THOMAS J. REYNOLDS and JAMES P. JOLLY*

Three methods of gathering and evaluating value profiles for use in market segmentation are compared. Different reliability estimates are found to produce different conclusions as to the relative test-retest reliability of the three methods. However, the most appropriate statistical measure, Kendall's tau_B, reveals that Likert rating scales are significantly less reliable than either rank ordering or paired comparison procedures. No differences in subject interest are found among the methods, but Likert scales and rank ordering require less response time than do paired comparisons. Implications for method choice and future research are discussed.

Measuring Personal Values: An Evaluation of Alternative Methods

Munson and McIntyre (1979), in noting the need to study cultural differences as a basis for developing international marketing strategy, investigated alternative elicitation methods for the measurement of personal values (Rokeach 1973). The primary importance of such research is the potential application for segmentation purposes across international markets. The specific foci of their investigation were the measurement concerns, namely the issues of reliability and validity, that would offer insight into the selection of data-gathering methods to be used in obtaining value profiles.

The elicitation procedures that served as the basis for comparison were:

RANK: Assigning numerical ranks per Rokeach.
RATE: Likert ratings on seven-point scales.
ANCHOR: An anchored approach to the seven-point Likert scale which required the respondent to position the values having the lowest and highest importance at the 1 and 7 points, respectively.

The authors reported several findings related to reliability, namely:

1. The anchored approach had the poorest average test-retest reliability, which was significantly lower than that of the ranking procedure.
2. Although somewhat lower in test-retest reliability than the ranking procedure, the standard Likert scale was *not* significantly less reliable.

Munson and McIntyre conclude on the basis of these findings that the Likert rating approach is an appropriate replacement for the original rank ordering procedure recommended by Rokeach.[1]

Several issues of interest come to mind when one considers the Munson and McIntyre findings. First, and most simply, the measure of correlation (Spearman) used to analyze the resulting rank orders for test-retest reliability is not the most appropriate. As Hays (1973, p. 796) notes, the Spearman correlation measure is a biased estimator, except in a few rather special conditions. Kendall's tau, however, provides an unbiased estimate of the true population correlation. This difference is accentuated when the analysis is undertaken for moderate sample sizes, which is the case in the Munson-McIntyre study. Thus, the degree of generalization permitted from the use of the Spearman coefficient remains in question.

*Thomas J. Reynolds is Associate Professor, School of Management, University of Texas at Dallas and James P. Jolly is a doctoral student.

The authors gratefully acknowledge the many useful suggestions and substantial contributions provided by the reviewers. Correspondence should be addressed to the senior author at Young and Rubicam, New York.

[1] For the specifics of their method and the complete results, the original work should be consulted.

Journal of Marketing Research
Vol. XVII (November 1980), 531–6

A second issue is the choice of the elicitation procedures to be compared. Unfortunately, the methods selected permit only relative comparisons across techniques which one would expect to be less than optimal. Specifically, a more meaningful design would be the comparison of these elicitation techniques with a procedure generally regarded as providing maximal reliability, thus setting an absolute standard or benchmark to which the other methods could be contrasted. This reasoning leads one to consider a complete paired comparison approach from which a rank order can be easily derived. The rationale for paired comparisons centers on the fact that the amount of information processing for each comparison task is minimal (Mosteller 1951), resulting in an "easier" judgment task and therefore in less error in judgment. This choice is reinforced by the fewer assumptions relative to the decision-making process required for the simple paired comparison task (Kendall and Smith 1940). The desire to obtain the benchmark perspective and hence a more realistic comparison across elicitation techniques was one motivating force directing our research.

As Munson and McIntyre may have been misled in their findings of equivalence between the ranking and rating methods because of the inappropriateness of Spearman's rho as a measure of reliability and the insensitive analytical procedures which provided only relative comparisons, our study was performed to partially replicate and extend their work. Specifically, the study involved a comparison of the test-retest reliabilities of the RANK, RATE, and PAIRED methods, the last referring to complete (all possible) paired comparisons which served as a benchmark for comparative purposes. The ANCHOR method was not selected for comparison because of its poor performance in the original study. Our study also goes beyond Munson and McIntyre's in that Kendall's tau_B also was computed as a reliability estimate to investigate differences between that method and Spearman's rho. In addition to the reliability estimates, measures of subject interest and time were obtained for each subject with the hope that we could reach a more comprehensive conclusion by considering the tradeoffs between the statistical concerns and those of a more practical nature. The following null hypotheses reflect these issues.

