CONNOTATIVELY CONSISTENT AND REVERSED CONNOTATIVELY INCONSISTENT ITEMS ARE NOT FULLY EQUIVALENT: GENERALIZABILITY STUDY

LEI CHANG
University of Central Florida

This article redefines what has been referred to as "negatively worded items" in the literature. The new term—"connotatively inconsistent items"—is more nearly accurate because it has a broader base for generalization. Using generalizability theory with a sample of 102 graduate students, the study showed that connotatively consistent and reversed connotatively inconsistent items were not fully equivalent.

The wording of an item is defined as connotatively consistent (CC) or connotatively inconsistent (CI) when the connotation of the item agrees or disagrees, respectively, with that shared by the majority of the items making up a scale or test. This definition expands the meaning of "negatively worded items" used in the literature in that an item does not have to be grammatically or semantically negative to be connotatively inconsistent.

In a Likert-type scale, the order of scale options associated with the CI items is usually reversed prior to data analysis. For instance, such reversals have to be done before (a) computing an estimate of the internal consistency of a scale, (b) making composite scores, or (c) comparing item or scale means. An unstated assumption of this practice is that the connotations of the items have been made uniform by reversing the CI items. In other words, after the CI items have been reversed, all items used for the data analysis bear the same connotation with respect to an underlying construct whether or not the items were originally consistent or inconsistent with the connotation of the construct. This assumption, however, needs empirical verification for two reasons.

The author thanks two anonymous reviewers, William Michael, and Raymond Wolfe for their helpful suggestions on an earlier version of this article.

Educational and Psychological Measurement, Vol. 55 No. 6, December 1995 991-997

First, do people process CC versus CI information differently? The corresponding measurement question is whether or not CC and CI items are perceived to bear the same intended indications of the construct being measured. In comparing four modes of item presentation, Ahlawat (1985) found that "semantically negative and positive item contents do not measure essentially the same construct" (p. 98). He speculated that negative grammar causes cognitive confusion during information processing. Similarly, Mook, Kleijn, and van der Ploeg (1991) identified two factors associated with the connotations of the items when a 20-item anxiety test and a 20-item depression test were factor analyzed together. Each test had 10 CC and 10 CI items, respectively.

Second, is there an ordering effect in selecting different scale options on a Likert-type scale? Endorsement of an item that is either consistent or inconsistent with the connotation of the test implies using the scale options appearing on the two opposite ends of a Likert-type scale. If respondents are particularly drawn to scale options at either end of a scale (primacy or recency effect), a "1" endorsing a CI item on a 7-point Likert-type scale, for example, would not truly equal "7" had the item been CC. Thus reversing the CI item later would not make it equivalent to a CC item. Related research has shown an ordering effect on rating behaviors (e.g., Chase, 1969; Klockars & Yamagishi, 1988; Wildt & Mazis, 1978) and, particularly, a primacy (Carp, 1974; Chan, 1991; Payne, 1972) and recency (McClendon, 1986) effect. In several studies, symptom-negatively worded items were found to have higher means than symptom-positively worded items in several psychological tests (Bernstein & Eveland, 1982; Bonke, Smorenburg, van der Ent, & Spielberger, 1987; DeVito & Kubis, 1983; Mook et al., 1991).

The purpose of the present study was to examine through use of generalizability theory the extent to which scores obtained from a test can be generalized to different measurement conditions represented by consistent and inconsistent connotations of item wording.

Method

Subjects were a convenience sample of 102 master's students in education enrolled in different sections of a research methods course. The students were predominantly White and 80% female. Eight items measuring perceived utility of quantitative methodology and quantitative efficacy were administered two times 1 week apart to these students. The items were taken from the Quantitative Attitude Questionnaire (QAQ) (Chang, in press), which was designed for and validated by graduate students in education and psychology. The wording of four of the eight items were inconsistent with the connotation of the QAQ. The eight items were administered with other QAQ items using a 6-point Likert-type scale. At the second administration, the eight items were rewritten to be the opposites of their original connotations. Thus there were

two observations for each item, one of the CC wording (e.g., "I'm good with math") and the other of the CI counterpart (e.g., "I'm bad with math"). Scale options of the CI items were reversed prior to data analysis.

