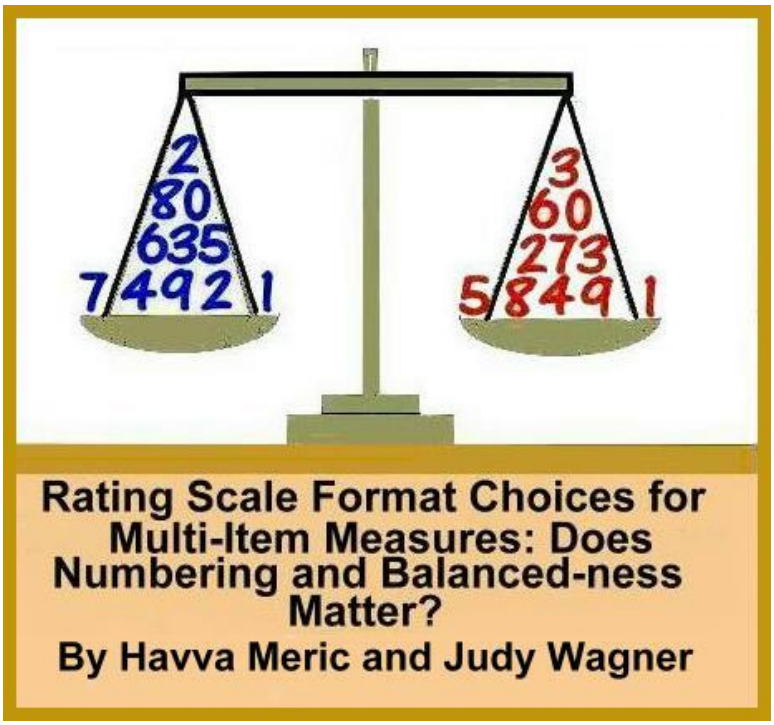

Peer reviewed

Havva Meric (merich@mail.ecu.edu) is an Associate Professor at East Carolina University. Judy Wagner (wagnerj@mail.ecu.edu) is an Assistant Professor at East Carolina University.

Abstract

Knowledge of the influence of variations in rating scale characteristics generally is limited by the restriction of study to single scale items. This experiment broadens the scope in this research area by studying the impact of varying numerical assignment to scale points and scale balanced-ness on the responses and internal consistency reliability of a well established multi-item instrument. Results indicate that this numerical format manipulation significantly influenced mean responses but not scale reliability. The balanced-ness manipulation also did not alter reliability, but positively unbalanced scales produced higher means than balanced scales. Implications and suggestions for future research are discussed.

Introduction

Rating scales are important for the construction of measurement instruments in both applied and basic business research. Because of this frequent reliance on rating scales, it is crucial for researchers to understand how the properties of these measures may influence various aspects of responses gathered from their use. Recent thinking indicates that once respondents are given a questionnaire, all context features of that questionnaire become relevant information and are used in judging, interpreting, and deciding how to respond to the alternatives provided. In some sense, the context of the questionnaire becomes part of an ongoing conversation between the respondent and the researcher. Therefore, while words constitute an important source of question-meaning for respondents, so also do all other formal features or nonverbal language (numbers, symbols or graphics) that are used in self-administered questionnaires (Schwarz 1996; Christian and Dillman 2004).

Some empirical evidence exists to suggest that altering a variety of characteristics of rating scales may significantly affect outcomes. Topics such as scale category descriptors, negatively worded scale items, left-hand versus right-hand bias, and numerical assignments to scale points have received some attention in the literature (e.g. Wildt and Mazis 1978; Friedman 1988; Friedman, Herskovitz and Pollack 1994; Friedman, Wilamowski, and Friedman 1981; Schwarz, et al. 1991).

While the extant literature provides some knowledge regarding a number of aspects of rating scales, much remains unknown about this topic. As a result, yet unstudied characteristics of rating scales are commonly used in a variety of ways that could potentially influence respondents' answers and ultimately bias reported findings. The scope of existing research in this area has primarily been restricted; first, to manipulations of single scale items and second, to evaluations of how format properties of rating scales influence response distributions (e.g. Schwarz, et al. 1991), with psychometric issues such as reliability and validity generally neglected. To date, we could find only one scale-effect study (Johnson, Bristow and Schneider 2004) that had deviated from the norm of single scale items by incorporating a multi-item instrument in the experimental design and could thus examine properties such as internal consistency among items. Additional scale-effect research using multi-item measures is needed to better understand whether, and if so how, this type of scale may behave differently than single-item scales when scale characteristics are varied. The aim of the current study is to begin to fill this gap.

Better understanding of scale format influences on multi-item scale properties is important and will assist future researchers in constructing better measures, particularly in disciplines such as marketing and organizational behavior, where multi-item rating scales are commonly used. In an era of sophisticated statistical techniques such as causal modeling, it is important that business researchers are ever aware that complex statistical results are only as accurate as

their measurement instruments. Drawing on this need for further research, this study seeks to uniquely contribute to and extend previous rating scale-format research by evaluating the effects of manipulating assignment of numerical values to scale points and balanced-ness of scale labels on response distributions, as well as on the internal consistency reliability, of a well established multi-item scale.

Literature Review

Due to the wide spread use of rating scales in measurement instruments in both applied and basic business research, the properties of rating scales have been subject to investigation by some researchers. For a comprehensive discussion of rating scales' general and psychometric properties and justification for their use, the reader is referred to Dawes and Smith (1985) and Churchill and Peter (1984). Findings that have a particular relevance for the current study are discussed below.

