[Transaction]

Journal of the Korean Society
of Clothing and Textiles
Vol. 34, No. 6 (2010) p.913~927

The Influence of Likert Scale Format on Response Result, Validity, and Reliability of Scale -Using Scales Measuring Economic Shopping Orientation-

Saehee Kim†

Div. of Fashion & Beauty, Busan Kyungsang College

Received March 23, 2010; Revised April 21, 2010; Accepted May 13, 2010

Abstract

This study investigates the influence of Likert scale formats such as the number of response categories and the inclusion of a mid-point from a methodological point of view using instruments that measure a fashion-marketing-related subject. Using a self-administered questionnaire, 201 respondents rated their economic clothing shopping orientation on three formats of scales that differed only in the number of response categories (ranging from 5 to 7) from February 8 to February 12, 2010. Descriptive statistics, Spearman's rank order correlation, t-test, exploratory factor analysis, confirmatory factor analysis, Pearson's correlation, and Cronbach's alpha were used in the analysis. The results are as follows. First, three scale formats were generally suitable for use due to validity and reliability. Second, the response results varied with the number of categories and the inclusion of a mid-point, although the differences were statistically insignificant (with only a few cases that differed). Third, construct validity was more secure in scales with fewer categories, whereas convergent and discriminant validity was generally good in all scale formats. Fourth, reliability coefficients were higher in scales with more categories. Fifth, the number of categories was of greater importance to instrument design than the inclusion of a mid-point. Implications for appropriate scale designs are suggested in this study.

Key words: Likert scale, Number of response categories, Mid-point category, Validity, Reliability

I. Introduction

Many researchers and marketers have surveyed consumers to understand them for fashion marketing. Various methods are used to understand consumer psychology and behavior. Among those, the most widely used is surveying consumers through questionnaires. This is because standardized questionnaires can improve the result comparison and the ease and accuracy of responding (Chae, 2005).

Various formats of measuring instrument are used in questionnaire design. Among those, the Likert type scale, the representative type of rating scale, is the most widely used for its time and cost efficiency and ease of composition. The purpose of a rating scale is to allow respondents to express both the direction and strength of their opinion about a topic (Garland, 1991). Researchers assume that the psychological "amount" of a respondent's attitude is marked on the Likert scale and that "amount" is represented consistently on the scale (Bae, 2002).

As rating scales are among the most widely used measuring instruments in social sciences including marketing, it is therefore not surprising that a great deal of researches have been devoted to the effects of variations in rating scale format, including differences in the number of response categories (Preston & Colman, 2000). Consistency of a respondent can

†Corresponding author
E-mail: saykim@bsks.ac.kr

be easily distorted by the number of categories or by whether the neutral response category is included in a scale (Bae, 2002).

However, most of previous studies on this issue have been performed in the field of statistics, social science, education, and so on, whereas no attempts have been made so far to examine the number or format of response categories systematically in the fashion marketing research field. Researches in fashion marketing field have concentrated on measuring and analyzing consumer attitudes, opinions, or behaviors rather than issues related to scale design. However, examining the validity and reliability of scale formats is necessary because a scale is the way to understand consumer behavior. If research results on same subject are different from one another, the difference might be the result of problems with the scale that was used.

Thus, this study aims to investigate the influence of the Likert scale format on survey results from a methodological point of view using a fashion marketing related instrument. The result might suggest ways to design more appropriate scales that would contribute to better research validity and reliability.

II. Theoretical Background

1. Number of Response Categories in a Likert Scale

The optimal number of response categories for a Likert scale has not been determined definitely. Too few categories might create difficulty in investigating respondents' attitudes precisely and analyzing the data (Chae, 2005). Too many categories might yield more precise investigations into the attitude of the respondent, but they may induce fatigue and unreliable responses (Kim, 2001). Guilford (1954) and Komorita and Graham (1965) indicated that if we employ too few response categories, our scale will be a coarse instrument and much of the discriminative power that raters are capable of will be lost. Conversely, if we employ too many categories, we could grade a scale so finely that it would be beyond the raters' limited powers of discrimination (Matell & Jacoby, 1971).

There have been studies that have compared the survey results of scales with different number of response categories. The specific results of those studies have not been completely consistent.

First, researches have found that the validity and reliability of a scale increased with increasing numbers of response categories. As for validity, Loken et al. (1987) examined the criterion validity of various scales and found 11-point scales to be superior to 3- or 4-point scales. Chang (1994) found approximately similar criterion validity for 4- and 6-point scales, but higher convergent validity were found for 6-point scales.

As for reliability, Lehman and Britney's study (as cited in Boote, 1981) found test-retest reliability increased with increasing numbers of response categories. Hancock and Klockars (1991) found that items in 9-point scale correlated better than items in 5-point scale. Alwin (1997) explained this with information theory, stating that rating scales with more response categories transmit a greater amount of information and are therefore inherently more precise in their measurement.