A three-facet partially nested generalizability study was conducted to determine the possible error variance contributed by the wording, CC versus CI, of items. In this design, people ($N = 102$) were the object of measurement. The three facets were wording ($N = 2$), scale ($N = 2$), and item ($N = 4$). Items were nested within scales, and the rest of the facets were crossed with each other. The wording facet was considered fixed because for the purpose of this study, CC and CI exhaust all the wording conditions in the universe. The scale facet was first considered random, assuming that the two scales were randomly sampled from an infinite universe of (four-item) scales or tests. It was subsequently fixed to evaluate better the wording effect within each scale. The item facet was random.

Results

The purpose of a G study usually is to estimate variance components—universe score or true score variance as well as different error variances resulting from different measurement conditions. These variance components are presented in Table 1.

To determine the exchangeability of the two kinds of item wordings in the G study phase, variance components involving the wording facet are of primary interest. The main effect of wording, which was almost zero ($\hat{\sigma}^2(w) = .01155$, or 0.5%), indicated that the mean score averaging over persons, items, and scales was the same for the two wordings. Because the 6-point scale contained items of both wordings, when averaging over people and items, the two wordings should have the same means unless respondents uniformly rated all items high or low (because of response-ordering effects or information-processing differences) when using one wording or the other. The present result indicated the lack of an overall wording effect.

The purpose of measurement is to capture the normative individual differences that, in this study, are represented by the person facet. An interaction between the person facet and another facet represents errors in transmitting the individual differences as a result of the observation conditions sampled within that facet. The relatively large variance component from the interaction between person and wording ($\hat{\sigma}^2(pw) = .1159$, or 5.5%) indicated that, to some degree, individual differences averaging over items and scales differed across the two wordings. In other words, the relative standing of a person averaged over items and scales may be high when obtained by CC wording but low when obtained by CI wording, or vice versa. Similarly, $\hat{\sigma}^2(pws)$, representing about 9% of the expected variance, which was relatively high, revealed that the unwanted influence of wordings (on people's relative standings) was also different across scales. The quantity,

Table 1
Variance Components From Mixed p × *w* × (i:s) *Design With Fixed* w

	df	MS	$\hat{\sigma}^2$	$(\hat{\sigma}^2)$%
Person (*p*	101	7.819	.23601*	11.1
Wording (*w*)	1	23.061	.01155	0.5
Scale (*s*)	1	201.885	.22187*	10.4
Item (*i:s*)	6	17.638	.07572	3.5
pw	101	2.215	.11590	5.4
ps	101	4.043	.55655*	26.1
ws	1	12.706	.02469	1.2
pi:s	606	0.848	.16938	7.9
wi:s	6	1.853	.01317	0.6
pws	101	1.288	.19467	9.1
pwi:s	606	0.509	.50889	23.9

Note. All facets were assumed random and variance estimates were obtained by applying the following equations:

$\hat{\sigma}^2(pwi{:}s) = \text{MS}(pwi{:}s)$

$\hat{\sigma}^2(pws) = [\text{MS}(pws) - \text{MS}(pwi{:}s)]/n_i$

$\hat{\sigma}^2(pi{:}s) = [\text{MS}(pi{:}s) - \text{MS}(pwi{:}s)]/n_w$

$\hat{\sigma}^2(wi{:}s) = [\text{MS}(wi{:}s) - \text{MS}(pwi{:}s)]/n_p$

$\hat{\sigma}^2(pw) = [\text{MS}(pw) - \text{MS}(pws)]/n_i n_s$

$\hat{\sigma}^2(ps) = [\text{MS}(ps) - \text{MS}(pi{:}s) - \text{MS}(pws) + \text{MS}(pwi{:}s)]/n_i n_w$

$\hat{\sigma}^2(ws) = [\text{MS}(ws) - \text{MS}(wi{:}s) - \text{MS}(pws) + \text{MS}(pwi{:}s)]/n_i n_p$

$\hat{\sigma}^2(i{:}s) = [\text{MS}(i{:}s) - \text{MS}(wi{:}s) - \text{MS}(pi{:}s) + \text{MS}(pwi{:}s)]/n_w n_p$