Statistical

H_1: The reliability of the RANK and RATE procedures is equal to that of the benchmark method, PAIRED.

Practical

H_2: RANK and RATE require the same amount of respondent time as PAIRED.

H_3: RANK and RATE are perceived to be equally as interesting as PAIRED.

Table 1
STIMULUS ITEMS

A comfortable life (a prosperous life)[a]
An exciting life (a stimulating, active life)[a]
A sense of accomplishment (lasting contribution)
A world at peace (free of war and conflict)[a]
A world of beauty (beauty of nature and the arts)[a]
Equality (brotherhood, equal opportunity for all)[a]
Family security (taking care of loved ones)
Freedom (independence, free choice)[a]
Happiness (contentedness)[a]
Inner harmony (freedom from inner conflict)
Mature love (sexual and spiritual intimacy)[a]
National security (protection from attack)[a]
Pleasure (an enjoyable, leisurely life)[a]
Salvation (saved, eternal life)[a]
Self-respect (self-esteem)
Social recognition (respect, admiration)[a]
True friendship (close compansionship)
Wisdom (a mature understanding of life)

[a]Item was used for the $n = 12$ stimulus set size (N12).

METHOD

Sample and Procedures

The participants in the first study were 54 undergraduate business students enrolled in an introductory marketing class. Subjects were told they would be participating in a marketing research demonstration of different methods of obtaining preference orders on a group of stimuli. Their average age was 28.5 years; 32 men and 22 women participated.

The stimuli chosen for the actual preference ordering were Rokeach's (1973) 18 "final" or "terminal" values. For reasons of economy and parsimony Rokeach's "instrumental" values were not included.[2] The complete set of stimuli is listed in Table 1.

Design and Measures

At the time of their first visit, subjects were assigned randomly to one of six conditions based on the 2 × 3 design denoting size of stimulus set (SIZE) and elicitation method (METHOD). In the $n = 18$ stimuli condition (N18) all 18 values constituted the stimulus set. The $n = 12$ stimuli condition (N12) was composed of a subset of 12 values randomly selected from the original 18. The different set sizes were included to ensure generalizability of the research findings.

Elicitation Methods

Subjects also were assigned randomly to one of three levels of the second factor, method of measurement. The three levels were RANK, RATE, and PAIRED.

The first two measures were gathered from conventional paper and pencil instruments. The first, RANK, as recommended by Rokeach (1973) consisted of a

[2]In the Munson and McIntyre study terminal values were shown to be more reliable than instrumental values for all techniques. Therefore, terminal values were thought to be sufficient for inclusion here.

Table 2
INTERCORRELATIONS, MEANS, AND STANDARD DEVIATIONS OF THE DEPENDENT VARIABLES

	(1)	(2)	(3)	(4)	*M*	*SD*	*N*
(1) TIME	—	−.15	−.14	−.18	12.98	13.10	54
(2) INTEREST		—	.32	.34	5.52	1.09	54
(3) RHO			—	.96	.73	.18	54
(4) TAU_B				—	.63	.18	54

randomly ordered list of the 18 (or 12) stimulus items. The subjects were instructed to rank the values in order of the items' perceived importance to them. A 1 was used to represent the most important item and an 18 (or 12) to represent the least important.

In the RATE condition subjects were presented with a series of 18 (12) seven-point Likert rating scales, one scale for each of the stimulus items. The scales ranged from 1 = "not at all important" to 7 = "extremely important."

The paired comparisons (PAIRED) condition was administered through the use of a minicomputer. Subjects were seated at a computer terminal and, after being familiarized with the terminal and keyboard, were shown pairs of stimuli. The pairs were presented one at a time and subjects responded by pressing either a 1 or a 2 corresponding to the position of the stimulus item which they considered to be more important. The number of pairs differed according to the sample size factor, 153 for N18 and 66 for N12.

