EFFECT OF ITEM REVERSAL AND ANCHOR CHOICE IN LIKERT SCALES

Alan D. Moore and Suzanne Young
University of Wyoming

It is usually recommended that we reverse the direction in about half the items in Likert scales and that we anchor the rating responses with the terms, "strongly disagree" through "strongly agree." To test the effects of item reversal and anchor word choice we conducted a 2x2 factorial experiment. Subjects were 84 students enrolled in a sophomore-level computer course for prospective teachers. A significant interaction of the two effects was found for a scale measuring attitudes toward computers. When item response anchors were asymmetric, the ratings on reversed item stems were higher than those with direct stems. In contrast, when item response anchors were symmetric, average ratings for reversed and direct item stems were not different. When constructing Likert scales it is recommended that a) direct item stems should be written; b) the response option continuum should be symmetric from strongly agree to strongly disagree, with the same number of positive and negative response options; and c) the direction of these response options should be randomly reversed, with about half the items SA to SD and half SD to SA. Scales formed following these procedures should be more easily understood by respondents, should have higher reliability and factorial simplicity, and should provide some protection against response set and acquiescent response behaviors.

One of the most common methods of constructing scales for the measurement of attitudes, sentiments, and opinions is the Likert scale (Likert, 1932). In this method a set of statements about an idea, issue, or object is written. Respondents are asked to assign a rating to each statement reflecting their degree of agreement with the statement. Typically, a 5 is assigned to "strongly agree" and a 1 is assigned to "strongly disagree." The score on the scale is simply the sum or average of the ratings across all the statements.

It is usually recommended that the direction of item stems be reversed in about half the items in a Likert scale. (Allen and Yen, 1979; Crocker and Algina, 1986; Likert, 1932; Nunnally, 1978; Ross, Wright and Anderson, 1983; Worthen, Whiute, Fan and Sudweeks, 1999). While this is intended to avoid a response set, stereotyped response among respondents, or acquiescent response set, the presentation of reversed items in a scale presents other problems. Often, when attempting to reverse the sense of an item, negative syntax is introduced which may be confusing for respondents. For example, it may be difficult to reason through the process of disagreeing with a statement like, "Not all children can learn." Also, in some groups, there may sensitivity to the reversed sense of questions. There may be a perception that these items are a criticism. The mere presentation of a statement that is counter to some desired goal or socially acceptable view may be perceived by respondents and gatekeepers as politically motivated or threatening.

It is also common practice to anchor the rating responses with the terms, "strongly disagree, disagree, neutral, agree, strongly agree." In Likert's (1932) original article describing the method, he said, "[s]o far as the measurement of the attitude is concerned, it is quite immaterial what the extremes of the attitude continuum are called; the important fact is that persons do differ quantitatively in their attitudes, some being more toward one extreme, some more toward the other" (p. 48). But researchers take great liberty in choosing these anchor terms, depending on the context of the question and the population being questioned.

Several researchers have studied the effect of item reversal. Pilotte and Gable (1990) studied the factor structure of computer anxiety scale data gathered from high school students. Using confirmatory factor analysis, they concluded that different latent traits of anxiety were measured by sets of items containing either positive or negative item stems.

Barnette (1997) studied the effects of item stem reversal and item response reversal in a 2x3 randomized experiment. A scale measuring attitudes toward year-round schooling was completed by 687 undergraduate students, graduate students, and inservice teachers. Item means were lower for direct-worded surveys compared to those with half direct and half-reversed item stems. Also, lower reliability estimates were associated with half direct, half reversed item stems, while higher reliabilities were found with scales containing direct item stems. He recommended that direct wording be used in item stems, but that half the response anchors should reversed. This would result in scales with higher reliability but would still guard against acquiescent and response set behaviors.

Lam and Klockers (1982) obtained data from 375 freshmen students at the University of Washington. Four randomized groups were asked to rate the quality of their educational experience. In a factorial design, one factor was type of anchor points. The questionnaires of half the students had only two bipolar response anchors labeled while those of the other half had each scaling point labeled. The second factor was symmetry of the item response choices -- offering more scaling points at one end of the scale. They found that the type of anchor did not result in different mean ratings, but that asymmetric response choices changed the average responses. Evidently raters do respond to the semantic meaning of the anchors.