Label and Position Effects

Research focusing on verbal scale properties has highlighted the importance of the selection and positioning of labels. Krosnick and Berent (1990) found verbal scales with labels for all scale points to be more reliable than scales with labels only at the endpoints. Various studies report that whether verbal descriptors are used only at endpoints or at every scale point also may affect the distribution of the data obtained (Rohrmann 1998; Wegener 1983). Further, it is important that researchers creating interval or ordinal scales carefully select scale point labels that represent equal intervals or that are ordinal in nature (e.g. Jones and Thurstone 1955; Myers and Warner 1968; Wildt and Mazis 1978). In general, these findings provide evidence that an individual's cognitive responses are influenced by the labels that are provided to them (Rohrmann 1998; Wildt and Mazis 1978).

Wildt and Mazis (1978) were the first investigators to ask if subjects respond to the label relative to the endpoints of the scale. They found that both the category labels and the relative position of the descriptors altered response distributions. Friedman, Cohen and Amoo (2003) further tested the label and the position effects and ascertain again that the category labels, rather than label positions relative to the endpoints, made an impact on the distribution of responses.

There has been some previous research that has investigated the possibility of a left-right bias in verbal rating scales. Findings support the existence of a bias toward the left-side of the scale. For example, Holmes (1974), Friedman, Friedman and Gluck (1988) and Belson (1966) found that the respondents were more likely to use the negative end of the scale if presented first (on the left-hand side), rather than last (on the right-hand side). Friedman, Herskovitz and Pollack (1994) revealed that a Likert scale with the positive label, "strongly agree," on the left side created a greater degree of respondent leniency than a Likert scale with the negative label, "strongly disagree," on the left side.

Balanced vs. Unbalanced Scales

Typically, rating scales used in research have been balanced. Balanced scales are composed of an equal, or balanced, number of favorable and unfavorable labels forming an equal-interval continuum anchored by opposite poles, with or without midpoints. A few studies have investigated the impact of balanced versus unbalanced scales on response distributions and variability. Findings have been mixed. Wildt and Mazis (1978) and Friedman, Wilamowski and Friedman (1981) found that unbalanced scales did not produce the same distribution results as balanced scales. Wildt and Mazis (1978) argue that both the position of the label in the hierarchy of categories provided and the psychological meaning of the label descriptor influence how respondents answer questions. Friedman, Wilamowaki and Friedman (1981) reported that a positively unbalanced scale produced a higher mean than a balanced scale and that a balanced scale produced a higher mean than a negatively balanced scale. Contrary results for balanced and unbalanced rating scales exist in educational research. For example, Lam and Klockars (1982) reported that negatively packed scales (majority of the scale points corresponding to negative labels) produced the highest mean rating of instructors and that positively packed scales (majority of the scale points corresponding to positive labels) produced the lowest mean ratings. To our knowledge, no work to date has extended the scope of this topic to include an investigation of the impact of balanced or unbalanced rating scale labeling on psychometric scale properties such as reliability measures for either single rating scale item measures or for a multi-item measurement instrument.

Numerical Format

Past research has also investigated the effects of varying numerical value assignment to scale points. Schwarz et al. (1991) experimented by either assigning numerical values of 0 to 10 or of -5 to +5 to 11 point scales. Results showed very different response distributions for equivalent positions on the left-hand half of the scale. When numerical end values were 0 and 10, 34 percent of respondents chose values between 0 and 5, where as, when numerical end values were -5 to +5, only 13 percent of respondents chose values between -5 to 0. Values ranging from 0 to 10 appeared to influence respondent cognition differently by suggesting an “absence or presence of an attribute. When the label “not at all successful” was combined with numeric value 0, respondents seemed to interpret it to reflect the absence of success. However, when the same label was combined with the numeric value of -5, subjects may have interpreted it to reflect the presence of failure. This suggests that the choice of the numerical values assigned to scale end points may create different interpretations, therefore creating different distributions, and may even impose artificial restrictions in item variance.

Schwarz et al. (1998) investigated the impact of numeric values on respondents’ interpretation of vague quantifiers, such as “rarely.” They report that respondents may have interpreted “rarely” as indicating a lower frequency when

paired with the numeric value of 0 (on a scale from 0 to 10) than when paired with the numeric value of 1 (on a scale ranging between 1 and 11). The authors concluded that the impact of numeric values is not restricted to use of a unipolar (e.g. 1 to 11) scale or a bipolar (+5 to -5) scale. It seemed that respondents cognitively related numerical values to verbal labels, drawing on the meaning of this connection and using these interpretations when responding.

The aforementioned research suggests a strong impact on response distributions of varying numerical value assignments to verbal label points of a rating scale. There is not sufficient research in this area of inquiry to understand when or why this happens. Therefore, it is impossible to generate a rule of thumb regarding for what type and when numerical value assignments will impact response distributions in a specific way. It seems that when designing a questionnaire, most researchers typically select a certain type of numerical value assignment based on their personal preference. For example, when discussing this issue, the current authors discovered that one of them prefers Strongly Agree to be assigned “6” and Agree to be assigned “5” and so on while the other author prefers the opposite, with Strongly Agree being assigned “1”, Agree “2”, and so on. It is a personal preference, and one can see both types of numerical value assignments in actual practice. However, the thought that these variations in numerical assignments do not make any difference in responses may be a very dangerous assumption, if in reality, respondents draw different meanings from associations between verbal labels and a variety of numerical value assignments. Thus, there seems to be an important need for research to understand the connection between different types of numerical value assignments, their impact on response distributions, as well as their influence on other properties of rating scales, such as reliability measures.