Dependent and Demographic Measures

After the subjects had completed the experimental procedure they were asked to respond to a one-item measure designed to ascertain their INTEREST in the task. The TIME required to complete the task also was recorded, along with age and sex information. The same procedure was repeated during the second session approximately two weeks later. The correlation between the orders derived from the first and second sessions provided the measure of test-retest reliability. Table 2 shows the intercorrelations and descriptive statistics for the dependent variables—INTEREST, TIME, and the RELIABILITY measures. Table 3 lists the descriptive statistics for those dependent measures for each of the experimental cells.

Analyses

To assess the test-retest reliability across techniques, following Munson and McIntyre, we computed a Spearman's rho rank order correlation coefficient between the Time 1 and Time 2 orders for each subject. In addition, a Kendall's tau_B rank order coefficient was computed.[3] Transformations of these two mea-

[3] Tau_B was used instead of tau_A because of the presence of tied ranks in the data.

Table 3
MEANS, STANDARD DEVIATIONS, AND SAMPLE SIZES FOR THE EXPERIMENTAL GROUPS

		N12				*N18*				*Total*		
Method	*DV*	*M*	*SD*	*N*		*M*	*SD*	*N*		*M*	*SD*	*N*
Paired	Rho (*z*)	.77	.24	9		.67	.24	10		.72	.28	19
	Tau_B (tau_B')	.66	.25	9		.57	.21	10		.61	.23	19
	Time	20.56	3.71	9		35.50	9.73	10		28.42	10.59	19
	Interest	6.11	.78	9		5.00	1.56	10		5.53	1.35	19
Rank	Rho (*z*)	.76	.18	8		.78	.11	10		.77	.14	18
	Tau_B (tau_B')	.62	.19	8		.63	.12	10		.63	.15	18
	Time	5.33	1.32	8		5.10	1.37	10		5.21	1.32	18
	Interest	5.78	.67	8		5.40	1.08	10		5.58	.90	18
Rate	Rho (*z*)	.75	.13	8		.66	.16	9		.70	.15	17
	Tau_b (tau_B')	.69	.13	8		.62	.16	9		.65	.15	17
	Time	3.38	1.41	8		4.56	1.88	9		4.00	1.73	17
	Interest	5.50	1.20	8		5.44	.88	9		5.47	1.01	17
Total	Rho (*z*)	.76	.18	25		.70	.18	29				
	Tau_B (tau_B')	.66	.19	25		.61	.16	29				
	Time	10.00	8.22	25		15.41	15.88	29				
	Interest	5.81	.90	25		5.28	1.19	29				

Note: *z* and Tau_B' are linear transformations of the reliability measures and are described in footnote 3.

sures were then used as dependent variables in two separate 2 × 3 analyses of variance of SIZE and METHOD. Table 3 also contains descriptive statistics for these transformations.

RESULTS

Reliability

Comparing the mean reliability estimates (see Table 3) for each of the three METHODS provides several interesting points. First, if we look only at the Spearman as our estimate of reliability the results are almost identical to those of Munson and McIntyre in that the three METHODS have approximately equal reliabilities. Furthermore, the tau_B's also appear to be about equal across METHODS. However, scanning mean values can be very misleading because of differences in scales of measurement and instability of error variances. Only through the analyses of variance of the appropriate transformations of these reliabilities can meaningful comparisons be made. The results of these analyses, presented in Table 4, indicate that the only significant effect in either of the 2 × 3, SIZE × METHOD analyses is due to the main effect for METHOD. Because, as expected, neither the SIZE nor the interaction effect reached significance, the two SIZE conditions were collapsed and a separate *post hoc* analysis was performed on the combined sample for each dependent reliability measure. These one-way *post hoc* analyses of METHOD by Scheffé's (1959) procedure are reported in Table 5.