Researches also found that both the validity and the reliability of scale were improved as the number of categories increased (Alwin, 1997; Andrews, 1984; Finn, 1972; Garner, 1960; Guilford, 1954; Komorita & Graham, 1965; Lozano et al., 2008). For example, Andrews (1984) found that as the number of categories increased by 2, 3, 4, 5, 7, 9-19, and 20, the validity was improved and error was diminished.

Second, many other researchers have confirmed that reliability, validity, and other scale qualities are largely independent of increasing numbers of response categories (Boote, 1981; Brown et al., 1991; Green & Rao, 1970; Jenkins & Taber, 1977; Komorita, 1963; Matell & Jacoby, 1971; Peabody, 1962; Preston & Colman., 2000; Schutz & Rucker, 1975). For example, Green and Rao (1970) found that information retrieval is maximized by using six or seven response categories, with little extra information being gained by increasing the number of categories beyond seven. Matell and Jacoby (1971) carried out a thorough

empirical study comparing scales with varying numbers of response categories (from 2 to 19) and concluded that as few as two response categories may be adequate in practice. They suggested that both reliability and validity are independent of the number of response categories. Schutz and Rucker (1975) found in their study of response patterns that the number of available response categories does not materially affect the cognitive structure derived from the results. Clarke (2000) found that increasing the number of scale categories from three to five reduced extreme responses, but beyond five categories there was little effect. Preston and Colman (2000) investigated several indices of reliability, validity, and discriminating power of various scale formats (ranging from 2 to 11). As a result, the 2-, 3-, and 4-point scales performed relatively poorly, and indices were significantly higher for scales with more response categories-up to about 7. The rest (8~11) performed relatively poorly. Several researchers have also addressed that the optimal number of scale categories is content specific and a function of the conditions of measurement (Cox, 1980; Friedman et al., 1981; Komorita, 1963; Matell & Jacoby, 1971; Wildt & Mazis, 1978).

On the other hand, there have been researches that suggest the optimum number of response categories. However, the results differ amongst the studies. The recommended numbers of categories were 5 (Jenkins & Taber, 1977; Lissitz & Green, 1975; McKelvie, 1978; Remmers & Ewart, 1941), 7 (Finn, 1972; Nunnally, 1978; Ramsay, 1973), 4~7 (Lozano et al., 2008), 5~ 6 (McKelvei, 1978), 7~9 (Cox, 1980), and 6~10 (Preston & Colman, 2000). By examining the previous researches, 5 to 7 response categories seem to be generally appropriate for a Likert scale. That is, they seem to result in valid and reliable outputs even though the previous results are not completely consistent.

2. Inclusion of a Mid-point Category in a Likert Scale

Another issue regarding the Likert scale format is whether a mid-point category is included in the scale. Several studies have been devoted to this issue, and there is a substantial discrepancy in their results.

First, we will address the preference for a scale with a mid-point. A mid-point is useful in the case of apathy or refusal to respond. Respondents might use a mid-point when they have no idea about the question (Chae, 2005). On the other hand, there is an opinion that mid-points can be useful in precise expression of respondents' attitudes (Komorita, 1963; Son & Chae, 2008).

Second, we will address the preference for a scale without a mid-point. There are those with the opinion that a mid-point might create the possibility of information loss (Converse & Presser, 1986; Nunnally, 1978) and respondents tend to respond more precisely after careful consideration when the mid-point is eliminated. Garland (1991) stated that market researchers would typically prefer respondents to make a definite choice rather than choose neutral or intermediate positions on a scale. A scale without a mid-point would be preferable, provided it does not affect the validity or reliability of the responses. Andrews (1984) found that a 3-point scale is less reliable than a 2-point scale, and there is no evidence that a 5-point scale is more reliable than a 4-point scale.

Third, there is also an opinion that mid-point inclusion and the validity/reliability of the scale are not related. Komorita (1963) and Jacoby and Matell (1971) found that a mid-point did not affect reliability. Saris (1988) found that a mid-point is useful only in a scale with many categories-over 7-point. Andrews (1984) and Bae (2002) compared 5- and 6-point scales and found that little difference was observed in reliability and validity. Particularly, these studies had limitations in that they were comparing only 5- and 6-point scales and it was indeterminate if the result came from mid-point inclusion or an increased number of categories.

Inclusion of a mid-point is also related to other issues like social desirability, cultural characteristics, response tendency, and meaning interpretation of respondents. First, Garland (1991) provided some evidence that social desirability bias, which can arise from respondents' desires to please the interviewer or appear helpful or to not be seen giving what they perceived to be a socially unacceptable answer, can be

minimized by eliminating the mid-point category from Likert scales. This implies the need for investigation into the mid-point inclusion in areas such as fashion or shopping that are sometimes considered to be extravagant.