Taken together, the reviewed literature suggests that all “formal context” features of a questionnaire (including response alternatives provided with choice of label descriptors, number values, order of labels, etc.) help respondents determine what is expected of them in a survey interview. Given this evidence for the influence of context features, it is important that any contextual variations in scale construction that may impact distribution of results and/or psychometric properties be empirically investigated as they may potentially impact measurement validity.

The Study

The current study focuses on two context features of rating scales that merit greater attention and understanding, numerical format and balanced-ness. The first objective of this research is to extend understanding of the effects of variations in numerical value assignments to rating scale points on response distributions and an important psychometric property, the internal consistency reliability measure. In marketing research, it is common to see the assignment of numerical values to scale points in either an ascending order (1= completely satisfied to 6=completely dissatisfied) or a descending order (6=completely satisfied to 1=completely dissatisfied). This practice seems to assume that it makes no difference whether

numerical values are assigned in either fashion. For example, Aiello, Czepiel and Rosenberg (1977) and Friedman (1988) use a descending numerical assignment format while Johnson and Schneider (2004) and Schwarz et al. (1991) use an ascending numerical assignment format.

A review of the relevant literature reveals no empirical investigation of the impact of assigning ascending or descending numerical values to scale points. Yet, if earlier results relating to numerical values (Schwarz et. al 1998) are correct and respondents are able to associate numerical values with categorical labels in their response tasks, then it is possible that ascending and descending assignment of numerical values to rating scale points may produce different response distributions. In addition, internal consistency reliability of a rating scale may be affected. If on the other hand, ascending and descending numerical value assignments do not produce different response patterns, then one might conclude that, in this particular case, respondents are paying more attention to the verbal descriptors as contextual information than to numerical value assignments. By the same token, if the reliability of a multi-item scale is not influenced by using ascending vs. descending numerical format, then that particular scale might be considered robust to those variations in scale format. In either case, the knowledge will be important to future researchers.

A second objective of this study is to extend current understanding of the effect of balanced and unbalanced scale formats on response distributions and the reliability structure of rating scales. The topic of balanced versus unbalanced scales has received very little attention in the literature, and their effect on reliability of a multi-item scale has yet to be addressed. Balanced scales may be appropriate for most variables of interest. Yet, there are instances where researchers are known to prefer the use of unbalanced scales. For example, Friedman et al. (1981) point out Gerber Food's use of the following categories: "excellent," "very good," "good," "not so good," and "poor." Gerber's aim in using this unbalanced and positively packed scale may relate to the desire to obtain results that better distinguish between ratings of "good" and "excellent." Though very limited in number, studies that have examined balanced vs. unbalanced scales generally support the notion that positively unbalanced scales may create higher mean ratings (Wildt and Mazis 1978; Friedman, Wilamowski and Friedman 1981). Because the extant research on scale balanced-ness used single item rating scales and therefore did not evaluate reliability, it remains an empirical question as to how a multi-item measurement scale would behave when the scale format is unbalanced.

To our knowledge, with one exception, research focusing on the characteristics of rating scales has used only single item measures in their various manipulations. The exception, Johnson, Bristow and Schneider (2004), examined the impact of negatively worded and double negatively worded rating scale items on the internal reliability and factor structure of four of the items from the seven-item Fashion Consciousness Scale (Lumpkin and Darden 1982). Their results indicated that both unfavorable and double-negative wording lowered the internal reliability of that particular scale when compared to the control condition that used positively

worded items. Thus, in the one known study that used a multi-item scale to examine variations in rating-scale characteristics, the internal reliability of the scale was shown to be compromised in some cases. Therefore, due to the dearth of research into rating-scale-properties using a multi-item scale and the very common use of multi-item scales in research, a third and very important objective of the current study is to expand our understanding of how variation of context features impacts this scale form. Specifically, this research empirically tests the effect of the context variables of numerical format and balanced-ness on subject responses and scale reliability using a well known and widely researched multi-item scale, CETSCALE (Shimp and Sharma 1987).

Sample

Four versions of self-administered questionnaires were randomly assigned and administered to undergraduate students enrolled in business classes at a large university in a college town in the eastern United States. Students were deemed appropriate respondents for this experiment because it involved opinions relating to whether an individual should restrict purchases to American products, and students are experienced consumers. All questionnaires were distributed and completed during normal class periods. In total, 240 questionnaires were distributed, and 224 usable forms were obtained. The final sample consisted of 112 females and 112 males.

Methodology

CETSCALE , a well established multi-item scale, was chosen as the basis for the manipulations of the two contextual scale formats of this study. CETSCALE, developed by Shimp and Sharma in 1987, is the most well-known scale measuring consumers' ethnocentric tendencies. The term “consumer ethnocentrism” was first applied by Shimp and Sharma (1987) to represent “the beliefs that are held by American consumers about the appropriateness, indeed morality, of purchasing foreign made products.” Although originally developed as a measure of American consumers' ethnocentric tendencies, CETSCALE was subsequently applied and its psychometric properties validated internationally in Japan, France, and Germany (Netemeyer, Durvasula and Lichtenstein 1991), Korea (Sharma, Shimp and Shin 1995), Russia (Durvasula, Craig and Netemeyer 1997), and China (Klein, Ettenson and Morris 1998). The original CETSCALE consists of 17 items and uses a modified Likert scale format (1=strongly disagree and 7=strongly agree) without a midpoint and with only endpoint labels. The 17 items of CETSCALE and a number of items requesting basic demographic information for the sample were included in the paper and pencil questionnaire for the current research.

