliability of the three mean squares*

Another form of reliability is the interform reliability involving the correlation between the values of an item index taken from two different forms of the same attitude scale. In the present study, two forms were constructed out of an identical attitude scale, two forms differing only in the order of attitude statements. Although the two forms could have been given to one and the same group of subjects, in this study the two forms were given to two separate sample of subjects. There, the obtained interform reliability figures are probably smaller than what they would be if the two forms were given to the same subject.

Of the three mean squares, only the between-subjects mean square and the deviation mean square showed statistically significent reliability coefficients ($r=.61$, $df=17$, $p<.01$; $r=.85$, $df=17$, $p<.01$, one-tail tests). This statistically significant interform reliability for between-subjects mean square renders further support to the notion that this quantity (item sensitivity) indeed measures item specific characteristics.

As for the between-subject mean square, the statistically significant interform reliability obtained is even more significant because the two forms were given to two separate groups of subjects. No signifcant interform reliability was expected for within-subject mean square because the latter measures certain item characteristics that involve not only the criterial but also its neighboring other items. When the item order is changed, a given item's neighboring items would also change and as a result interform reliability is expected to be low or negligible.

This is what has been found. The significant interform reliabiity obtained for the scale position mean square was expected because two factors in the mean square quantity ($=3nd^2$), namely n and item scale position, would contribute toward a positive correlation. More specifically, the overall normal shape of attitude distribution will remain the two forms, and hence the shape of the distribution of n's. The quantity d is related to item scale position, which should remain constant across the two forms.

Inspection of Figure 2 seems to show that items that are overly popular or overly unpopular beyond their normal capacity to draw endorsements expected on the basis of their scale position remain as such despite changes in the order of statements and subjects. (Incidentally, Figure 2 also shows that the group of subjects who received Form B had a stronger son preference attitude than the one which received Form A.)

This group difference is perhaps due to the fact that the second group (Form A) contained more female students than the first. The second group contained 16 females as against 3 females in the first group.)

Discussion

The most important findings coming from this study are that within-subject mean square and between-subjects mean square show statistically significant test-retest reliability, and that as expected, the magnitude of between-subjects mean square was not affected by changes in item order whereas that of within-subject mean square was as seen through interform reliability of these two measures. These results are supportive of the initial notion that the between-subjects mean square measures something that is unique to the cons criterial item per se whereas the within-subject mean square is an item index which is influenced not only by the criterial but also by other, neighboring items.

One disturbing finding has to do with the correlation between the mean squares and item scale value. At least in Form A, both the within-subject mean square were positively correlated with item scale position (the scale position of the criterial item). What these correlations mean is not at all clear, and these correlations came as a surprise. Since the same correlations were not significant in form B, one perhaps should not attach any significance to the observed correlations. There were also indications in the data that the overly popular items as opposed to overly unpopular items might be associated with larger within-subjects mean squares, but more data are needed before any definite conclusion can be drawn.

The fact that between-subjects sum of squares occupies only one-fourth of the total sum of squares of an endorsement set on the average attests to the usefulness of between-subjects mean square as a purer alternative to other bases of item selection such as Thurstone's test of irrelevance in the second stage of Thurstone-type attitude scale construction. As noted earlier in the introduction, Thurstone's test of irrelevance basically relies on the total variance of an endorsement set. Isolation of between-subjects sum of squares from the total sum of squares and separating it from the other components means that one has a purer item index which does away with noises which comprise the other components. Even if one has the mean to isolate this purer component from the rest, the effect of separation this component will not be great if the ratio of noise to signal is small. The present study showed that on the average, each endorsement set contains sums of squares about 75% of which represent noise as far as information useful for the selection of items is concerned.

References

Cha, J. H. An analysis of Thurstone's test of irrelevance. *Research Notes* (Korean Institute for Research in the Behavioral Sciences), 1973, *2*, 135-143. (In Korean)

Cha, J. H., Kong, C. J., & Lee, E. O. Report on the construction of a son preference attitude scale. *Research Notes*, 1973, *2*, 168-172. (In Korean)

Cha, J. H., & Lee, E. O. Sensitivity: A new item index for construction of Thurstone-type attitude scales. *Korean Journal of Psychology*, 1974, *2*, 1-11. (In Korean)

Edwards, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.

Ferguson, L. W. *Personality measurement*. New York: McGraw-Hill, 1952.

Sherif, M., Sherif, C., & Nebergall, N. C. *Attitude and attitude change*. Philadelphia: Saunders, 1965.

Thurstone, L. L., & Chave, E. J. *The measurement of attitude*. Chicago: University of Chicago Press, 1929.