Second, there might be a cross-cultural difference in response styles. According to Chen et al. (1995), East Asian students such as those from Japan and China were more likely as compared to their North American counterparts to use the mid-point on Likert scales. That is, individualism seemed to be negatively related to mid-point selection.

Third, Matell and Jacoby (1972) demonstrated that as the number of response categories increased respondents' selection of the mid-point category decreased. Matell and Jacoby (1972) advised either eliminating the mid-point or using a scale with many categories when attempting to minimize the selection of the mid-point category.

Fourth, there have been researchers who posited that the mid-point category might be interpreted differently by different respondents, and sometimes differently than intended (Hofacker, 1984). Stone (2004) stated that a middle response choice of "3" can reflect a decision not to prefer either end, a lack of information to make a decision, or an unwillingness to commit to a definite response.

Kulas et al. (2008) examined whether the mid-point option in Likert scales indicates a moderate stance on a question or whether it is a "dumping ground" for unsure or non-applicable responses. Specifically, they identified mid-point dysfunction. In addition, they found that respondents used the mid-point as a non-applicable proxy, even under implicit instructions to "skip if you do not know." These researches imply careful inclusion of the mid-point in a Likert scale.

III. Research Purpose and Research Problems

After reviewing previous researches, it was found that though there have been many of researches on scale format, no attempt to consider the issue was devoted to the field of fashion marketing research. As previous authors have asserted, the optimum number of categories or other related results depends on the research subject and contents. Therefore, this study aims to investigate the effect of a scale format that employs a fashion-marketing-related scale. In particular, this study will look into the effects of both the number of categories and the inclusion of a mid-point, which makes it different from most of the previous studies investigated either of the two issues. That is, this study will provide a thorough assessment of descriptive results, validity, and reliability of scores from rating scales that vary in number of response categories and mid-point inclusion. The research problems are as follows.

Research Problem 1. Are there differences in response results according to the number of categories and the inclusion of a mid-point in a Likert scale?

Research Problem 2. Is there an influence on the validity of scales according to the number of categories and the inclusion of a mid-point in a Likert scale?

Research Problem 3. Is there an influence on the reliability of scales according to the number of categories and the inclusion of a mid-point in the Likert scale?

IV. Methods

1. Scale Design

This study aims to investigate the influence of scale formats on response results, validity, and reliability of scales, employing instruments measuring fashion consumer behavior. Thus, a concept which is frequently surveyed in the field of fashion marketing research was needed to be selected for scale design. Numerous researches in this field have employed the concept clothing shopping orientation as a major variable to identify consumer groups or to investigate consumer characteristics. Therefore, a scale for clothing shopping orientation was considered to be appropriate for this study. There is a standardized scale for clothing shopping orientation developed by Kim and Rhee (2004), which is composed of 31 items (10 items

measuring economic shopping orientation, 15 items measuring hedonic shopping orientation, and 6 items measuring convenient shopping orientation). As too many items might be boring to respondents who have to respond to repeated items in this study, ten items measuring economic shopping orientation were chosen for this study.

In designing the scale formats, this study selected 5-, 6-, and 7-point Likert scales in consideration of the fact that more than 74% of scales used in domestic research fields are 5-, 6-, and 7-point scales (Kim, 2001). This was also because that previous researches generally showed that the optimal number of categories ranged from 5~7.

To investigate differences in response results, validity, and reliability by the number of categories, a 5-point Likert scale (1: strongly disagree, 2: disagree, 3: neutral, 4: agree, 5: strongly agree) and a 7-point Likert scale (1: strongly disagree, 2: disagree, 3: slightly disagree, 4: neutral, 5: slightly agree, 6: agree, 7: strongly agree) were composed. To investigate the influence of the inclusion of a mid-point, a 6-point Likert scale (1: strongly disagree, 2: disagree, 3: slightly disagree, 4: slightly agree, 5: agree, 6: strongly agree) was composed to compare with scales including a mid-point. In order to prevent distorting the interpretation of a mid-point's effect with the effect of the number of categories, the 6-point scale (without a mid-point) was compared with both the smaller and the larger scales that included a mid-point (5- and 7-point scales).

Each version was composed of 10 items measuring economic shopping orientation, and the contents of the items in each version were identical. Each respondent filled out a questionnaire that consisted of three versions of rating scales.

Response categories were marked using verbal expressions instead of numeric expressions. This was because numeric measurement might not reflect the natural psychological process (Windchil & Wells, 1996) and the mid-point marked in numeric expression can be interpreted with various meanings (Hofacker, 1984; Kulas et al., 2008; Stone, 2004).

