RESPONSE-ORDER EFFECTS IN LIKERT-TYPE SCALES

JASON C. CHAN
University of Texas at Austin

The present study argued that the meaning of verbal labels of a Likert-type response scale was affected by the presentation order of the scale labels. It was proposed that subjects tended to choose the first alternative acceptable to them from among the ordered response categories so that a primacy effect was predicted. Findings supported the hypothesis. In addition, this response-order effect interfered with the threshold values, with factor structures estimated by factor analysis based on polychoric correlations, and with the item and person parameters estimated by the graded response model. Practical implications of the response-order effects were discussed.

In psychological research, the response scale for Likert-items is usually presented first with the most positive descriptor (e.g., strongly agree, agree, neutral, disagree, strongly disagree). If the presentation order is reversed (e.g., strongly disagree, disagree, neutral, agree, strongly agree), will the subjects' responses be affected? If the answer is "yes," toward what direction will these responses shift? Will this shift affect the estimation of item parameters, latent traits, and latent structure in factor analysis and item response theory (IRT)? Attempting to answer these questions is the purpose of the present study.

Serial Position Effects

If scale labels/descriptors have a standard, objective, and context-free meaning, then reversing their presentation order should not

affect subjects' responses. This view was implicitly or explicitly assumed by some researchers who tried to assign fixed ranking or numerical values to those frequently used scale quantifiers (e.g., Bass, Cascio, and O'connor, 1974; Hakel, 1968). This view, however, has been severely challenged.

The meaning of the verbal label of a scale may depend upon the contexts of the label. One important context to be considered is the position of the label. Chase (1969) suggested that the meaning of the scale adjectives was determined by the relative position of the adjective in a group of response categories rather than by the "standard" definition of the scale labels. This suggestion was consistent with Wildt and Mazis' (1978) findings that both label and location had an impact on subjects' responses. Klockars and Yamagishi (1988) also found that the meaning of the labeled position was defined as a compromise between the label itself and the relative position. Consequently, as stated by Worcester and Burns (1975), "the problem is not just that different words mean different things but that the same word can be made to mean different things as the context changes (p. 182)."

Reversing the order of response labels implies not only changing the position of the scale label but also reversing the sequence of subjects' information-processing. In this situation, a primacy or a recency effect was frequently considered. As early as 1929, Mathews found that subjects were more likely to endorse response options printed on the left side of the page than those on the right side. This primacy effect was also found when respondents were asked to make preference choices from a long list of political candidates (Brook and Upton, 1974; Mueller, 1970), from a list of 16 radio programs (Becker, 1954), or from a set of children's qualities (Krosnick and Alwin, 1987). Belson (1966) asked his subjects to rate 36 questions about television on five kinds of intensity scales. He found a clear primacy effect: The categories at the positive end of scales in the traditional order (favorable to unfavorable) or at the negative end of scales in reversed order received greater endorsement when they were presented first than when presented last. This effect was replicated by Payne (1972) in a postal survey, although Payne maintained that the effect was smaller than that realized by other investigations. In regard to evaluative items, Carp (1974) found that either the positive or negative end of the scale drew more responses when it was presented first.

In contrast, some researchers have also reported recency effects, i.e., a response was more likely to be selected when it was given last or near the bottom of a list. Those recency effects tended to occur

for dichotomous items (McClendon, 1986; Schuman and Presser, 1981) or for response options presented in oral form (Rugg and Cantril, 1944). Finally, a few researchers have failed to obtain clear evidence for serial position effects of response options (Johnson, 1981; Powers, Morrow, Goudy, and Keith, 1977).

Krosnick and Alwin (1987) developed a theorectical model to explain the underlying processes of responding to a scale. According to this model, Simon's (1957) satisfying principle is relevant for visually presented options. People usually respond with the first satisfactory or acceptable option in order to minimize cognitive costs rather than perform an exhaustive search for optimal solutions. Consequently, when the list of options is long and when a number of options is similar or equally appealing, a primacy effect is predicted. Orally presented options tend to produce recency effects because the capacity of short-term auditory memory is limited.

The present study employed 5-point intensity scales presented in written form. As it is usually difficult for subjects to differentiate adjacent categories (e.g., slightly strong agreement versus moderately strong agreement) in finely anchored scales, a primacy effect was predicted.

Form-Resistant Psychometric Parameters

Even though it was known that the form of response scales might have some effects on responses, a "form-resistant correlation" hypothesis (Stouffer and DeVinney, 1949) was usually assumed. If the form of response scales has a constant effect on every subject who responds to the same form, this hypothesis may be tenable because Pearson correlations are scale-free. However, it seems dangerous to place uncritical reliance on the just stated assumption. Krosnick and Alwin (1987; 1988) have tried to test this hypothesis empirically for different form effects. They found that changes in response order altered both the variances and the covariances of the items. However, they also found that the loadings of the items on the latent factor and the variance of the factor were invariant with regard to item/response order.

