DOCUMENT RESUME

ED 397 109 TM 025 241

AUTHOR Chang, Lei
TITLE Dependability of Anchoring Labels of Likert-Type Scales.
PUB DATE Apr 96
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Generalizability Theory; *Graduate Students; Higher Education; Individual Differences; *Likert Scales; Research Design; Responses; *Scaling Scores; *Test Construction
IDENTIFIERS *Anchoring Devices

ABSTRACT

This study uses generalizability theory 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 effect generalizability-study design. In one sample, 173 graduate students in education were administered 7 items measuring attitudes toward quantitative methodology. The other sample consisted of 108 graduate students in education who responded to the 8-item Life Orientation Test (M. F. Scheier and C. S. Carver, 1985). From both samples, variance components associated with labeling were found to be trivial, contributing little to the observed score variance. The dependability of anchoring labels was maintained for both normative and absolute interpretations of individual differences with respect to what was being measured. Two plausible explanations were provided. Respondents could primarily be using the numerical information in rating a Likert-type scale or could treat both the scale numerals and verbal labels as representing ordinal rather than equidistant relations. (Contains 2 tables and 29 references.) (Author/SLD)

Dependability of Anchoring Labels of Likert-type Scales

Lei Chang

College of Education

University of Central Florida

Orlando, FL 32816-1250

Inquiries regarding this manuscript should be addressed to Lei Chang, Department of Educational Foundations, University of Central Florida, Orlando, FL 32816-1250.

Paper presented at the annual meeting of the American Education Research Association, New York City, April, 1996

Abstract

This study uses generalizability theory 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 effect 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 8-item Life Orientation Test (Scheier & Carver, 1985). From both samples, variance components associated with labeling were found to be trivial, contributing little to the observed score variance. Two plausible explanations were provided. Respondents could primarily be using the numerical information in rating a Likert-type scale or could treat both the scale numerals and verbal labels as representing ordinal rather than equidistant relations.

Dependability of Anchoring Labels of Likert-type Scales

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 change the anchoring labels adopted by 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) where 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 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 are 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 while 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 (Finn, 1972; Dixon, Bobo, & Stevick, 1984; 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" (pp. 198). Similarly, Boote (1981) found no difference in test-retest reliability between labeled and unlabeled scales. However, Peters and McCormick (1966) reported that job-task

anchored scales had higher reliability than scales that did not have labels.

Anchoring labels have also 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 the connotative valency of which are perceived to be different. The initial purpose was to determine a set of verbal labels the placement of which represents an equal interval distance (Cliff, 1959; Bass, 1968; Bass, Cascio, & O'Connor, 1974; Spector, 1976). For example, Bass (1968), having 71 undergraduate students rate the distances among 28 adverbs of frequency, 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 (Lam & Klockars, 1982; French-Lazovik & Gibson, 1984; Klockars & Yamagishi, 1988) to see if such manipulation affects the mean and variance of the resulting distribution. The findings indicate the influence of the use of different anchoring labels on distributional characteristics of the scale.

These studies are limited to comparing labels associated

with odd numbers of scale points. Some researchers suspect 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 over odd numbered scales (Matell & Jacoby, 1972; McKilvie, 1978). These observations indicate the possibility that an accurate verbal description of numerical distances is especially 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 mid-point of a scale, may be viewed as more pejorative than neutral. Using a word that is perceived below the mid-point to anchor it forces the distribution of responses to shift to the higher end of the scale. A mean of such responses will be higher than it would otherwise be. This particular difficulty in labeling the mid-point 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 compares different verbal labels anchoring a 4-point and a 6-point scale.

Method

Application of Generalizability Theory

This study uses generalizability theory to evaluate the dependability of scale labels. The design of a generalizability study represents a researcher's conceptualization of a possible measurement situation, i.e., what is the object of measurement and what are 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, which are associated with a population of persons (which are the objects of measurement). This measurement conceptualization is represented by a crossed random effect design, Persons by Items by Labels, or p x i x 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. Expectations can be taken over the conditions of a facet or the population of persons. For example, $\mu_{1} \equiv E_p E_i X_{pil}$. There is a variance associated with each set of expected values. 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 are associated with a single condition and a single object of measurement. These estimates provide information for making decisions on more efficient measurement procedures. Such decisions constitute what is referred to as a d-study or d-study considerations (Brennan, 1983) if the same data set is used for both the d- and g-study. Such is the approach of the present study.