Scale versions were presented in the following order: 5-, 6-, and 7-point. To avoid memory effects that might arise from responding to identical item contents repeatedly, presentation order of items in each version was randomized. In addition, other question items measuring demographic characteristics and clothing selection criteria were inserted between each version, although those questions were not "actual" research variables in this study. Furthermore, each version was located on a separate page and respondents were forbidden to review a page they had already responded to eliminate the possibility of response regulating for consistency.

2. Data Collection

Questionnaires were administered to 210 respondents in Korea from February 8 to February 12, 2010. Internet survey system was used to control errors. That is, the questionnaire constitutes identical items only different in the number of response categories that respondents might review their own prior responses for response consistency. Questionnaires in document form can not control this kind of error. Thus, internet survey was employed which could design the response system blocking page return. Convenience sampling was carried out to obtain even data in the distribution of gender, age, and occupation. A total of 201 out of 210 questionnaires administered were analyzed.

Among the respondents, 51.2% were males and 48.8% were females; 56.2% were students, 17.4% were office workers, 8.5% were housekeepers, 5.5% were professionals, 4.5% were sales/service workers, 4.0% were self-employed, 3.5% were part-time workers, and 0.5% was unemployed. The ages were distributed evenly from 15 to 58. As for monthly expenditure for clothing, 17.9% were "less than 50,000 won", 32.8% were "50,000 won~less than 100,000 won", 27.9% were "100,000 won~less than 200,000 won", 16.4% were "200,000 won~less than 300,000 won", 3.0% were "300,000 won~less than 500,000 won", and 2.0% were "more than 500,000 won".

3. Data Analysis

To design the analysis process, this study referred to previous researches (Bae, 2002; Dawes, 2008; Friedman et al., 1981; Preston & Colman, 2000), and

selectively chose appropriate statistical methods for each research problems. First, to analyze the response results according to scale format, descriptive statistics, Spearman's rank order correlation, and t-tests were used. Second, validity was tested through construct validity and convergent/discriminant validity. Exploratory factor analysis and confirmatory factor analysis were used for construct validity, and Pearson's correlation coefficients were used for convergent/discriminant validity. Third, to test reliability, Cronbach's alpha was used. As for statistics programs, AMOS was used for confirmatory factor analysis, and SPSS was used for the rest of the analysis.

V. Results

1. Response Results according to the Number of Categories and Mid-point Inclusion

To investigate the response results, descriptive statistics, mean rank, and mean difference by gender were analyzed. <Table 1> shows the results for each item in the three scale formats.

1) Response Results according to the Number of Categories

To investigate the differences in response results according to the number of categories, the 5- and 7-point scales were compared. The 6-point scale excludes the mid-point, so the response results of the 6-point scale might be affected by that fact. Thus, the 6-point scale was used only to investigate the mid-point effect.

The median and category length of each scale are different such that comparing the means or standard deviations of the two scales is not meaningful. Thus, the two scales were compared by general distribution of data and rank of mean score.

All items in both scales showed a mean score over the value of the median (5-point: 3, 7-point: 4). The ranks of mean scores were slightly different between the two scales. However, the difference was not significant as the result of Spearman's rank order correlation. This implies that response results are slightly different in mean score according to the number of categories, although the difference is not significant.

All items except for one item in 5-point scale showed a negative skewness score; overall, the responses tended to lean towards the right of the scale, i.e., agreement. Specifically, responses to eight items in the 7-point scale generally more tended to lean toward the right (i.e., agreement) than corresponding items in the 5-point scale. These results imply that if researchers want a more definite expression (agreement or disagreement) of opinion, attitude, or any behavior, items with many categories might be more favorable and effective.

Next, as a result of the t-test to find if the scales result in identical differences between groups, significant differences were found in items 1, 6, 8, and 10 in the 5-point scale and items 5, 6, 8, and 10 in the 7-point scale. That is, although the items that the two groups showed significant differences in were mostly identical, there were different results for two items (item 1 and item 5). This implies that according to the number of categories in a scale, the survey results that reveal group differences are not identical.

2) Response Results according to Mid-point Inclusion

To investigate the difference in response results resulting from mid-point inclusion, the 6-point scale was compared with the 5-point and 7-point scales. All items in the three scales showed mean scores over the median (5-point: 3, 6-point: 3.5, 7-point: 4). The ranks of mean scores were slightly different among the three scales. However, the difference was not significant as the result of Spearman's rank order correlation. This implies that response results differ slightly in mean score resulting from the inclusion of a mid-point, although the difference is not significant.