Based on assigning successive integers to response categories and obtaining the Pearson correlations, the just cited factor analysis assumed that the observed variables were measured by equal interval scales and were normally distributed. However, Likert-type scales are more reasonably treated as ordinal scales which measure an underlying continuous variable or trait. In this case, individuals' responses are jointly determined by the level of the latent trait and

the threshold value of the response scales. In the LISREL VI program (Jöreskog and Sörbom, 1984), the threshold values are estimated from the inverse of the normal distribution function, and then polychoric correlations are computed. Finally, the program performs confirmatory factor analysis on the matrix of polychoric correlations. The present study will examine the impacts of reversing the response scale order on factor structures estimated by factor analysis with polychoric correlations.

In dealing with ordinal multiple indicators which are assumed to measure one latent dimension, Samejima's (1969) graded response model is an appropriate alternative to factor analysis. This model is interesting in the current study because it has the property of invariance of person and item parameters if it fits the data. This property is known as "item-free person measurement" and "sample-free test calibration" in IRT models (Wright, 1967). The present study predicted that response order effects would produce item bias.

Method

Subjects

A total of 102 students (49 males, 53 females) from two senior high schools in Taipei city participated in this study as part of the activities for counseling and guidance classes. Their ages ranged from 15 to 17.

Instruments

Five items from the Personal Distress (PD) Scale, a subscale of the Interpersonal Reactivity Index (Davis, 1980), were translated into Mandarin by Chan (1986). One example of these items is "When I see someone who badly needs help in an emergency, I go to pieces." Each item was accompanied by a 5-point Likert-type response scale.

Procedures

In the first administration of the five items, the response scale labels were from the positive end to the negative end (from left to right) as follows: "describes me very well (4)," "describes me quite well (3)," "describes me well (2)," "describes me slightly well (1)," and "does not describe me well (0)." The scale form with this response order will be called the "positive form." Five weeks later,

in the repeated administration of the same test, the just stated response order was reversed. The latter form will be called the "reversed form."

Results

Mean Differences

If successive integers were assigned to response categories and summed over the items to obtain the total score, a systematic pattern of mean differences was evident. The positive response order tended to produce higher means on four out of five items compared to those on the reversed order. An overall multivariate test indicated that these mean differences between two forms of response order across five items were significant, $F(5, 96) = 3.47$, $p < .01$. Subsequent univariate t tests indicated that, among the four predicted mean differences, two were significant, $t(101) = 4.13$, $p < .001$ and $t(101) = 2.09$, $p < .05$; one was marginally significant, $t(101) = 1.96$, $p < .053$; and the remaining one was not significant. When those effects on items were accumulated at the test level, the total score was significantly higher for the positive response order than for the reversed order, $t(101) = 3.31$, $p < .001$. Those effects supported the primacy effect hypothesis. No significant recency effect was found.

Factor Analysis

Path coefficients and residuals estimated by LISREL with polychoric correlations were presented in the left half of Table 1. Because the chi-square value provided by LISREL VI was not correct with polychoric correlations, it was ignored. However, the other two goodness-of-fit measures, the Adjusted Goodness-of-fit Index (AGFI) and the Root Mean-square Residual (RMR), suggested that the one-factor model fit the positive-form data fairly well, $AGFI = .968$, $RMR = .034$, whereas it did not fit the reversed-form data adequately, $AGFI = .799$, $RMR = .086$. These results were confirmed by the factor analysis with Pearson correlations: The one-factor model fit the positive-form data very well, $Chi\text{-}square(5) = 2.82$, $p < .727$, whereas it was rejected as an inappropriate model for the reversed-form data, $Chi\text{-}square(5) = 24.01$, $p < .001$.

The threshold values estimated for computing polychoric correlations were presented in the left half of Table 2. It is interesting to

TABLE 1
Discrimination Parameters Estimated by the Graded Response Model and Path Coefficients Estimated by a One-Factor Model

	Path Coefficients				Discrimination in IRT			
	Pos. Form		Rev. Form		Pos. Form		Rev. Form	
Item	LV	Res.	LV	Res.	a	SE	a	SE
1.	.811	.343	.906	.179	2.550	.411	3.157	.545
2.	.418	.825	.564	.682	.829	.247	1.183	.247
3.	.811	.343	.881	.223	2.504	.399	3.011	.602
4.	.692	.521	.747	.442	1.639	.330	1.916	.393
5.	.576	.668	.510	.740	1.170	.308	1.059	.300

Note. Pos. = Positive; Rev. = Reversed; LV = Latent Variable; Res. = Residuals; a = Discrimination parameters; SE = Standard error of estimates.

note that the threshold values for the positive form were systematically lower than those for the reversed form—a finding suggesting that the positive form was "easier" than the reversed form. This result converged with that of the IRT analysis described in the next paragraph.