The experiment used a 2 (numerical format) x 2 (balanced-ness format) randomized block between-subjects design. Gender was the blocking variable. Numerical format was manipulated as ascending (1 to 6) or descending (6 to 1) numerical assignments to scale points. Balanced-ness was manipulated as either a

balanced scale consisting of an equal number of favorable and unfavorable responses or an unbalanced scale with more favorable than unfavorable scale points. To enable a strong manipulation of balanced-ness for this study, a six point modified Likert-type rating scale with all scale points labeled by verbal descriptors and without the provision of a midpoint was utilized. The descriptor indicating the greatest degree of agreement was always presented on the left side of the scale. Table 1 displays the manipulations.

Table 1

Experimental Manipulations

Example question: "Purchasing foreign-made products is un-American".

Manipulation	**Scale Responses**					
Ascending/Balanced	*Very Strongly Agree* 1	*Strongly Agree* 2	*Agree* 3	*Disagree* 4	*Strongly Disagree* 5	*Very Strongly Disagree* 6
Descending/Balanced	*Very Strongly Agree* 6	*Strongly Agree* 5	*Agree* 4	*Disagree* 3	*Strongly Disagree* 2	*Very Strongly Disagree* 1
Ascending/Unbalanced	*Very Strongly Agree* 1	*Strongly Agree* 2	*Agree* 3	*Some-what Agree* 4	*Disagree* 5	*Very Strongly Disagree* 6
Descending/Unbalanced	*Very Strongly Agree* 6	*Strongly Agree* 5	*Agree* 4	*Some-what Agree* 3	*Disagree* 2	*Very Strongly Disagree* 1

Analysis and Results

For the analysis, the data were recoded so that the order of scores for both scales was from 1 to 6. In other words, responses in the descending treatment were reverse scored to enable statistical comparisons with responses in the ascending treatment. Because of this coding scheme, lower scores indicate higher degrees of ethnocentrism while higher scores denote less ethnocentrism. Descriptive statistics for the four experimental treatments are displayed in Table 2. Total CETSCALE sum score mean ratings, standard deviations, individual scale item mean ratings and standard deviations are provided. In each case the cell size was 56; comprised of an equal number of male (28) and female (28) respondents. Differences in means and standard deviations are discussed in a later section. The proportion of

completed questionnaires did not differ significantly by treatment (X^2=3.5, df = 3, sig. = .32).

Table 2

Descriptive Statistics

Scale Item	Item Mean		Item Std. Dev.		Item Mean		Item Std. Dev.	
	Ascend	Descend	Ascend	Descend	Balanced	Unbal	Balanced	Unbal
1	3.86	4.08	1.125	1.169	3.63	4.30	1.131	1.073
2	4.00	4.03	1.153	1.111	3.84	4.21	1.135	1.100
3	2.92	3.12	.916	.949	2.98	3.08	.771	1.083
4	3.75	3.81	.964	1.091	3.50	4.04	.910	1.069
5	4.54	4.71	.973	.972	4.22	5.04	.887	.884
6	4.26	4.38	.904	.822	4.07	4.57	.791	.867
7	4.36	4.56	.846	.889	4.12	4.83	.825	.770
8	3.85	3.93	.966	.972	3.70	4.15	.957	.932
9	3.94	4.07	.979	1.059	3.79	4.26	.963	1.029
10	4.19	4.35	1.025	.833	3.99	4.59	.963	.812
11	4.28	4.32	.879	.782	4.10	4.53	.838	.771
12	4.12	4.22	.911	.866	3.85	4.53	.851	.794
13	3.64	3.82	.880	1.052	3.59	3.88	.906	1.020
14	4.58	4.64	.864	.844	4.25	5.00	.844	.684
15	4.12	4.44	1.052	.911	3.94	4.60	.903	.981
16	3.92	4.07	.969	.979	3.81	4.21	.945	.972
17	4.53	4.57	.937	.855	4.28	4.85	.872	.830
Total Mean Score	68.76	71.59			65.65	74.68		
SD			11.860	12.353			11.560	11.065

Means Ratings and Variance

One objective of this research was to determine whether scale responses would be influenced by the manipulations of the ascending or descending numerical formats and the balanced or unbalanced response labels used in the experiment. A GLM analysis (see Table 3) was used to determine if significant mean differences existed between the four experimental groups. Following a blocked design, an initial analysis included gender in the model. Gender was not significant (F_1=.880, p=.35), and therefore is not part of the final model. The overall final model was significant (F_3=13.684, p< .000). The Levene statistic for the GLM model was

nonsignificant ($F_{3,220}$=.099, p=.96), suggesting that the variances were similar for all experimental conditions. Because this is an exploratory study, a full model was estimated.