All items in the 6-point scale showed negative skewness scores as did those in both the 5- and 7-point scales. That is, responses tended to lean toward the right of the scale, i.e., agreement. After comparing the skewness scores of the three scales, seven items in the 6-point scale showed the most negative scores. This implies that eliminating the mid-point results in more right (agreement) oriented responses. In addition, the percentage of mid-point selection

Table 1. Mean, standard deviation, mean rank, skewness, and independent sample t-tests for ten items by scale format

Item[a]	5-point scale						6-point scale						7-point scale					
	Total			Group Difference			Total			Group Difference			Total			Group Difference		
	Mean (S.D.)	Rank	Skew-ness	Male	Female	t-test	Total	Rank	Skew-ness	Male	Female	t-test	Total	Rank	Skew-ness	Male	Female	t-test
I1	3.54 (0.84)	3	-.49	3.43 (.89)	3.66 (.77)	-2.00*	4.18 (1.02)	3	-.80	4.14 (1.04)	4.22 (1.00)	-0.62	4.85 (1.27)	3	-.64	4.81 (1.21)	4.90 (1.34)	-0.51
I2	3.29 (0.90)	9	-.18	3.30 (.91)	3.28 (.89)	0.2	3.96 (1.03)	10	-.41	3.95 (1.06)	3.97 (1.01)	-0.12	4.53 (1.25)	10	-.29	4.53 (1.20)	4.52 (1.31)	0.08
I3	3.26 (1.00)	10	-.18	3.18 (1.06)	3.35 (.92)	-1.16	3.98 (1.11)	8	-.36	3.93 (1.18)	4.03 (1.03)	-0.63	4.57 (1.28)	8	-.34	4.64 (1.28)	4.49 (1.29)	0.83
I4	3.33 (0.92)	7	-.10	3.37 (1.03)	3.30 (.80)	0.56	3.99 (0.98)	7	-.39	3.96 (1.00)	4.02 (.97)	-0.43	4.63 (1.19)	7	-.16	4.56 (1.23)	4.69 (1.14)	-0.78
I5	3.36 (0.87)	5	-.22	3.25 (.83)	3.47 (.91)	-1.77	3.98 (1.13)	9	-.39	3.78 (1.13)	4.18 (1.11)	-2.58**	4.69 (1.33)	6	-.25	4.46 (1.30)	4.93 (1.33)	-2.55**
I6	3.31 (0.88)	8	.05	3.50 (.93)	3.11 (.78)	3.23***	4.11 (1.04)	4	-.09	4.23 (1.12)	3.98 (.93)	1.74	4.76 (1.27)	5	-.29	5.01 (1.21)	4.49 (1.29)	2.94***
I7	3.36 (0.79)	6	-.49	3.31 (.79)	3.41 (.80)	-0.87	4.02 (1.09)	6	-.57	3.92 (1.07)	4.13 (1.10)	-1.37	4.56 (1.32)	9	-.35	4.55 (1.30)	4.56 (1.35)	-0.04
I8	3.61 (0.82)	2	-.46	3.40 (.89)	3.83 (.69)	-3.80***	4.37 (1.10)	1	-.48	4.11 (1.15)	4.64 (.99)	-3.55***	4.99 (1.30)	2	-.27	4.64 (1.27)	5.36 (1.24)	-4.04***
I9	3.62 (0.74)	1	-.47	3.59 (.77)	3.65 (.70)	-0.58	4.36 (.98)	2	-.62	4.31 (1.05)	4.42 (.90)	-0.78	5.12 (1.15)	1	-.74	5.15 (1.22)	5.10 (1.09)	0.27
I10	3.42 (0.84)	4	-.21	3.26 (.80)	3.59 (.85)	-2.83**	4.04 (1.10)	5	-.11	3.78 (1.04)	4.33 (1.09)	-3.66***	4.81 (1.33)	4	-.36	4.53 (1.29)	5.09 (1.32)	-3.03***

*$p \leq .05$, **$p \leq .01$, ***$p \leq .001$

[a]Items were marked numerically due to the space limitation. Specific contents of items can be found in <Table 2>.

was higher in the 5-point scale than in the 7-point scale except for one item. This result is in line with previous research (Matell & Jacoby, 1972).

Next, as the result of the t-test to find the difference between male and female groups, significant differences were found in items 5, 8, and 10 in the 6-point scale. This is different from the results in the 5- (items 1, 6, 8, and 10) and 7-point (items 5, 6, 8, and 10) scales. The 5- and 6-point scales showed different results for three items; one item was different between the 7- and 6-point scales. This implies that the survey results for revealing group differences are not identical according to the inclusion of a mid-point, and the difference is greater between the 5- and 6-point scales.

2. Scale Validity according to the Number of Categories and Mid-point Inclusion

Scale validity was assessed in several ways. To examine the validity of scales, exploratory factor analysis was carried out first <Table 2>, and then confirmatory factor analysis was performed afterwards (Fig. 1)-(Fig. 2), (Table 3). Pearson's correlation coefficients between items were also analyzed to examine the convergent and discriminant validity of the scales (Table 4).