IRT Analysis

Item and person parameters were estimated through the MULTILOG program (Thissen, 1986). Discrimination and difficulty parameters, which are conceptually corresponding to factor and threshold values in the factor analysis, were presented in the right half of Table 1 and Table 2, respectively. The two sets of discrimination parameters for positive and reversed forms had a Pearson r of .9794, whereas the two sets of difficulty parameters had a Pearson r of .8515. From these two tables, it was shown that items with positive form of response scales were less discriminative and "easier" than those with reversed form of response scales. These results were consistent with the previous findings from factor analysis. Person parameters for the 102 people were also estimated. With the positive-form test, person parameters were estimated to have a mean of -0.024 and a standard deviation of 0.886. On the other hand, with the reversed-form, a mean of -0.230 and a standard deviation of 0.950 resulted. The two means differed significantly from each other, $t(101) = 2.81$, $p < .01$. This result implied that the two forms of tests produced different estimation of the latent trait of the same individuals. The two sets of person parameters had a Pearson r of .6835.

TABLE 2
Estimated Threshold Values from Factor Analysis and Item Difficulty Parameters from the Graded Response Model

Item	Factor Analysis T1	T2	T2-T1	The Graded Response Model b1	(SE)	b2	(SE)	b2-b1
1.	−1.19	−1.13	.06	−1.427	(.230)	−1.289	(.184)	.138
	−.27	−.06	.21	−.411	(.141)	−.133	(.137)	.278
	.25	.37	.12	.223	(.151)	.403	(.138)	.207
	.86	1.09	.23	1.024	(.190)	1.291	(.194)	.087
2.	−1.05	−.88	.17	−2.401	(.806)	−1.589	(.425)	.812
	−.51	.06	.57	−1.185	(.463)	.050	(.255)	1.235
	−.02	.62	.64	−.106	(.348)	1.041	(.362)	1.147
	.75	1.18	.43	1.665	(.609)	2.141	(.556)	.476
3.	−1.24	−1.00	.24	−1.503	(.228)	−1.160	(.170)	.343
	−.05	.24	.29	−.149	(.148)	.218	(.144)	.367
	.46	.56	.10	.502	(.160)	.606	(.155)	.104
	1.09	1.35	.26	1.340	(.233)	1.671	(.264)	.331
4.	−.38	−.26	.12	−.622	(.214)	−.424	(.184)	.198
	.49	.62	.13	.600	(.226)	.753	(.201)	.153
	1.05	1.29	.24	1.482	(.313)	1.742	(.362)	.260
	2.06	1.56	−.50	3.108	(.720)	2.242	(.493)	−.866
5.	.27	.21	−.06	.417	(.269)	.393	(.309)	−.024
	1.05	1.18	.13	1.836	(.471)	2.260	(.634)	.424
	1.76	1.65	−.11	3.244	(1.109)	3.276	(.923)	.032
	*	1.89	*	*	(*)	3.861	(1.17)	*

Note. T1 (or T2) = Threshold values for the positive (or reversed) form; b1 (or b2) = Difficulty parameters for the positive (or reversed) form; SE = Standard error of estimates; * = No person responded to the highest category, so that number of threshold is reduced for this item.

Discussion

The present study revealed that the order of response scale labels had a primacy effect on subjects' choices of the alternatives in Likert-type attitude scales. This effect altered the factor structures underlying both the Pearson and the polychoric correlation matrices of the two scale forms. In addition, threshold values estimated by the LISREL VI program and difficulty parameters estimated by the graded response model revealed that items in the positive form were "easier" than those in the reversed form. When the graded response model was used, individuals responding to the items with positive response order were also estimated to have "higher" levels of the latent trait than the same persons who responded to the items with reversed order.

How is it possible that the one-factor model fitted the positive-form data better than the reversed-form data while the positive-form

items had lower factor loadings and discriminative ability than the reversed-form items? It is possible because the data points in a correlation matrix are not completely independent from one another. If one of the correlations in a matrix is peculiar (e.g., too high), then the one-factor model will fit the data poorly even though the estimated loadings and discrimination of each variable are all high.

For IRT models, the sample size of the present study may be too small to produce an accurate estimation of the item parameters. In addition, five items are also insufficient to yield a stable estimation of person parameters. However, as long as the patterns of the results are revealing, they deserve further studies with more items and larger sample size.