$\hat{\sigma}^2(p) = [\text{MS}(p) - \text{MS}(ps) - \text{MS}(pw) + \text{MS}(pws)]/n_w n_i n_s$

$\hat{\sigma}^2(w) = [\text{MS}(w) - \text{MS}(pw) - \text{MS}(ws) + \text{MS}(pws)]/n_p n_i n_s$

$\hat{\sigma}^2(s) = [\text{MS}(s) - \text{MS}(i{:}s) - \text{MS}(ws) + \text{MS}(ps) - \text{MS}(pwi{:}s) + \text{MS}(wi{:}s) + \text{MS}(pi{:}s) + \text{MS}(pws)]/n_p n_i n_w$

$\hat{\sigma}^2(p)^* = \hat{\sigma}^2(p) + \hat{\sigma}^2(pw)/n_w$

$\hat{\sigma}^2(s)^* = \hat{\sigma}^2(s) + \hat{\sigma}^2(ws)/n_w$

$\hat{\sigma}^2(ps)^* = \hat{\sigma}^2(ps) + \hat{\sigma}^2(psw)/n_w$

$\hat{\sigma}^2(wi{:}s)$, which was trivial (.0131, or 0.6%), suggested that wordings did not affect the factorial structure of the items. In other words, the differences among items, $\hat{\sigma}^2(i{:}s)$, were maintained for both wordings. Finally, $\hat{\sigma}^2(pwi{:}s)$ was unexplainable in generalizability theory. It represented a combination of random error and systematic variances unaccounted for by the design employed.

The interaction between person and scale was the largest among all the variance components ($\hat{\sigma}^2(ps)$ = .5565, or 26%). This large variance component indicated that relative standings of people differed substantially on the two scales. In factor analysis terminology, there were two distinct factors. This substantial difference across scales called for separate analyses for each scale (Shavelson & Webb, 1991). Two-facet $p \times w \times i$ mixed designs were analyzed for each of the two scales. Similar but stronger results were obtained. They are reported in Table 2.

Table 2
Variance Estimates From Mixed p × *w* × *i Design With Fixed* w

Source	Scale 1		Scale 2	
	$\hat{\sigma}^2$	$(\hat{\sigma}^2)$%	$\hat{\sigma}^2$	$(\hat{\sigma}^2)$%
Person (*p*)	.73413	34	.53678	27
Wording (*w*)	.00100	0	.08038	4
Item (*i*)	.14880	7	.17204	8
pw	.35266	16	.26848	14
pi	.27098	12	.26424	13
wi	.16646	8	.17230	9
pwi	.50134	23	.51643	25

Again, the main effect of wording had little or no variance. But the Person × Wording interaction, which had a large variance component (Scale 1: $\hat{\sigma}^2(pw) = .3526$, or 16%; Scale 2: $\hat{\sigma}^2(pw) = .2684$, or 14%), demonstrated that the relative positions of people were affected by the different wordings.

As wordings were found to confound the normative individual differences, a subsequent practical question was which wording, CC versus CI, had higher estimates of reliability. To answer the question, four single facet G studies (where people were crossed with items) and the subsequent D studies having the same design were conducted within each of the two scales and two wordings. Generalizability coefficients reflecting decisions for using different numbers of items are reported in Table 3. In the design of the study, they were equivalent to forecasts of internal consistency reliability using the Spearman-Brown formula. Overall, scales made up of CC items had higher reliability estimates than did those consisting of reversed CI counterparts.

Discussion

This study represented an initial effort to examine the psychometric assumptions for a widely used practice, namely, using CI items and reversing their scale options prior to data analysis. The results cast doubt on the legitimacy of this practice. Reversed CI items were not fully exchangeable with their CC counterparts. One recommendation is to avoid the use of CI items when possible. When such items must be used, they should be carefully examined. A prudent approach would be to check first whether the means and variances of the reversed CI items are comparable with those of the CC items bearing on the same content or construct. Any such differences should then be taken into consideration in subsequent data interpretation.

