DEPENDABILITY OF ANCHORING LABELS OF LIKERT-TYPE SCALES

LEI CHANG
Chinese University of Hong Kong

Generalizability theory was used to examine the dependability of anchoring labels of Likert-type scales. Variance components associated with labeling were estimated in two samples using a two-facet, random effects g-study design. In one sample, 173 graduate students in education were administered seven items measuring attitudes toward quantitative methodology. The other sample consisted of 108 graduate students in education who responded to the eight-item Life Orientation Test. From both samples, variance components associated with labeling were found to be minute, contributing little to the observed score variance. One plausible explanation was that respondents could primarily be using the numerical information in rating a Likert-type scale.

Researchers sometimes use different verbal labels to anchor the scale points associated with different items of a psychological test. For example, in the 28-item Achievement Anxiety Test (AAT, Alpert & Haber, 1960), six sets of descriptors were used to label the five-point scale associated with different items. Researchers may also subsequently change the anchoring labels used originally on an instrument. This kind of instrument modification is often not disclosed in a study (Huck & Jacko, 1974). According to Huck and Jacko (1974), the AAT (Alpert & Haber, 1960) was used in three investigations (Walsh, 1968, 1969; Walsh, Engbretson, & O'Brien, 1968) in which the six different sets of anchoring labels were changed into one constant set. This change was not reported in these studies.

As another example, anchoring labels of the Self-Consciousness Scale (Fenigstein, Scheier, & Buss, 1975; Scheier & Carver, 1985) have been changed from *extremely characteristic of me—extremely uncharacteristic of*

Please send inquiries regarding this article to Lei Chang, Department of Educational Psychology, Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

Educational and Psychological Measurement, Vol. 57 No. 5, October 1997 800-807
© 1997 Sage Publications, Inc.

from the SAGE Social Science Collections. All Rights Reserved.

me to *a lot like me—a lot unlike me*. In these situations, are the different labels exchangeable, or do they add, erroneously, to the observed variance of the measurement? The present study was intended to answer this question by examining the dependability of the anchoring labels associated with Likert-type scale points.

Most of the existing studies compared scales that were fully labeled, labeled at two ends, and not labeled. The findings have been mixed. Using a 17-item faculty evaluation questionnaire, Newstead and Arnold (1989) compared a five-point scale anchored by these three forms. Each form was given to an independent sample of approximately 50 undergraduate students. The unlabeled scale was found to produce the highest means, whereas the scale having verbal labels at two ends had the lowest means. There was no difference in the variances produced by the three labeling formats. As a proxy for reliability comparison, they had the subjects use the three scales to rate six items where there was an objectively correct answer. The unlabeled scale showed the highest degree of accuracy, whereas the fully labeled scale was least accurate.

In contrast, Huck and Jacko (1974) reported that fully labeled scales resulted in higher means than did scales having labels at two ends. Other studies, however, observed no difference in means among the three scale formats (Dixon, Bobo, & Stevick, 1984; Finn, 1972; Wyatt & Meyers, 1987), although some of these researchers reported differences in variance (Dixon et al., 1984; Wyatt & Meyers, 1987). To add to the confusion, Frisbie and Brandenburg (1979) found no difference in one set of items between fully labeled and end-labeled scales and, for another set of items, higher means for the end-labeled scale.

As part of an investigation of numbers of scale options, McKelvie (1978) also compared labeled versus unlabeled scales of five and seven scale points and concluded that "neither reliability nor validity are influenced by the presence of verbal anchors" (p. 198). Similarly, Boote (1981) found no difference in test-retest reliability between labeled and unlabeled scales. However, Peters and McCormick (1966) reported that scores on job- or task-anchored scales had higher reliability than scores on scales that did not have labels.

Anchoring labels also have been studied in terms of scaling or assigning scale values to the anchoring labels. Researchers suspect that the central tendency of the distribution may shift due to the use of different anchoring labels having connotative valencies that are perceived to be different. Initial studies were conducted to determine a set of verbal labels that represented an equal interval distance (Bass, 1968; Bass, Cascio, & O'Connor, 1974; Cliff, 1959; Spector, 1976). For example, Bass (1968) had 71 undergraduate students rate the distances among 28 adverbs of frequency and found that *always*, *very often*, *fairly often*, *sometimes*, *seldom*, and *never* approximated an equidistant relation to each other. In another study, the following valuative

phrases were found to be evenly spaced and symmetric about the midpoint: *very poor*, *need improvement*, *satisfactory*, *good*, and *very good* (Lam & Klockars, 1982). Researchers have subsequently tried to manipulate the choices of the anchoring labels and their locations on the numerical scale (French-Lazovik & Gibson, 1984; Klockars & Yamagishi, 1988; Lam & Klockars, 1982) to see if such manipulation affects the mean and variance of the resulting distribution. The effects in using different anchoring labels on distributional characteristics of the scale have been identified.