Examination of Tables 4 and 5 illustrates the importance of the choice of a reliability measure on one's conclusions about the various value elicitation methods. The first panel of Table 4 reports the ANOVA results for Spearman's rho. If this were the only result available, as indeed it was for Munson and McIntyre, we might be led to conclude that, as there is no significant METHOD effect, the rating scale method is as reliable as the other techniques. However, the final panel, for Kendall's tau_B, clearly indicates alternative implications because of the significant METHOD effect. To further investigate this effect one must examine Table 5. Here it is clear that the significant METHOD effect can be attributed to the difference between the RATE condition and the other two relatively equal conditions. Therefore, *if* we use Kendall's tau_B as our criterion, we must reject H_1 for RATE, as this procedure is significantly less reliable than the benchmark. Thus, with Spearman's rho as our criterion of reliability our results lead us to agree with the results of Munson and McIntyre; however, with the more appropriate Kendall measure, the data clearly *do not* support their recommendation for the use of Likert rating scales to obtain value profiles. The RANK condition, notably, is not significantly less reliable than the benchmark, PAIRED, and therefore H_1 cannot be rejected for that technique. In terms of test-retest reliability, then, rank ordering does appear to be a desirable technique, even in comparison with a theoretically maximal technique.

Table 4
ANALYSES OF VARIANCE ON RELIABILITY MEASURES

Source	*df*	*MS*	*F*
Spearman (z)			
Method	2	.25	1.12
Size	1	.47	2.12
Method × Size	2	.03	.14
Error	48	.22	
Tau_B (tau_B')			
Method	2	.35	5.71[a]
Size	1	.13	2.13
Method × Size	2	.02	.40
Error	48	.06	

[a] $p < .01$.

Note: The reliability measures used as dependent variables in these analyses are transformed values. See footnote 3 for explanation.

$$z = 1/2 \log_e \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right)$$

$$\text{Tau}_B' = \sqrt{\frac{2 \arcsin \cdot \text{concordant pairs}}{\sqrt{\text{total} - 1/2 \Sigma t_1(t_1 - 1)}\ \sqrt{\text{total} - 1/2 \Sigma t_2(t_2 - 1)}}}$$

Table 5
POST HOC (SCHEFFE) ONE-WAY ANALYSES OF VARIANCE OF RELIABILITY MEASURES

Reliability measure	*Mean transformed reliability*	*Significant difference*
Z		
Paired	1.08	None
Rank	1.15	
Rate	.93	
Tau_B'		
Paired	2.20	Rate vs. paired and rank
Rank	2.27	
Rate	2.00	

Note: The values reported here are transformed values. See footnote 3 for explanation.

$$z = 1/2 \log_e \left(\frac{1 + r_{xy}}{1 - r_{xy}} \right)$$

$$\text{Tau}_B' = \sqrt{\frac{2 \arcsin \cdot \text{concordant pairs}}{\sqrt{\text{total} - 1/2 \Sigma t_1(t_1 - 1)}\ \sqrt{\text{total} - 1/2 \Sigma t_2(t_2 - 1)}}}$$

Practical Considerations

Table 6 gives the results of the multivariate analysis of variance (MANOVA) on the dependent measures,

Table 6
MULTIVARIATE ANALYSIS OF VARIANCE ON COMPLETION TIME AND INTEREST

Dependent variable	Source							
	Size		Method		Size × Method		Model	
	df	*F*	*df*	*F*	*df*	*F*	*df*	*F*
Time[a]	1	31.03[b]	3	239.62[b]	3	4.33[c]	7	108.98[b]
Interest	1	4.02[d]	3	.05	3	.66	7	.88
MANOVA	2,81	14.55[b]	6,162	57.06[b]	6,162	2.41[d]		

[a]The natural logarithm of time was used in this analysis to stabilize error variances.
[b]$p < .001$.
[c]$p < .01$.
[d]$p < .05$.

INTEREST and the log of TIME.[4] Two points are important.

First, for the dependent variable TIME, the significant main effect for METHOD is definitive in relation to our second null hypothesis, H_2, which predicts no differences in required respondent time between either RANK or RATE and the benchmark, PAIRED. Because of the significant MANOVA *F*, a *post hoc* Scheffe test was again performed, this time with METHOD as the independent variable and the log of TIME as the dependent variable.[5] The results of this test reveal that both RANK and RATE required less respondent time than did PAIRED ($p < .0001$). Thus, H_2 can be rejected for both the RANK and RATE techniques. The results further reveal the RATE procedure to be significantly less time consuming than RANK. One might be tempted to interpret this finding as an argument for the use of the rating scale method. However, the *practical* significance of the approximately one-minute reduction in time appears to be questionable in comparison with the saving of more than 20 minutes which these two techniques provide over paired comparisons.