A d-study design is associated with a measurement procedure and a universe of generalization. A measurement procedure can be seen as an application of a measurement conceptualization achieved in a g-study, or, more specifically, as a data collection design to guide the sampling of conditions from the universe of admissible observations. A particular measurement procedure usually involves sampling multiple conditions from each facet for each object of measurement. A universe of generalization consists of the set of all such combinations of sets of multiple conditions. Because almost never do researchers apply more than one labeling scheme to the measurement of an item, the purpose of this study is to determine error variance associated with using a single kind of anchoring label but not with a mean score from a sample of labeling conditions. The

study is not concerned with determining and reducing error variance associated with the item facet. For simplicity, one condition is assumed to be sampled from the item facet as well. Thus, in the present study, the universe of generalization is exactly the same as the universe of admissible observations. That is not only does the latter exhausts all the conditions of the former because the d- and g-study designs are of the same structure, but both universes are associated with single conditions and single objects of measurement. For all practical purposes, the d- and g-studies are indistinguishable in the current investigation. However, to comply with generalizability analysis conventions, uppercase letters connoting mean scores are still used in the current investigation to denote d-study designs or the fact that one condition is sampled for both the item and label facets.

The dependability of the labeling condition is evaluated by computing a generalizability coefficient 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.

$$E\hat{\rho}^2 = \hat{\sigma}^2(p) + \hat{\sigma}^2(pI) \quad / \quad \hat{\sigma}^2(p) + \hat{\sigma}^2(pI) + [\hat{\sigma}^2(pL) + \hat{\sigma}^2(pIL)]$$

Dependability coefficient (Brennan & Kane, 1977) for domain-referenced interpretation of measurement is also evaluated to determine labeling consistency in transmitting persons' domain status or absolute level on a item.

$$\hat{\Phi} = \hat{\sigma}^2(p) + \hat{\sigma}^2(pI) \quad / \quad \hat{\sigma}^2(p) + \hat{\sigma}^2(pI) + [\hat{\sigma}^2(L) + \hat{\sigma}^2(pL) + \hat{\sigma}^2(pIL)]$$

These coefficients are not the focus of this study because they do not represent realistic situations where any test or measurement almost always involves more than one item. Instead, this study focus on an examination of the error variances associated with the labeling facet.

Errors Associated with Labeling

In the present investigation of the dependability across labeling conditions, one source of inconsistency or error variance is $\sigma^2(pL)$. It indicates that the relative standings of persons, averaging over items, are different when different labels are applied. That is, the deviation of a person's item mean from the mean over all persons, $\mu_p - \mu$, is different for different labels. For example, person 1, averaging over items, will score higher than person 2 when one kind of labels is used to anchor the scale options but, when another set of labels is used, obtains the same or lower mean score in relation to person 2. Thus, observed individual differences (with respect to an attribute that is being measured) are unstable due to different labeling conditions of the measurement.

Another source of inconsistency is $\sigma^2(L)$ which represents variability among μ_L when averaging over persons and items. In other words, scores for all persons on all items will be higher for some labels than others. Thus, even though the relative standing of persons is comparable across labels, e.g., person 1 scores higher than person 2 on all labels, the absolute scores of persons are not comparable; all persons score higher on one label

than another. Sample estimates of total or mean scores will not be comparable when different labels are used to anchor response options.

Item-label interaction, $\sigma^2(IL)$, indicates the extent to which item mean scores (over persons) would be rank ordered differently depending on the labeling scheme used. In this case, not all items are consistently registered higher or lower by one kind of label versus another kind; such variability is $\sigma^2(L)$. $\sigma^2(IL)$ indicates that some items will be registered higher by the same label than other items. $\sigma^2(IL)$ is assumed to be independent from $\sigma^2(p)$ which represents individual differences averaging over items and labels. The assumption implies that the item-label inconsistencies affect individual differences in a consistent manner. If the label-item combination effect ($\sigma^2(IL)$) influences, in a different manner, different persons' perceptions of a trait being measured, the resulting variability in the responses is contained in the residual term, $\sigma^2(pIL)$. In the current design where there is one entry at each cell, the residual term represents a three-way interaction which is confounded by other unexplained sources of variation. The conventional application of G-theory leaves the residual term unexplained.

Subjects, Measures, and Procedures

The above variance components associated with the label condition were estimated in two samples using the previously described p x I x L design (which is the same as p x i x l). In

one sample, 173 graduate students in education were administered seven items on a 4-point scale that measured attitudes toward quantitative methodology. The four-point scale of the items were anchored by two kinds of labels. One labeling has 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 one week apart.

The other sample consisted of 108 graduate students in education who responded to the 8-item Life Orientation Test (Scheier & Carver, 1985). A 6-point scale was used which was either fully labeled for all the points or labeled only at two ends. Subjects responded to the items using both labeling formats. The two administrations were one week apart. Order of administrations of the two labels was mixed.

Results and Discussion

Table 1 contains, for both measures, the means and standard deviations of the items obtained from two labeling formats and the correlations between the same items of the two labels. For both measures, the means and standard deviations were consistent across the two labels. Table 2 contains, for both measures, the variance components from the two-facet random effect d-study design, p x I x L.

For both measures, $\hat{\sigma}^2(pL)$, which indicates interference of

labeling on the relative standings of persons, averaging over items, is moderate. It accounts for four and six percent of the observed variance for the two measurements, respectively. 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, $\hat{\sigma}^2(L)$, is almost zero for both measures (negative variance is treated as zero). Averaging over persons and items, there is almost no variance among μ_L. In other words, the same means or total scores will be obtained using different labels. This result shows the dependability of labeling for an absolute (domain-referenced) interpretation of observations. Finally, $\hat{\sigma}^2(IL)$ represents less than one percent of the observed variance for both measurements. There is little inconsistency among combinations of a label and an item. Thus, item calibration is consistent for different labels.