Since there are conflicting results on the effect of item reversal and the effect of anchor asymmetry, further study is needed. The purpose of our study was to investigate the effect of item stem reversal in Likert scales and the effect of choice of anchor terms. Our questions were a) what difference in ratings results when items are either reversed or direct when scaling attitudes using a Likert scale; and b) what is the difference in ratings when response anchor terms "don't agree" through "very much agree" are replaced by "strongly disagree" to "strongly agree? It is important to learn what the actual effect of reversal is. If the effect is large, we must construct scales that include these reversed items in spite of objections based upon inferred threat or respondent confusion. However, if this effect is small or zero, we may be able to relax the requirement that about half the items be reversed. It is also important to have some empirical estimate of the potential effect of a particular choice of anchor terms. If there is little difference in ratings depending on a particular choice of anchor items then the choice may rightly be a matter of personal taste.

METHOD

In order to test these effects, a 2x2 factorial experiment was conducted. During March and April, 1999, a sample of undergraduate teacher education students was asked to participate in the study. These students were enrolled in a sophomore-level computer course for prospective teachers, Teaching with Microcomputers, at the University of Wyoming, during Spring semester, 1999.

A scale measuring attitudes toward computers, the Computer Attitude Survey for Prospective Teachers (Young, 1994) was modified for use in the study. This instrument was revised and pilot tested on a sample similar to the present sample in Spring semester, 1998. The scale consists of two sections. The first section, the Likert scale, contains 30 statements relating to the use of computers in instruction in which respondents are asked to circle a number from 1 to 5, according to whether they don't agree, slightly agree, tend to agree, generally agree, or very much agree with a statement. On this scale, the anchor terms are not symmetric, that is, there are more anchor words on the "agree" end of the scale than on the "disagree." About half the items (13 of 30) are reversed, consistent with standard Likert scale design methods. The second section contains 9 items that gather demographic information and information about computer use, comforts, and interest. An internal consistency reliability estimate of 0.94 was obtained using Cronbach's alpha.

For the present study three additional forms of the instrument were constructed to meet the design of the 2x2 factorial experiment. Form C was the unmodified instrument. Form B contained the same items but the reversed items were changed back to direct statements. Forms A and D con-

tained the same items as B and C, respectively, but the response anchors "don't agree, slightly agree, tend to agree, generally agree, and very much agree" were changed to "strongly disagree, disagree, neutral, agree, strongly agree."

Each student in the three sections of Teaching with Microcomputers received one of four forms of the Computer Attitude Survey for Prospective Teachers. The forms were randomly assigned to the students by spiraling the four forms as they were distributed to students in each class. Students were assured that their responses were anonymous and that they had a right to refuse to participate in the study. They were allowed 15 minutes in class to complete the instrument. At the end of the time period the researchers collected the instruments.

The sample consisted of 84 students. Twenty-nine were in the Monday section, 34 in the Tuesday, and 21 in the Thursday section. The median age in this highly skewed distribution was 20 years old, with 14% 18 years old, 25% 19 years old, 18% 20 years old, 13% 21-22 years old, and 30% 23 years and older. Ages ranged from 18 to 43. The majority of the students were female (70%). About half (52%) were elementary education majors, 38% were secondary education majors, 6% were junior high or middle school education majors, and 4% reported "Other."

RESULTS

The Likert scales were scored by reversing the appropriate item responses on forms C and D, then adding the item responses for the first 30 items. The means, standard deviations, and sizes of the four experimental groups are displayed in Table 1. It shows that the mean attitude scores for the

Table 1
Means and Standard Deviations of Experimental Groups

	Reversed			Not Reversed			Total		
	M	SD	n	M	SD	n	M	SD	n
Symmetric	110.95	12.05	21	107.59	10.14	22	109.23	11.11	43
Asymmetric	110.19	15.64	21	104.25	17.42	20	107.29	16.60	41
Total	110.5	13.79	42	106.00	14.00	42	108.29	14.00	84

Table 2
Analysis of Variance for Attitude, Reversal, and Symmetry

Source	SS	df	MS	F	p
Reversal	453.747	1	453.747	2.310	.133
Symmetry	88.273	1	88.273	.449	.505
Interaction	34.879	1	34.879	.178	.675
Error	15717.259	80	196.466		
Total	16279.143	83			

Table 3
Means and Standard Deviations of Experimental Groups on Reversed Items Only

	Reversed			Not Reversed			Total		
	M	SD	n	M	SD	n	M	SD	n
Symmetric	43.48	6.14	21	42.23	4.80	22	42.84	5.47	43
Asymmetric	48.571	7.70	21	40.56	7.78	20	44.66	8.65	41
Total	46.02	7.35	42	41.43	6.37	42	43.73	7.22	84

four experimental groups were quite similar.