1) Scale Validity according to the Number of Categories

Exploratory factor analysis resulted in three factors respectively in both the 5- and 7-point scales. However, the loaded items were a little bit different. That is, item 7 ("I like stores with convenient facilities and recreational spaces") loaded on the "rational" factor in the 5-point scale, whereas it loaded on the "low price and convenience oriented" factor in the 7-point scale. In addition, item 4 ("I carefully plan what to purchase before shopping") loaded on the "planning" factor in the 5-point scale whereas it loaded on the "rational" factor in the 7-point scale. On the other hand, item 7 and item 4 showed substantial factor loading on other factors as well. That is, the conceptual structures of scales were different between the two scales with a different number of categories. Furthermore, this implies that the grouping of items is less definite in the scale with many categories, and thereby the data reduction effect by exploratory factor analysis is also less effective. Factors accounted for 71% of variance in the 7-point scale and 60% in the 5-point scale, respectively. That is, the more categories there were, the more the variance was explained.

To investigate the construct validity of the two scales more in depth, confirmatory factor analysis was carried out following the results of exploratory factor analysis (Fig. 1)-(Fig. 2), (Table 3). To investigate how the observed variables represent the latent variables, each observed variable loaded on only one latent variable on which they had loaded in the exploratory factor analysis. Data from the 5-point scale were analyzed using Model 1 <Fig. 1>, whereas data from the 7-point scale were analyzed using both Model 1 and 2 (Fig. 1)-(Fig. 2). This was done to compare the confirmatory factor analysis results (e.g., fit indices and coefficients) of scale formats in an identical model (Model 1), as well as in respective original model (e. g., 5-point scale: Model 1, 7-point scale: Model 2). In this case, Model 1 was selected as "identical model" for comparison because two scales (e. g., 5- and 6-point scale) among three scales showed identical factor construct which corresponds to Model 1. The results are shown in <Table 3>.

Fit indices to models were generally good or moderate. The 5-point scale showed slightly better fit indices than the 7-point scale. In particular, the GFIs of the 7-point scale to both Model 1 and Model 2 were slightly lower than .90, which is usually an acceptable fit index. All of the coefficients were significant in both of the two scales. Generally, the coefficients were slightly higher in the 7-point scale. Thus, a scale with a small number of categories seems to be more appropriate for measuring conceptual constructs more validly, whereas a scale with more categories seems to be more appropriate for observed variables to account for more of latent variables.

On the other hand, convergent validity and discriminant validity were assessed by examining the correlations of items. <Table 4> shows the intercorrelations between items from the three scale formats.

A specific item was assumed to show convergent

Table 2. Exploratory factor analysis of each scale format

Factors and Items	Factor loading
5-point scale	
Factor 1: Rational	
I10: I browse styles and prices in many stores before purchasing	.82
I8: I shop around many stores before purchasing to find items with good quality and design	.82
I5: After shopping, I evaluate purchased item to determine if I shopped reasonably	.61
I7: I like stores with convenient facilities and recreational spaces	.45
Eigenvalue: 3.56% Variance Explained: 35.63%	
Factor 2: Planning	
I6: I purchase only necessary items without impulse shopping	.73
I3: I allocate budget for clothing in advance of shopping	.68
I9: I try to select the most appropriate item within the budget	.68
I4: I carefully plan what to purchase before shopping	.63
Eigenvalue: 1.38% Variance Explained: 13.77%	
Factor 3: Low-price oriented	
I2: I like shopping at outlets or off-price stores	.87
I1: I usually visit stores selling low priced or bargain items	.83
Eigenvalue: 1.08% Variance Explained: 10.75%	
6-point scale	
Factor 1: Planning	
I3: I allocate budget for clothing in advance of shopping	.81
I9: I try to select the most appropriate item within the budget	.81
I6: I purchase only necessary items without impulse shopping	.68
I4: I carefully plan what to purchase before shopping	.63
Eigenvalue: 4.17% Variance Explained: 41.66%	
Factor 2: Rational	
I8: I shop around many stores before purchasing to find items with good quality and design	.86
I10: I browse styles and prices in many stores before purchasing	.79
I7: I like stores with convenient facilities and recreational spaces	.59
I5: After shopping, I evaluate purchased item to determine if I shopped reasonably	.53
Eigenvalue: 1.36% Variance Explained: 13.57%	
Factor 3: Low-price oriented	
I1: I usually visit stores selling low priced or bargain items	.87
I2: I like shopping at outlets or off-price stores	.86
Eigenvalue: 1.12% Variance Explained: 11.21%	
7-point scale	
Factor 1: Rational	
I8: I shop around many stores before purchasing to find items with good quality and design	.84
I10: I browse styles and prices in many stores before purchasing	.81
I5: After shopping, I evaluate purchased item to determine if I shopped reasonably	.78
I4: I carefully plan what to purchase before shopping	.66
Eigenvalue: 4.82% Variance Explained: 48.23%	
Factor 2: Planning	
I6: I purchase only necessary items without impulse shopping	.85
I9: I try to select the most appropriate item within the budget	.75
I3: I allocate budget for clothing in advance of shopping	.68
Eigenvalue: 1.23% Variance Explained: 12.33%	
Factor 3: Low-price and convenience oriented	
I2: I like shopping at outlets or off-price stores	.88
I1: I usually visit stores selling low priced or bargain items	.83
I7: I like stores with convenient facilities and recreational spaces	.45
Eigenvalue: 1.12% Variance Explained: 11.21%	