Second, INTEREST is found to be significantly affected only by the SIZE condition. An examination of means reveals that within each METHOD condition INTEREST decreased as the number of stimuli increased. However, the Scheffe tests reveal no significant differences in interest among any of the three methods in either of the SIZE conditions; thus, on this basis, H_3 cannot be rejected for either RANK or RATE.

DISCUSSION AND CONCLUSIONS

Munson and McIntyre (1979) concluded that Likert rating scales are an appropriate replacement for the rank ordering procedure of obtaining value profiles originally recommended by Rokeach (1973). However, their use of Spearman's rho instead of Kendall's tau as a measure of test-retest reliability and the fact that the design of their study permitted only relative comparisons cast doubt on the generalizability of their findings.

Our study demonstrates that the conclusions reached by using rho as a measure of test-retest reliability can be very different from those drawn from the more appropriate transformation of Kendall's tau. The study also reveals that when this appropriate statistic is used, and is used in a design allowing comparison of the procedures with a theoretically maximal procedure, paired comparisons, the endorsement of the RATE procedure is unwarranted. The rating scale method proves to be significantly less reliable than either the rank ordering or the paired comparison procedure. Moreover, the practical advantages of Likert scales

[4]The actual rank order correlation coefficients cannot be used as dependent variables in an analysis of variance because of differences in scales of measurement and instability of error variances. These differences violate interval scale assumptions and make direct comparison of correlations meaningless. For this reason, appropriate transformations were used which create linear scales upon which values can be compared accurately across methods. In the case of Spearman's rho the appropriate transformation is Fisher's *r* to *z*:

$$z = 1/2 \log_e \left(\frac{1 + rxy}{1 - rxy} \right).$$

For the Kendall measure, because tau_B is equal to a ratio of proportions, a transformation recommended by Winer (1971, p. 400) was applied.

$$\text{Tau}_B = \frac{\text{concordant pairs} - \text{discordant pairs}}{\sqrt{\text{total pairs} - 1/2 \Sigma t_1(t_1 - 1)}\ \sqrt{\text{total pairs} - 1/2 \Sigma t_2(t_2 - 1)}}$$

where t_1 and t_2 are the numbers of tied observations in each group of ties on the Time 1 and Time 2 ranks, respectively. Therefore, the transformation yields

$$\text{Tau}_B' = \frac{2 \arcsin \cdot \text{concordant pairs}}{\sqrt{\text{total} - 1/2 \Sigma t_1(t_1 - 1)}\ \sqrt{\text{total} - 1/2 \Sigma t_2(t_2 - 1)}}.$$

[5]A logarithmic transformation was used to stabilize error variances.

referred to by Munson and McIntyre are found to be nonexistent in terms of subject interest in the task and, at best, of questionable *practical* significance in terms of the respondent time required.

In conclusion, our data do not support Munson and McIntyre's recommendations for the use of Likert scales. Although a definitive statement cannot yet be made as to the most appropriate data collection method for use with value profiles, clearly the abandonment of Rokeach's original ranking technique for the "quick and dirty" Likert scale method appears unwarranted.

REFERENCES

Hays, W. L. (1973), *Statistics for the Social Sciences*. New York: Holt, Rinehart and Winston, Inc.

Kendall, M. G. and B. Smith (1940), "On the Method of Paired Comparisons," *Biometrika*, 31, 324–45.

Mosteller, F. (1951), "Remarks on the Method of Paired Comparisons: III. A Test of Significance for Paired Comparisons When Equal Standard Deviations and Equal Correlations are Assumed," *Psychometrika*, 16, 207–18.

Munson, J. M. and S. H. McIntyre (1979), "Developing Practical Procedures for the Measurement of Personal Values in Cross-Cultural Marketing," *Journal of Marketing Research*, 16 (February), 48–52.

Rokeach, M. (1973), *The Nature of Values*. New York: Free Press.

Scheffé, H. A. (1959), *The Analysis of Variance*. New York: John Wiley & Sons, Inc.

Winer, B. J. (1971), *Statistical Principles in Experimental Design*. New York: McGraw-Hill Book Company.