These studies were limited to comparing labels associated with odd numbers of scale points. Some researchers suspected that the middle category in an odd-numbered scale makes room for a response set (Bendig, 1954; Cronbach, 1950; Goldberg, 1981; Nunnally, 1967). Even-numbered scales were found to have higher reliability than odd-numbered scales (Bendig, 1954; Masters, 1974) and, thus, were preferred to odd-numbered scales (McKilvie, 1978; Matell & Jacoby, 1972). These observations indicate that an accurate verbal description of numerical distances is more difficult to achieve when labeling the middle point than other points of a scale. For example, as French-Lazovik and Gibson (1984) pointed out, "average," which is often used to anchor the midpoint of a scale, may be viewed as more pejorative than neutral. Using a word that is perceived to be below the midpoint to anchor the midpoint forces the distribution of responses to shift to the higher end of the scale. The mean of such responses will be higher than it would otherwise be. This particular difficulty in labeling the midpoint may have led to the research findings of mean differences between labeled and unlabeled scales as well as among the differently labeled scales. Given this speculation and the fact that existing studies have not examined labeling of even-numbered scales, the present study compared different verbal labels anchoring a four-point and a six-point scale.

Method

Application of Generalizability Theory

The present study used generalizability theory (Brennan, 1983; Cronanbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991) to evaluate the dependability of scale labels. The design of a generalizability study represents a researcher's conceptualization of a possible measurement situation, that is, the object of measurement and the conditions under which observations are made. In the present investigation of anchoring labels, the measurement situation involves persons (object of measurement) responding to any items (the item facet) the scale of which is labeled by verbal descriptors (the label facet). This results in a two-facet universe of admissible observations, items and labels, that are associated with a population of persons (which are the

objects of measurement). This measurement conceptualization is represented by a crossed random-effects design, persons by items by labels, or $p \times i \times l$. Under this design, a particular measurement observation, X_{pil}, denotes the observed score of any person in the population on any item-label combination in the universe of admissible observations. Under the present design, there are seven variance components making up the observed score variance:

$$\sigma^2(X_{pli}) = \sigma^2(p) + \sigma^2(i) + \sigma^2(l) + \sigma^2(pi) + \sigma^2(pl) + \sigma^2(il) + \sigma^2(pil).$$

These variances are the expected squared deviations of the means associated with a condition from the grand mean. For example, $\sigma^2(l) = E(\mu_l - \mu)^2$. These variance components are estimated by equating them to their observed mean squares in ANOVA, for example, $\hat{\sigma}^2(l) = [MS(l) - MS(pl) - MS(il) + MS(pil)] / n_p n_i$.

The variance estimates from this g-study design were associated with a single condition and a single object of measurement. These estimates provided information for making decisions on more efficient measurement procedures. Such decisions constitute what is referred to as a *d-study*.

In the current investigation, the d-study was concerned with determining error variance associated with the labeling facet but was not concerned with determining variance due to the item facet. Because researchers rarely apply more than one labeling scheme to the measurement of an item, variance estimation in the d-study was associated with using a single kind of anchoring label but not with a mean score from a sample of labeling conditions. Thus, the d-study estimates of variance components in relation to labeling were identical to those obtained in the g-study.

The dependability of the labeling condition was evaluated by computing a generalizability coefficient ($E\hat{\rho}^2$) with item being fixed. The question being addressed by such a coefficient is how well a person's relative standing on the same item obtained by one labeling scheme can be generalized to other labeling schemes.

The dependability coefficient for domain-referenced interpretation of measurement ($\hat{\varphi}$, Brennan & Kane, 1977) was also evaluated to determine labeling consistency in transmitting persons' domain status or absolute level on an item.

In the current design where there was one entry at each cell, the residual term represented a three-way interaction that was confounded by other unexplained sources of variation. The conventional application of G-theory leaves the residual term unexplained.

Subjects, Measures, and Procedures

The variance components associated with the labeling condition were estimated in two samples using the previously described $p \times i \times l$ design. In

one sample, 173 graduate students in education from an U.S. university were administered seven items on a four-point scale that measured attitudes toward quantitative methodology. The four-point scale of the items was anchored by two kinds of labels. One labeling had 1 = *disagree*, 2 = *somewhat disagree*, 3 = *somewhat agree*, 4 = *agree*. In the other labeling, 1 = *strongly disagree*, 2 = *disagree*, 3 = *agree*, 4 = *strongly agree*. Subjects responded to the items twice using the two kinds of labels. The order of administrations of the two labels was mixed among students. The two administrations were 1 week apart. Thus, this was a 173 by 7 by 2 design.