The interaction of numerical format and balance was nonsignificant (F_1=1.583, p=.21). There was a main effect for numerical format (F_1=3.358, p=.068). Subjects' total mean scores were significantly higher for descending (M=71.59) versus ascending (M=68.76) numerical assignment. This indicates that subjects have responded as being relatively less ethnocentric in the descending treatment than in the ascending condition. Scale balanced-ness also influenced ratings (F_1=36.044, p< .000). Positively unbalanced scales produced significantly higher ratings (M=74.68) compared to balanced scales (M=65.65). In other words, subjects in the unbalanced scale condition reported being relatively less ethnocentric than those in the balanced treatment. Possible reasons for these findings are presented later in the discussion section.

Table 3

Analysis of Variance

Source	**Type III Sum Of Squares**	**df**	**F**	**Sig.**
Corrected Model	5187.451	3	13.684	.000
Intercept	1102731.323	1	8726.827	.000
Ascending_Descending	424.310	1	3.358	**.068**
Balanced_Unbalanced	4554.528	1	36.044	**.000**
Asc_Des x Bal_Unbal	200.008	1	1.583	.210
Error	27799.438	220		
Total	1135773.000	224		
Corrected Total	32986.888	223		
R Squared = .157	Adjusted R Sq= .146			

Internal Consistency Reliability

We were also interested in whether or not the differences in numerical format and scale balance would affect the internal consistency reliability of a multi-item scale. The Chronbach's alpha is presented below:

Ascending	∂ = .94
Descending	∂ = .95
Balanced	∂ = .95
Unbalanced	∂ = .94

The similar and relatively high reliabilities for all four suggest that the reliability of the CETSCALE is unaffected by the particular manipulations of numerical format and balanced-ness used here. Further, these reliabilities are almost identical to those reported by Shimp and Sharma (1987), which ranged from .94 to .96.

Discussion

The focus of this work was to address an under-researched, but critically important, area of study. In particular, this experiment sought to extend existing knowledge of the ways in which scale format may affect reported empirical results. We tested whether or not varying the numerical format of scale responses by using either ascending or descending numerical values or by constructing balanced or unbalanced scale formats would introduce differences in response distributions of obtained data. Additionally, the study examined whether these manipulations influenced the internal consistency reliability of a multi-item scale.

First, the type of numerical format manipulated here did influence mean responses. Results indicate that descending numerical assignment to the scale points produced significantly higher total scores than ascending numerical assignment to the scale points. This finding remains consistent when individual item means are inspected as well (Table 2). Assignment of descending numerical values to the scale points results in a higher mean score for each of the 17 CETSCALE items.

Further, we inspected the frequency of extreme responses (far left, “very strongly agree” and far right, “very strongly disagree.”) Results show that respondents used the far right of the scale far less frequently for the ascending numerical condition (only 1 out of 17 items) than for the descending numerical treatment (16 of 17 scale items). When the label “very strongly disagree” is assigned “1” rather than “6”, respondents seem to be relatively more willing to choose the lower numbered end of the scale. The emergence of such different patterns of scale usage may be caused by respondent expectations for balance, symmetry, and congruity in question format. Human beings are rational, and they attempt to create a cognitive balance that they believe to be rational (Festinger 1957; Aronson 1997). Assuming a need for congruity, one might expect that certain combinations of verbal labels and numeric value assignments might fit together more naturally in a cognitive sense than other pairings. Explicitly, labels expressing the greatest degree of agreement, “very strongly agree”, or “strongly agree,” might seem more congruent with larger numbers such as “6”, and “5” than with smaller numbers of “1” and “2.” In everyday life, larger amounts of an attribute are typically related to the assignment of larger numbers. For example, the heavier or the warmer of two objects is generally associated with a higher number of pounds or degrees, respectively, than are the lighter or cooler of the twosomes.

In the current study, using the labels “very strongly agree” and “strongly agree,” which indicate the greatest degree of agreement, with the smaller numerical

assignments of “1”and “2” and assigning the labels “very strongly disagree” and “strongly disagree,” which specify the least extent of agreement, to larger numerical values of “6” and “5” might have represented a cognitive mismatch to respondents. Therefore, this study’s finding of higher total mean scores (less ethnocentrism) for the descending numerical condition may be the result of respondents having experienced cognitive incongruity with the ascending numerical treatment.

In addition to mean responses, this research also examined the influence of ascending and descending numerical formats on psychometric properties. Contrary to the vast majority of extant studies that used single item measures, we used an established multi-item scale in our investigation. Varying numerical assignment in an ascending or descending fashion did not significantly influence the internal consistency reliability as measured by Cronbach’s alpha. This suggests that, at least for CETSCALE, this manipulation of numerical assignments does not impact its internal consistency reliability. Therefore, CETSCALE is robust to this type of variation.