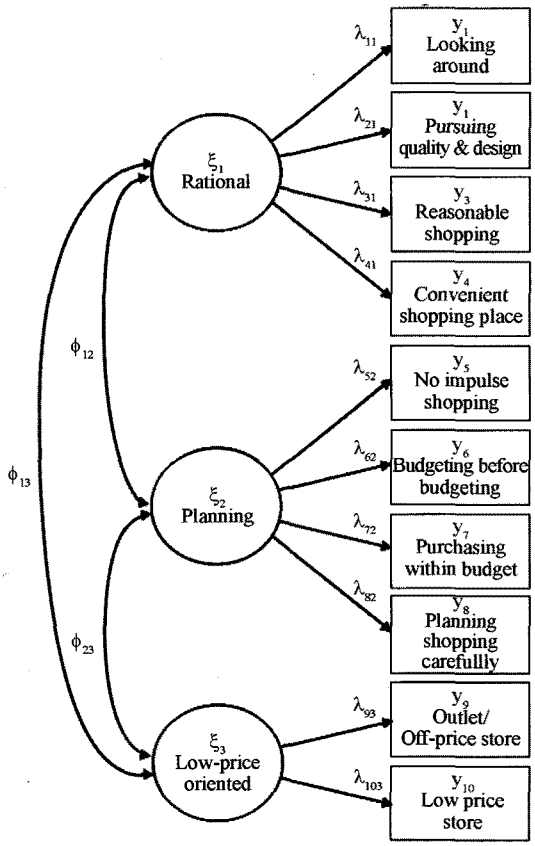

Fig. 1. Structural model of economic shopping orientation in 5- and 6-point scales (Model 1).

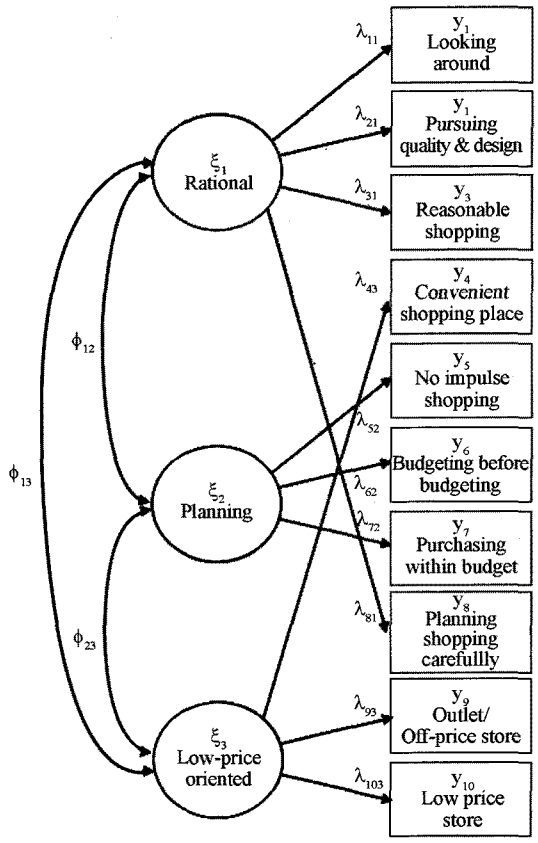

Fig. 2. Structural model of economic shopping orientation in 7-point scale (Model 2).

validity to the extent that it is correlated with other items measuring the same content in other scale formats. In addition, a specific item was assumed to show discriminant validity to the extent that it is not correlated with other items measuring other content in other scale formats. As a result, the convergent validity of both scales was confirmed, because all of the correlations between two corresponding items that measured the same content in the 5- and 7-point scales were significant and higher than for other correlation coefficients. Furthermore, the correlations between items measuring different content were generally low such that the discriminant validity of the scales was confirmed. Occasionally, relatively high correlations between items measuring different content were observed, but these were items loading on the same factors. Thus, both of the two scales exhibited convergent validity and discriminant validity regardless of the number of categories.

In this section, validity of scales with different number of categories was investigated through several statistical methods. Putting together the results, this study's results seem to support the opinion that the relationship between the number of categories and a scale's validity is independent (Boote, 1981; Brown et al., 1991; Green & Rao, 1970; Jenkins & Taber, 1977; Komorita, 1963; Matell & Jacoby, 1971; Peabody, 1962; Preston & Colman., 2000; Schutz & Rucker, 1975).