One limitation of the study was that wording was confounded with retest. Because the two administrations of the same test were 1 week apart, memory

Table 3
D Study Results From G Study of the Same Design

		G study	Decision study		
n'_i =		1	4	8	12
G coefficient ($\hat{\rho}_{Rel}$)[a]					
Scale 1	CC items	.62	.87	.93	.95
	CI items	.52	.81	.89	.93
Scale 2	CC items	.57	.84	.91	.94
	CI items	.38	.71	.83	.88
Relative error variance ($\hat{\sigma}^2_{Rel}$)[b]					
Scale 1	CC items	.66	.16	.08	.05
	CI items	.67	.17	.08	.06
Scale 2	CC items	.69	.17	.09	.06
	CI items	.68	.17	.09	.05
Universe score variance ($\hat{\sigma}^2_p$)					
Scale 1	CC items	1.09			
	CI items	.73			
Scale 2	CC items	.92			
	CI items	.43			

a. $\hat{\rho}_{Rel} = \hat{\sigma}^2_p/(\hat{\sigma}^2_p + \hat{\sigma}^2_{Rel})$.
b. $\hat{\sigma}^2_{Rel} = \hat{\sigma}^2_p/n'_i$.

and/or boredom effects could have confounded the finding in either direction (an outcome contributing to the consistency or inconsistency between CC and reversed CI items). Controlled experiments comparing matched groups or groups of random assignment should be conducted in future studies to determine whether the results triangulate on the findings obtained from the present crossed design. Because students' responses were collected by the instructor in a nonanonymous manner, ecological validity might be in question. However, such a response effect, if present, was not expected to affect the CC and CI wordings differentially. The sample size and number of items used in this study were relatively small. Precision of variance estimation can be especially improved in future studies by using more items.

Further research should also look into the cognitive characteristics in processing CI information as a possible contributor to the difference between CC and reversed CI items. It is unknown whether the inconsistency between CC and reversed CI items [$\sigma^2(pw)$] found in this study was due to ordering of response options or to idiosyncracies in processing connotatively inconsistent information. It is important to untangle this confounding situation in future research in that, if the inconsistency between CC and CI items were due to ordering of response options, such a finding would imply that tests involving only CC items may be contaminated by a response set. On the other hand, if there were a cognitive difference in responding to items of consistent

versus inconsistent connotations, the different response behaviors also might be different as a function of the content of the items. Items included in the present study were almost exclusively "opinion items" representing a person's point of view, whereas response patterns may be different in relation to "factual items" that reflect observable behaviors or phenomena. Items employed in further research also may be distinguished in terms of whether they address self versus others, or people versus things.

References

Ahlawat, K. S. (1985). On the negative valence items in self-report measures. *Journal of General Psychology, 112*(1), 89-99.

Bernstein, I. R., & Eveland, D. C. (1982). State vs. trait anxiety: A case study in confirmatory factor analysis. *Personality and Individual Differences, 3*, 361-372.

Bonke, B., Smorenburg, J. M., van der Ent, C. K., & Spielberger, C. D. (1987). Evidence of denial and item intensity specificity in the State-Trait Anxiety Inventory. *Personality and Individual Differences, 8*, 185-191.

Carp, F. M. (1974). Position effects on interview responses. *Journal of Gerontology, 29*, 581-587.

Chan, J. C. (1991). Response-order effects in Likert-type scales. *Educational and Psychological Measurement, 51*, 531-540.

Chang, L. (in press). The Quantitative Attitude Questionnaire: Instrument development and validation. *Educational and Psychological Measurement.*

Chase, C. I. (1969). Often is where you find it. *American Psychologist, 24*, 1043.

DeVito, A. J., & Kubis, J. F. (1983). Alternate forms of the State-Trait Anxiety Inventory. *Educational and Psychological Measurement, 43*, 729-734.

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

McClendon, M. J. (1986). Response-order effects for dichotomous questions. *Social Science Quarterly, 67*, 205-211.

Mook, J., Kleijn, W. C., & van der Ploeg, H. M. (1991). Symptom-positively and -negatively worded items in two popular self-report inventories of anxiety and depression. *Psychological Reports, 69*, 551-560.

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.

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

Wildt, A. R., & Mazis, M. B. (1978). Determinants of scale response: Label versus position. *Journal of Marketing Research, 15*, 261-267.