For one study, $E\hat{\rho}^2 = .6233$, $\hat{\Phi} = .6175$. For the other study, $E\hat{\rho}^2 = .4917$, $\hat{\Phi} = .4859$. The moderate coefficients are the direct result of the large $\hat{\sigma}^2(pIL)$ which will be greatly reduced if multiple items are sampled.

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 a normative and absolute interpretation of individual differences with respect to what is being measured. One 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. This finding also implies that researchers can free themselves from the concern and effort in choosing verbal labels the connotative intensity of which can be quantified into equidistant intervals. 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. This practical implication is particularly of value to researchers who sometimes find it necessary to change the number of scale points of an instrument to meet particular research needs. In such practice, adding or reducing the numerical scale points or interchanging between even and odd scale points often results in incomparability between the numerical equal distances and the psychological valencies of the labels anchoring the numerical points.

One weakness of the study is the possible memory effect of the subjects in responding to the same questionnaire twice which can not be determined or assessed given the way the study was designed and implemented. There is a need for further research that employs a nested design where respondents are randomly

assigned to different labeling scheems 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). Findings of the present study suggest two plausible explanations for respondents' rating behavior.

First, 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' response to the numerical information makes it negligible whether or not the connotative strength of different anchoring labels represents the

same equidistance underlined by numerals of a scale. Perhaps, this is partially why the earlier psychophysical research on scaling has been discontinued. It becomes somewhat insignificant to gauge the exact connotative intensity of verbal descriptors as long as they are underscored by numerals to which subjects respond.

The speculation that subjects use primarily the numerals but not the verbal labels when rating a Likert-type scale can find its support in some of the common designs of a survey instrument. Most of the instruments using a Likert-type scale have the anchoring labels appear only at the beginning of the instrument or at the beginning of each page of a questionnaire. In rating each item, the respondents are nearly exclusively exposed to numerals but not to the verbal labels. If efficiency or economy has been the reason behind not repeating the same set of labels for every item on an instrument, this practice has certainly also served the purpose of forcing respondents to respond to the equidistant numerals. Anyone who has responded to Likert-type scales probably can recall that the thought of "which number" rather than "which label" one checked for earlier items influenced his/her decision on "which number" but not "which label" to check for a later item.

The other speculation into the cognitive process of using Likert-type scales is that, instead of judging for an equidistance, subjects may well be responding to the order of attitudinal valency which usually is represented adequately by

both the numerals and the labels. Thus, even though "2 = disagree" versus "3 = agree" may represent a larger magnitude of attitude change than "3 = agree" versus "4 = strongly agree", subjects can ignore this discrepancy and simply respond to the ordinal information. In this case, the numerals are treated as representing rankings rather than the algebraic relations as they are intended. Then there is little discrepancy between the labels and the numerals they anchor; both of them indicate ordinal relations. With this explanation, there would be little problem in choosing anchoring labels that represent an ordering sequence rather than equal distance. Instead, a more serious potential problem lies in whether the data obtained from a Likert-type scale can be treated as interval for certain statistical and psychometrical analyses. This, however, is a different and long-standing issue.

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., & 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.

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.

Fin, 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.

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.

Table 1

Means, Standard Deviations, and Correlations

Item	Mean	SD	Mean	SD	Corr
		Six-Point Scale			
	Non-Labeled		Labeled		
1	4.14	1.44	4.53	1.06	.71
2	4.59	1.25	4.69	1.12	.63
3	4.[illegible]7	1.05	4.58	1.05	.67
4	4.84	1.02	4.80	1.09	.70
5	4.91	1.14	4.88	0.99	.52
6	4.97	1.05	4.97	0.97	.56
7	4.81	1.14	4.57	1.05	.66
8	4.87	1.24	4.96	1.05	.57
		Four-Point Scale			
	Label A		Label B		
1	3.18	0.67	3.30	0.77	.48
2	2.73	0.78	2.79	0.92	.69
3	2.90	0.72	2.80	0.82	.57
4	3.03	0.66	3.06	0.83	.52
5	2.98	0.74	3.26	0.93	.34
6	3.13	0.71	3.18	0.77	.39
7	3.16	0.72	3.31	0.80	.46

Table 2

Variance Components from p x I x L Random Effect Design

	SS	df	MS	$\hat{\sigma}^2$	$\hat{\sigma}^2$%
	Six-Point Scale, Fully-Labeled vs. 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
	Four-Point Scale, Two Kinds of Labels				
Person (p)	409.72007	172	2.3821	.10417	16.2
Label (L)	4.21181	1	4.2118	.00231	0.4
Item (I)	71.88687	6	11.9811	.03021	4.7
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