The other sample consisted of 108 graduate students in education from a U.S. university who responded to the eight-item Life Orientation Test (Scheier & Carver, 1985). A six-point scale was used that was either fully labeled (1 = *completely disagree*, 2 = *very much disagree*, 3 = *disagree*, 4 = *agree*, 5 = *very much agree*, 6 = *completely agree*) or labeled only at two ends (1 = *completely disagree*, 6 = *completely agree*). Subjects responded to the items using both labeling formats. The two administrations were 1 week apart. Order of administrations of the two labels was mixed. This was a 108 by 8 by 2 design.

Results

Table 1 presents, for both measures, the variance components from the two-facet random-effects design, $p \times i \times l$. For both measures, $\hat{\sigma}^2(pl)$, which indicates interference of labeling on the relative standings of persons averaging over items, was moderate. It accounted for 6% and 4% of the observed variance for the two measurements, respectively, as reported in Table 1. This result shows that labeling as a necessary condition for obtaining attitude measurement does not introduce much error in a normative interpretation of the observations. The main effect of labeling, $\hat{\sigma}^2(l)$, was almost zero for both measures (negative variance was treated as zero). Averaging over persons and items, there was almost no variance among μ_l. In other words, the same means or total scores will be obtained using different labels. This result demonstrates the dependability of labeling for an absolute (domain-referenced) interpretation of observations. Finally, $\hat{\sigma}^2(il)$ represented less than 1% of the observed variance for both measurements. There was little inconsistency among combinations of a label and an item. Thus, item calibration was consistent for different labels.

For the first study, $E\hat{\rho}^2$ was .4917 and $\hat{\varphi}$ was .4859. For the second study, $E\hat{\rho}^2$ was .6233 and $\hat{\varphi}$ was .6175. The moderate coefficients were the direct result of the large $\hat{\sigma}^2(pil)$, which will be greatly reduced if more items are sampled.

Table 1
Variance Components From p × i × l *Random Effects Design*

Source	*SS*	*df*	*MS*	$\hat{\sigma}^2$	$\hat{\sigma}^2\%$
Four-point scale, two kinds of labels					
Person (p)	409.72007	172	2.3821	.10417	16.2
Item (i)	71.88687	6	11.9811	.03021	4.7
Label (l)	4.21181	1	4.2118	.00231	0.4
pi	674.68456	1032	0.6538	.19275	30.0
pl	92.57391	172	0.5382	.03857	6.0
il	6.86623	6	1.1444	.00506	0.8
pil	276.84806	1032	0.2683	.26826	41.8
Six-point scale, fully-labeled versus end-labeled					
Person (p)	1059.68229	107	9.90357	.53824	41.7
Item (i)	68.69850	7	9.81407	.03539	2.7
Label (l)	0.48669	1	0.48669	-.00187	0.0
pi	664.98900	749	0.88784	.23545	18.3
pl	87.82581	107	0.82080	.05048	3.9
il	11.89757	7	1.69965	.01188	0.9
pil	312.28993	749	0.41694	.41694	32.3

Discussion

It is concluded that attitude measurement obtained from a Likert-type scale can be generalized across different anchoring labels. The dependability of anchoring labels is maintained both for relative and absolute interpretation of individual differences with respect to what is being measured. One potential practical implication is that researchers need not be overly concerned with the practice of using different labels to anchor Likert-type scales for items of the same or different instruments. As long as the numerical scale is clearly defined and consistent across items and tests, the labeling difference does not seem to contribute to the observed variance.

Furthermore, perhaps researchers can free themselves from the concern and effort in choosing verbal labels that are equal distant in connotative intensities. In the present study, the labels used to anchor the four-point scale represented unequal distances: In one set, 1 = *disagree* and 4 = *agree* and the distance between the two labels was three; in the other set, 2 = *disagree* and 3 = *agree* and the distance between the same two labels was one.

One weakness of the present study is the possible memory effect of the subjects in responding to the same questionnaire twice, which cannot be determined or assessed given the way the study was designed and imple-

mented. However, a counter-balanced delay was employed to minimize this effect. There is a need for further research that employs a nested design where respondents are randomly assigned to different labeling schemes to cross validate the findings of this study.