Analysis of Variance (see Table 2) shows no statistically significant relationship between attitude scores and either experimental variable (R^2 = .035, $F(3,80) = 0.954$, $p = .419$). The main effect reflecting the difference in attitude scores according to whether items were reversed or direct was not statistically significant, ($F(1, 80) = 2.310$, $p=.133$). Likewise, the main effect of anchor type was not statistically significant ($F(1,80) = 0.449$, $p=.505$). Furthermore, the interaction of reversal and anchor choice was also not statistically significant ($F(1,80) = .178$, $p=.675$). The standardized mean differences (effect sizes) were .326 for Reversal, .138 for Anchor Type, and .184 for the interaction. These were calculated by dividing the difference in means by the root mean square error from the factorial ANOVA.

Since the standardized mean difference for item reversal was the largest, and near the upper bound of Cohen's (1977) designation of a small effect size, we took a closer look at the responses to reversed items. Because we suspected that the effect of item reversal was masked by the 17 out of the 30 items that were direct in all forms, we studied a subscale consisting of only those items that were reversed in the original scale. We reasoned that if item reversal had no effect, the distributions of responses to these items would coincide, after the item scores were reversed, with those when the items were direct. Table 3 shows the means and standard deviations for each combination of symmetry and reversal.

The results of a 2x2 ANOVA of subscale scores with the factors item reversal and symmetry are displayed in Table 4. The main effect for symmetry was not significant. But, both the interaction of symmetry and reversal and the reversal main ef-

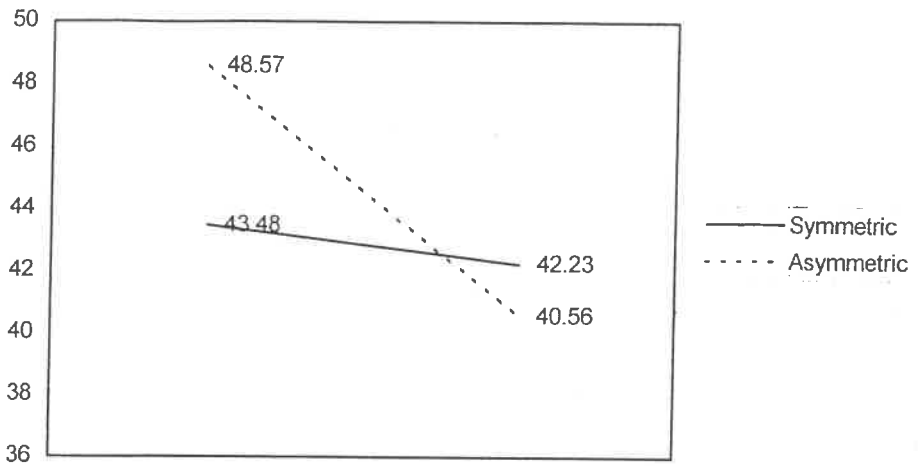
Table 4
Analysis of Variance for Attitude, Reversal, and Symmetry— Reversed Items Only

Source	SS	df	MS	F	p
Reversal	450.669	1	450.669	10.084	.002
Symmetry	61.263	1	61.263	1.371	.245
Interaction	240.528	1	240.528	5.382	.023
Error	3575.195	80	44.690		
Total	4320.702	83			

fect were significant (F=5.382, p=.023; and F=10.084, p=.002, respectively).

The significant interaction means that the effect of item reversal depends on whether the item anchors were symmetric or not. A graph of the nature of this interaction is displayed in Figure 1. Tests of simple effects show that there was no statistically significant difference between reversed (43.48) and direct (42.23) average subscales scores when the item anchors were symmetric (F = 0.375, p=0.542). But the mean subscale score was significantly higher for reversed items (48.57) compared to unreversed items (40.56) when asymmetric item anchors were used (F = 14.749, p <.001).