The present study also extended research in the area of balanced and unbalanced scale formats. As mentioned earlier, there have been a very limited number of studies that have investigated the influence of balanced-ness of a scale format in survey design. In the current experiment, positively unbalanced scales produced higher total score means than balanced scales. Respondents in the unbalanced treatment group were more willing to disagree (recall that due to data recoding, higher numbers indicate disagreement with individual scale items). Therefore, in the unbalanced condition, respondents were found to be significantly less ethnocentric than subjects in the balanced condition. This result does not agree with some previous balanced-ness research that reported that positively balanced scales produced higher scores (more agreement) for the attitude variable in question (Friedman and Leefer, 1981; Friedman, Cohen and Amoo, 2003). Yet, at least one study in the education domain found the opposite result. Lam and Klockars (1982) reported that negatively packed scales (majority of the scale points corresponding to negative labels) produced the highest mean ratings of instructors while positively packed scales (majority of the scale points corresponding to positive labels) produced lowest mean ratings. In other words, their results indicated that students were less lenient, or gave lower ratings, to their professors when positively packed scales were used. Our results are consistent with the Lam and Klockars’ (1982) finding for positively packed scales. Maybe, allowing fewer unfavorable categories prompted respondents to be more cognizant of unfavorable verbal labels, and, therefore, to be less lenient in the measurement of consumer ethnocentrism. These findings might also be explained by the change in the order position of the “Disagree” label in the balanced versus unbalanced conditions, or, even, by the fact that the number of categories by which respondents could express disagreement is more restricted in the unbalanced than in the balanced conditions.

The same directional result is also true for the individual scale item means (see Table 2). Unbalanced scale treatment results indicate consistently higher

individual mean scores for each of the 17 CETSCALE items. We also further inspected the response patterns at the very far left “very strongly agree” and very far right “very strongly disagree.” For the left hand side, the frequency with which respondents use that end of the scale is similar for both balanced and unbalanced conditions (only 11 more total responses in the unbalanced treatment). The same response pattern is true for the very far right hand of the scale for both treatments (only 12 more total responses for the unbalanced treatment). However, with further inspection of response patterns to each of one of the six scale points, a very different and interesting finding emerges. Respondents in the unbalanced condition are two to four times more likely to choose the two far right labels of the scale (“very strongly disagree” and “disagree”) than the respondents in the balanced condition. More preference for the two far right scale labels in the unbalanced condition holds true for each of the 17 CETSCALE items.

This result, where respondents report themselves as less ethnocentric when a positively unbalanced scale format is used, raises an important question about the impact of balanced-ness as a scale format variable and its relationship to leniency in measurement. There have been some concerns in marketing, in the past, where respondents have been known to be lenient in their ratings of service quality (for example, in the area of patient satisfaction), thus providing inflated scores that are problematic for differentiating between good and bad performance. The findings of this study suggest that it may well be beneficial to use a positively unbalanced scale format in service quality measurement instruments because of the likelihood that the positively unbalanced scale will discourage leniency.

The current study also examined the impact of balanced-ness on the internal consistency reliability of the selected multi-item instrument, CETSCALE. The positively unbalanced manipulation used here did not negatively affect the internal consistency reliability. CETSCALE was found to be robust in this regard. The internal consistency reliability of CETSCALE was relatively high, regardless of the experimental manipulation, and was also very similar to that reported in previous studies validating the CETSCALE in the United States (e.g. Shimp and Sharma 1987).

In summary, our study investigated whether numerical assignment or balanced-ness variations in scale format significantly influenced response distributions and the internal consistency reliability of a well known multi-item scale. “Borrowing” scales is a common practice in both applied and theoretical research in marketing. In some instances, these borrowed scales are modified to suit the preferences and needs of the individual researcher. These adaptations include using different numerical assignments to scale points, varying the number of scale points, at times reversing the position of end-points anchors, and in some instances using unbalanced scales. These variations are often not thought of as significant enough deviations from the original scales to potentially affect response distributions, internal consistency reliability measures or construct validity. The current research suggests that all of these aforementioned variations may indeed influence response

distributions and also raises the question of a potentially detrimental impact on nomological and construct validity of a multi item scale.

Future Research

While this experiment provided insight regarding how two specific aspects of rating-scale format may affect responses and ultimately research findings, perhaps a more significant contribution of this work is that it draws attention to the crucial need for more research of its type and raises many additional questions. For example, as reported above, results here suggest that a researcher's choice or preference for an ascending or descending numerical format may significantly influence the means and the distribution of the data obtained. This study's research instrument used labels and numbers for all scale points at the top of each page rather than repeating labels after each scale item. It is important to note that the numerical-format effect found in the current research is limited to the aforementioned modified Likert scale format. There are other preferences for scale format design for verbal rating scales. Sometimes, researchers repeat labels (and numbers) after each scale item. Other times, they would include the verbal labels and corresponding numerical representations as a part of their instructions and have the respondents enter a number in a single line (or a box) provided at the end of each scale item. Before generalization of our findings, additional research should address the question of whether these results hold across various scale format design choices (e.g., labels above each item, or circling numerals rather than checking boxes, etc.). Extensions such as these will be helpful in isolating and better understanding the numerical-format effect found here.

This study also highlights the need for additional research which examines scale balanced-ness. The literature reveals little work on this topic as well as contradictory findings for the few studies that do exist. As discussed above, results of the current work on scale balanced-ness may provide a possible way to deal with respondent leniency for some research topics when there is such concern. Only through future investigation can support for this notion be determined and the causes of the contradictory findings on scale balanced-ness be resolved.

Despite the proliferation of multi-item scales in research today, to our knowledge only one other study has used a multi-item scale or questioned whether the common practice of varying scale format decisions may influence its psychometric properties. Lam and Klockars (1982) found that reliability decreased for the Fashion Consciousness Scale with the use of negative and double negative wording. In our work, CETSCALE was shown to be robust with respect to internal consistency reliability for the two specific numerical and balanced-ness manipulations that we had designed. Thus, the two scale-effect studies that utilize multi-item scales have produced results that indicate internal consistency reliability may be affected negatively by certain decisions regarding scale format and other decisions may not impact reliability at all. It remains an empirical question as to what other scale format variations may consistently impact a multi-item scale's reliability

negatively and also whether CETSCALE or other established multi-item scales would be robust to other types of rating-scale variations.