Another weakness of the present study lies in the employment of small numbers of items. External validity of the findings can be improved in future studies using larger samples of items and respondents. In addition, further research should focus on the cognitive process of responding to a Likert-type scale. Despite the earlier psychophysical research on the connotative intensity of different adverbs and adjectives (e.g., Cliff, 1959), the exact meanings subjects assign to the response options when responding to a rating scale remain mostly unknown (Klockars & Yamagishi, 1988). At least one plausible explanation can be presented.

In the present study, the numerals associated with the two measurement scales were constant, whereas the labeling of the numerical points was manipulated. The finding that labeling did not add to the observed variance could suggest that subjects respond mostly to the numerical but not labeling information when rating the psychological valency of an item. If there is a discrepancy between the equal-distance relation intended by the scale and what is inadequately represented by the labels, such a discrepancy seems to be easily compensated for by the numerals underlying the scale. Because the numerals (i.e., 1, 2, 3 . . .) represent equidistant relations, subjects' responses to the numerical information reduce the relevancy of the differential connotative strength of the verbal labels.

References

Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. *Journal of Abnormal and Social Psychology, 61*, 207-215.

Bass, B. M. (1968). How to succeed in business according to business students and managers. *Journal of Applied Psychology, 52*(3), 254-262.

Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology, 59*(3), 313-320.

Bendig, A. W. (1954). Reliability of short rating scales and the heterogeneity of the rated stimuli. *Journal of Applied Psychology, 38*(3), 167-170.

Boote, A. S. (1981). Reliability testing of psychographic scales: Five-point or seven-point? Anchored or labeled? *Journal of Advertising Research, 21*(5), 53-60.

Brennan, R. L. (1983). *Elements of generalizability theory*. Iowa City, IA: ACT.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14*, 277-289.

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review, 66*, 27-44.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement, 10*, 3-31.

Cronbach, L. J., Gleser, G. C., Nanda, M. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability of scores and profiles*. New York: John Wiley.

Dixon, P. N., Bobo, M., & Stevick, R. A. (1984). Response differences and preferences for all category-defined and end-defined Likert formats. *Educational and Psychological Measurement, 44*, 61-66.

Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting Psychology, 45*(4), 522-527.

Finn, R. H. (1972). Effects of some variations in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement, 32*, 255-265.

French-Lazovik, G., & Gibson, C. L. (1984). Effects of verbally labeled anchor points on the distributional parameters of rating measures. *Applied Psychological Measurement, 8*(1), 49-57.

Frisbie, D. A., & Brandenburg, D. C. (1979). Equivalence of questionnaire items with varying response formats. *Journal of Educational Measurement, 16*, 43-48.

Goldberg, L. R. (1981). Unconfounding situational attributions from uncertain, neutral, and ambiguous ones: A psychometric analysis of descriptions of oneself and various types of others. *Journal of Personality and Social Psychology, 41*(3), 517-552.

Huck, S. W., & Jacko, E. J. (1974). Effect of varying the response format of the Alpert-Haber *Achievement Anxiety Test. Journal of Counseling Psychology. 21*(2), 159-163.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25*(2), 85-96.

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

Masters, J. R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement, 11*(1), 49-53.

Matell, M. S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? Effects of testing time and scale properties. *Journal of Applied Psychology, 56*, 506-509.

McKelvie, S. J. (1978). Graphic rating scales—How many categories? *British Journal of Psychology, 69*, 185-202.

Newstead, S. E., & Arnold, J. (1989). The effect of response format on ratings of teaching. *Educational and Psychological Measurement, 49*, 33-43.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw-Hill.

Peters, D. L., & McCormick, E. J. (1966). Comparative reliability of numerically anchored versus job-task anchored rating scales. *Journal of Applied Psychology, 50*, 92-96.

Scheier, M. F., & Carver, C. S. (1985). The Self-Consciousness Scale: A revised version for use with general populations. *Journal of Applied Social Psychology, 15*(8), 687-699.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Spector, P. E. (1976). Choosing response categories for summated rating scales. *Journal of Applied Psychology, 61*(3), 374-375.

Walsh, R. P. (1968) Some correlates of test-taking anxiety. *Psychological Reports, 22*, 449-450.

Walsh, R. P. (1969). Test-taking anxiety and psychological needs. *Psychological Reports, 25*, 83-86.

Walsh, R. P., Engbretson, R. O., & O'Brien, B. A. (1968). Anxiety and test-taking behavior. *Journal of Counseling Psychology, 15*, 572-575.

Wyatt, R. C., & Meyers, L. S. (1987). Psychometric properties of four 5-point Likert-type response scales. *Educational and Psychological Measurement, 47*, 27-35.