Figure 1. Interaction of Reversal and Symmetry



	Reversed	Direct
Symmetric	43.48	42.23
Asymmetric	48.57	40.56

Though the sample sizes were too small to allow Cronbach alpha internal consistency estimates of reliability for each of the four forms, a rough estimate of the reliability was obtained by calculating split half reliabilities using the Split-Half Reliability procedure in SPSS for Windows 9.0. After correcting for the full-scale length using the Spearman-Brown correction, these estimates were 0.91, 0.94, 0.80, and 0.92 for forms A, B, C, and D, respectively. Form C, which has a substantially lower estimate than the other three forms was the one in which about half the item stems were reversed and the item anchors were asymmetric.

DISCUSSION

The standard methods for the construction of Likert scales are called into question by this study. Though there was little effect of reversal or symmetry of anchors on the total attitude scale scores, the true effect of reversal was masked because 17 of the 30 item stems were the same in all forms. In looking at only those 13 reversed items, the most interesting finding was the interaction of response reversal and anchor symmetry. According to our results, the effect of item reversal is small when response anchors are symmetric, that is, when items are half positive and half negative. But when these anchors were not symmetric, the effect of item reversal was substantial. The standardized mean difference for this simple effect was 1.20 standard deviations. When a scale with asymmetric anchors contains some reversed items and some direct items, this interaction can be expected to lower the reliability of the scale, consistent with the findings of Pilotte and Gable (1990) and Barnette (1997).

Our results suggest that when response anchors are symmetric, the effect of item stem reversal is negligible. This conclusion supports the recommendation of Barnette (1997) that direct item stems should be used rather than reversing half of them, as traditionally recommended. Since the introduction of reversed items may increase the factorial complexity of the scale, scales with only direct item stems would be expected to be more reliable. Furthermore, if item response anchors are not symmetric, the effect of item reversal can have substantial effects on the item means, factor struc-

ture, and scale reliability. Barnette's suggestion that item stems should be direct, but about half the responses anchors should be reversed, in order to overcome response set behaviors, is attractive and should be studied more.

In conclusion, the best recommendation we can currently offer for the construction of Likert scales, based on our literature review and this study, are a) direct item stems should be written; b) the response option continuum should be symmetric from strongly agree to strongly disagree, with the same number of positive and negative response options; and c) the direction of these response options should be randomly reversed with about half the items SA to SD and half SD to SA. Scales formed following these procedures should be more easily understood by respondents, have higher reliability and factorial simplicity, and should provide some protection against response set and acquiescent response behaviors. We should not construct scales using reversed items when item response anchors are asymmetric.

REFERENCES

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory. Belmont, CA: Wadsworth, Inc.

Barnette, J. J. (1997). Effects of item and response set reversal on survey statistics. Paper presented at the annual meeting of the American Educational Research Association, March 24-28, 1997, Chicago, IL.

Cohen, J. (1977). Statistical Power Analysis for the Behavioral Sciences (Rev. Ed). New York: Academic Press.

Crocker, L., & Algina, J. (1986). Introduction to Classical and Modern Test Theory. New York: Holt, Rinehart and Winston, Inc.

Lam, T. C. & Klockers, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. Journal of Educational Measurement, 19 (4), 317-322.

Likert, R. (1932). A technique for the measurement of attitudes. Archives of Psychology, 140, 5-55.

Pilotte, W. J. & Gable, R. K. (1990). The impact of positive and negative item stems on the validity of a computer anxiety scale. Educational and Psychological Measurement, 50, 603-610.

Ross, P. H., Wright, J. D. & Anderson, A. B. (1983). Handbook of Survey Research. New York: Academic Press.

Nunnally, J. C. (1978). Psychometric Theory (2nd Ed.). New York: McGraw-Hill. Worthen, B. R, Whiute, K. R., Fan, X., & Sudweeks, R. R. (1999). Measurement and Assessment in the Schools (2nd Ed.). New York: Longman.

Young, S. (1994). Computer Attitude Survey for Prospective Teachers. Unpublished Manuscript, University of Northern Colorado at Greely.