The present study avoided the confounding factor of sampling fluctuation by using repeated samples. Nonetheless, as no control group was employed, the attribution to primacy effect might be weakened by some rival hypotheses such as maturity or selection. When one considers the systematic findings and their implications, however, a further comprehensive study of the response order effects is warranted.

What is the practical implication of the present findings? Researchers using Likert-type scales routinely recode or flip the response scales of the negatively worded items. The present findings suggested that the recoded scores would differ from the raw scores obtained from the actual situation where the order of the response scales was reversed. In a longitudinal study, a group-comparison study, or a test-retest experiment of attitude change using Likert-type scales or other similar types of questionnaires as instruments, researchers may accidently shift the response scale order and confound its effects with interesting variables. Some researchers may even explicitly assume that the effect of reversing the response scale is negligible. The present study has suggested that the effect may influence the raw-score means, both factor structures and threshold values estimated by the factor analysis, and item/person parameters estimated by an IRT model. Therefore, a primacy effect from this measurement factor needs to be avoided, considered, held constant, or separated from the interested effects by cautious researchers.

REFERENCES

Bass, B. M., Cascio, W. F., and O'connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology*, 59, 313–320.

Becker, S. L. (1954). Why an order effect. *Public Opinion Quarterly*, 18, 271–278.
Belson, W. A. (1966). The effects of reversing the presentation order of verbal rating scales. *Journal of Advertising Research*, 6, 30–37.
Brook, D. and Upton, G. J. G. (1974). Biases in local government elections due to position on the ballot paper. *Applied Statistics*, 23, 414–419.
Carp, F. M. (1974). Position effects on interview responses. *Journal of Gerontology*, 29, 581–587.
Chan, J. C. (1986). Sex, gender roles and empathy. Unpublished thesis. National Chengchi University, Taiwan, R. O. C.
Chase, C. I. (1969). Often is where you find it. *American Psychologist*, 24, 1043.
Davis, M. H. (1980). A multidimensional approach to individual differences in empathy. *Catalog of selected documents in psychology*, 10, 85.
Hakel, M. D. (1968). How often is often? *American Psychologist*, 23, 533–534.
Johnson, J. D. (1981). Effects of the order of presentation of evaluative dimensions for bipolar scales in four societies. *The Journal of Social Psychology*, 113, 21–27.
Jöreskog, K. G. and Sörbom, D. (1984). *LISREL VI: User's guide* (3rd ed.). Mooresville, IN: Scientific Software.
Klockars, A. J. and Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement*, 25, 85–96.
Krosnick, J. A. and Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly*, 51, 201–219.
Krosnick, J. A. and Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly*, 52, 526–538.
Mathews, C. O. (1929). The effect of the printed response words on an interest questionnaire. *Journal of Educational Psychology*, 30, 128–134.
McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly*, 67, 205–211.
Mueller, J. E. (1970). Choosing among 133 candidates. *Public Opinion Quarterly*, 34, 395–402.
Payne, J. D. (1972). The effects of reversing the order of verbal rating scales in a postal survey. *Journal of Market Research Society*, 14, 30–44.
Powers, E. A., Morrow, P., Goudy, W. J., and Keith, P. M. (1977). Serial order preference in survey research. *Public Opinion Quarterly*, 41, 80–85.
Rugg, D. and Cantril, H. (1944). The wording of questions. In H. Cantril (Ed.), *Gauging public opinion*, Chapter 2. Princeton, NJ: Princeton University Press.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, Monograph Supplement No. 17.

Schuman, H. and Presser, S. (1981). *Questions and answers in attitude surveys*. New York: Academic Press.
Simon, H. A. (1957). *Models of man*. New York: Wiley.
Stouffer, S. A. and DeVinney, L. C. (1949). How personal adjustment varied in the army—By background characteristics of the soldiers. In S. A. Stouffer, E. A. Schuman, L. C. Devinney, S. A. Star, and R. M. Williams (Eds.), *The American soldiers: Adjustment during army life*. Princeton, NJ: Princeton University Press.
Thissen, D. (1986). *Multilog: A user's guide*. Mooresville, IN: Scientific Software, Inc.
Wildt, A. R. and Mazis, M. B. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research*, 15, 261–267.
Worcester, R. M. and Burns, T. R. (1975). A statistical examination of the relative precision of verbal scales. *Journal of the Market Research Society*, 17, 181–196.
Wright, B. D. (1967). Sample-free test calibration and person measurement. Proceedings of the 1967 Invitational conference on Testing Problems. Princeton, NJ: Educational Testing Service, 1986.