Overall, given the central importance of construct validity in marketing research, academics and marketing practitioners alike need to know much more about the optimal choices for scale properties when designing their research instruments. Limited research has pursued this goal in recent decades. Future research should provide the renewed focus that this area of study merits.

References

Aiello, J. A. Czepiel, and L. J. Rosenberg, "Scaling the Heights of Consumer Satisfaction: An Evaluation of Alternative Measures," in *Proceedings of the Research Symposium on Consumer Satisfaction, Dissatisfaction, and Complaining Behavior,* Bloomington, ed. Ralph L. Day, l92-119, Indiana: Graduate School of Business Administration, Indiana University, 1977.

Aronson, E., "A Theory of Cognitive Dissonance," *American Journal of Psychology*, 110(1), 127-137, Spring, 1997.

Belson, W. A., "The Effect of Reversing the Presentation Order of Verbal Rating Scales," *Journal of Advertising Research*, Vol. 6, No. 4, pp. 30-37, 1966.

Christian, L. M. and D. A. Dillman, "The Influence of Graphical and Symbolic Language Manipulations on Responses to Self-Administered Questions," *Public Opinion Quarterly,* Vol. 68, No. 1, pp. 57-79, 2004.

Churchill, G. Jr. and J. P. Peter, "Research Design-Effects on the Reliability of Rating Scales: A Meta Analysis," *Journal of Marketing Research,* Vol. 21 (November), pp. 360-375, 1984.

Cox, E. P., "The Optimal Number of Response Alternatives for a Scale: A Review," *Journal of Marketing Research*, Vol. 17, pp. 407-422, 1980.

Dawes, R. M. and T. Smith, "Attitude and Opinion Measurement," in *Handbook of Social Psychology,* eds. G. Lindzey and E. Aronson, Vol. 2, pp. 509-566. New York: Random House, 1985.

Douglas, S. P. and E. J. Nijssen, "On the Use of Borrowed Scales in Cross-national Research: A Cautionary Note," *International Marketing Review*, Vol. 22, pp. 621-642, 2003.

Dunham T. C. and M. L. Davison, "Effects of Scale Anchors on Student Ratings of Instructors," *Applied Measurement in Education,* Vol. 4, No. 1, pp. 23-35, 1991.

Durvasula, S., A. J. Craig and R. G. Netemeyer, "A Cross-Cultural Comparison of Consumer Ethnocentrism in the United States and Russia," *Journal of International Consumer Marketing,* Vol. 9, No. 4, pp. 73-93, 1997.

Festinger, L. A Theory of Cognitive Dissonance. Stanford University Press: Stanford, CA. 1-31, 1957.

Friedman, H. H., "The Effects of Positive and Negative Wording in Responses to a Likert Scale," *Applied Marketing Research*, Vol. 28, No. 2, pp. 17-22, 1988.

Friedman, H. H., Y. Wilamowsky and L. W. Friedman, “Balanced and Unbalanced Rating Scales: A Comparison,” *Mid-Atlantic Journal of Business*, Vol. 19, pp. 1-7, 1981.

Friedman, H. H., P. J. Herksovitz and S. Pollack, “Biasing Effects of Scale-Checking Styles on Responses to a Likert Scale,” *Proceedings of the American Statistical Association Annual Conference: Survey Research Methods*, pp. 792-795, 1994.

Friedman, H. H., L.W. Friedman and B. Gluck, “The Effects of Scale-Checking Styles on Responses to a Semantic Differential Scale," *Journal of the Market Research Society*, Vol. 30, No. 4, pp. 477-481, 1988.

Friedman, H. H., D. Cohen and T. Amoo, “Label or Position: Which has the Greater Impact on Subjects’ Responses to a Rating Scale?, “*Journal of International Marketing and Marketing Research*, Vol. 28, No. 2, pp. 77-81, 2003.

Holmes, C., “A Statistical Evaluation of Rating Scales,” *Journal of the Market Research Society*, Vol. 16, No. 2, pp. 87-107, 1974.

Johnson, J. M., D. N. Bristow and K.C. Schneider, “Did You Not Understand the Question or Not? An Investigation of Negatively Worded Questions in Survey Research,” *Journal of Applied Business Research*, Vol. 20, No.1, pp. 75-86, 2004.

Jones, L.V. and L. L. Thurstone, “The Psychophysics of Semantics: An Experimental Investigation,” *Journal of Applied Psychology*, Vol. 39, No. 1, pp. 31-39, 1955.

Klein, J. G., R. Ettenson and M. D. Morris, “The Animosity Model of Foreign Product Purchase: An Empirical Test in the People's Republic of China,” *Journal of Marketing,* Vol. 62, No. 1, pp. 89-100, 1998.

Krosnick, J. A. and M. K. Berent, “The Impact of Verbal Labeling of Response Alternatives and Branching on Attitude Measurement Reliability,” Presentation: *The American Association for Public Opinion Research Annual Meeting,* 1990.

Lam, T. C. M. and Klockars, A. J., “The Influence of Labels and Positions in Rating Scales,” *Journal of Educational Measurement*, Vol. 19, pp. 312-322, 1982.

Lumpkin, James R. and William R. Darden, “Relating Television Preference Viewing to Shopping Orientations, Lifestyles, and Demographics,” *Journal of Advertising*, Vol. 11, No. 4, pp. 56-67.

Meredith, W., “Measurement Invariance, Factor Analysis and Factorial Invariance,” *Pyschometrika,* Vol. 58, pp. 525-543, 1993.

Myers, James H. and W. Gregory Warner, “Semantic Properties of Selected Evaluation Adjectives,” *Journal of Marketing Research*, Vol. 5, No. 4, pp. 409-412, 1968.

Netemeyer, R., S. Durvasula and D. R. Lichtenstein, “A Cross-National Assessment of the Reliability and Validity of the CETSCALE,” *Journal of Marketing Research*, Vol. 28, (August), pp. 320-327, 1991.

Rohrmann, B., “The Use of Verbal Scale Point Labels,“ in C. Norman & S. R. F. Job (Eds.), *7th International Congress on Noise as a Public Health Problem*: Sydney: N. E. Pty Ltd., Vol. 2, pp. 523-527, 1998.

Schwarz, N., “What Respondents Learn from Questionnaires: The Survey Interview and the Logic of Conversation: (1993 Morris Hansen Lecture),” *International Statistical Review*, Vol. 63, pp. 153-177, 1995.

Schwarz, N., *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation,* Lawrence Erlbaum, Mahwah, NJ, 1996.

Schwarz, N., C. E. Grayson, and B. Knauper, “Formal Features of Rating Scales and the Interpretation of Question Meaning,” *International Journal of Public Opinion Research,* Vol. 10, No. 2, pp. 177-183, 1998.

Schwarz, N., B. Knauper, H. J. Hipler, E. Noelle-Neumann, and L. Clark, “Numeric Values May Change the Meaning of Scale Labels,” *Public Opinion Quarterly*, Vol. 55, No. 4, pp. 570-582, 1991.

Sharma, S., T. A. Shimp, and J. Shin, “Consumer Ethnocentrism: A Test of Antecedents and Moderators,” *Journal of the Academy of Marketing Science,* Vol. 23, No. 1, pp. 26-37, 1995.

Shimp, T. and S. Sharma, "Consumer Ethnocentrism: Construction and Validation of the CETSCALE," *Journal of Marketing Research*, Vol. 24, (August), pp. 280-289, 1987.

Wegener, B., “Category-Rating and Magnitude Estimation Scaling Techniques: An Empirical Comparison,” *Social Methods and Research*, Vol. 12, pp. 31-75, 1983.

Wildt, A. R. and M. B. Mazis, “Determinants of Scale Response: Level versus Position,” *Journal of Marketing Research*, Vol. 15, No. 2, pp. 261-67, 1978.

Appendix A *

This survey is interested in your views about American-made products. Your truthful answers and your cooperation are greatly appreciated. Your responses will be kept in strictest confidence.

Section I

In this section, we are interested in your opinions about American-made products. Please read each statement carefully and on the line provided next to each statement, check (√) a box corresponding to your opinion.

	Very Strongly Agree (1)	Strongly Agree (2)	Agree (3)	Somewhat Agree (4)	Disagree (5)	Very Strongly Disagree (6)
1. American people should always buy American-made products instead of imports.	☐	☐	☐	☐	☐	☐
2. Only those products that are unavailable in the U.S. should be imported.	☐	☐	☐	☐	☐	☐
3. Buy American-made products. Keep America working.	☐	☐	☐	☐	☐	☐
4. American products, first, last, and foremost.	☐	☐	☐	☐	☐	☐
5. Purchasing foreign-made products is un-American.	☐	☐	☐	☐	☐	☐
6. It is not right to purchase foreign products, because it puts Americans out of jobs.	☐	☐	☐	☐	☐	☐
7. A real American should always buy American-made products.	☐	☐	☐	☐	☐	☐
8. We should purchase products manufactured in America instead of letting other countries get rich off us.	☐	☐	☐	☐	☐	☐
9. It is always best to purchase American products.	☐	☐	☐	☐	☐	☐
10. There should be very little trading or purchasing of goods from other countries unless out of necessity.	☐	☐	☐	☐	☐	☐
11. Americans should not buy foreign products, because this hurts American business and causes unemployment.	☐	☐	☐	☐	☐	☐
12. Curbs should be put on all imports.	☐	☐	☐	☐	☐	☐

(continue on the back)

	Very Strongly Agree (1)	Strongly Agree (2)	Agree (3)	Somewhat Agree (4)	Disagree (5)	Very Strongly Disagree (6)
13. It may cost me in the long-run but I prefer to support American products.	☐	☐	☐	☐	☐	☐
14. Foreigners should not be allowed to put their products on our markets.	☐	☐	☐	☐	☐	☐
15. Foreign products should be taxed heavily to reduce their entry into the U.S.	☐	☐	☐	☐	☐	☐
16. We should buy from foreign countries only those products that we cannot obtain within our own country.	☐	☐	☐	☐	☐	☐
17. American consumers who purchase products made in other countries are responsible for putting their fellow Americans out of work.	☐	☐	☐	☐	☐	☐

Section II

The following question is for classification purposes only. Your response will remain confidential.

1. What is your gender?	_____Male	_____Female

* This example questionnaire demonstrates the format used for the ascending/unbalanced condition. All items are from CETSCALE which was developed by Shimp and Sharma in 1987.