2) Scale Validity according to Mid-point Inclusion

As a result of exploratory factor analysis, three factors were extracted in the 6-point scale as with the 5-

Table 3. Model fits and coefficients of confirmatory factor analysis by scale format

Fit & Coefficients \ Scale format		5-point scale	6-point scale	7-point scale	
				Transformed Model[a]	Original Model[b]
Fit Indices	GFI	.93	.90	.87	.87
	AIC	120.92	158.60	182.90	213.49
	Chi-square	74.92*** (df=32)	112.60*** (df=32)	136.90*** (df=32)	167.49*** (df=32)
	RMR	.04	.08	.12	.14
Coefficient	λ_{11}	.77***	.82***	.92***	.89***
	λ_{21}	.76***	.72***	.80***	.80***
	λ_{31}	.52***	.65***	.67***	.68***
	λ_{41}	.40***	.48***	.45***	.47*** (λ_{43})[c]
Coefficient	λ_{52}	.35***	.46***	.53***	.62***
	λ_{62}	.68***	.83***	.81***	.74***
	λ_{72}	.73***	.74***	.76***	.88***
	λ_{82}	.64***	.74***	.86***	.77*** (λ_{81})[d]
	λ_{93}	.66***	.78***	.72***	.75***
	λ_{103}	.94***	.85***	.90***	.84***
	φ_{12}	.64***	.65***	.75***	.69***
	φ_{13}	.48***	.51***	.56***	.60***
	φ_{23}	.42**	.44***	.54***	.59***

**$p \leq .01$, ***$p \leq .001$
[a]Data analyzed based on Model 1 to compare with the results of 5- and 6-point scales.
[b]Data analyzed based on Model 2 (Original model for 7-point scale).
[c, d]Coefficient labels specific to Model 2.

and 7-point scales. The loaded items for each factor were identical to those of the 5-point scale. The explained variance was 66% in the 6-point scale, which was between the 5- and 7-point scales.

Confirmatory factor analysis was carried out following the results of the exploratory factor analysis (Table 3). Data were analyzed using Model 1 (Fig. 1). Fit indices were relatively acceptable, and those indices were in the middle of the 5- and 7-point scales. Coefficients of factor loading and correlations between latent variables were also in the middle of the 5- and 7-point scales. Thus, the number of categories seems to be more influential than the inclusion of a mid-point for establishing the construct validity of scales.

On the other hand, the results of the correlation analysis between items showed that the correlations between the 5- and 6-point items measuring the same content and those of the 7- and 6-point items measuring the same content were highest among all correlations (Table 4). Thus, the convergent validity was confirmed. In addition, the correlations between items measuring different content were generally low such that the discriminant validity of the scales was confirmed. Occasionally, relatively high correlations between items measuring different content were observed, but these were items loading on the same factors. Therefore, the three scales exhibited convergent validity and discriminant validity regardless of mid-point inclusion.

3. Scale Reliability according to the Number of Categories and Mid-point Inclusion

To establish consistency over the items within each scale format and each factor, the ratings derived from each scale were evaluated for reliability using Cronbach's alpha. <Table 5> shows the alpha coefficients for the internal consistency reliability analysis. The overall alpha coefficients of the three scale formats were all relatively high (above .79). The alpha coeffi-

Table 4. Intercorrelations between items

		Correlation coefficients between scales with difference in number of categories									
		5-point scale									
		I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
7-point scale	I1	.67**	.48**	.26**	.25**	.16*	.16*	.27**	.31**	.34**	.34**
	I2	.56**	.73**	.20**	.26**	.18*	.10	.27**	.24**	.28**	.22**
	I3	.27**	.26**	.72**	.59**	.33**	.33**	.28**	.28**	.53**	.37**
	I4	.34**	.24**	.55**	.65**	.45**	.34**	.30**	.49**	.50**	.52**
	I5	.27**	.25**	.40**	.35**	.71**	.17*	.32**	.49**	.41**	.47**
	I6	.17*	.13	.22**	.29**	.14	.70**	.23**	.14*	.40**	.16*
	I7	.25**	.29**	.23**	.20**	.22**	.10	.72**	.29**	.30**	.25**
	I8	.37**	.27**	.30**	.36**	.35**	.04	.32**	.77**	.33**	.63**
	I9	.31**	.22**	.48**	.44**	.28**	.40**	.31**	.33**	.72**	.34**
	I10	.41**	.30**	.38**	.40**	.43**	.09	.35**	.68**	.44**	.71**
		Correlation coefficients between scales with difference in mid-point inclusion									
		6-point scale									
		I1	I2	I3	I4	I5	I6	I7	I8	I9	I10
5-point scale	I1	.73**	.62**	.20**	.36**	.29**	.11**	.24**	.36**	.28**	.36**
	I2	.53**	.70**	.16*	.26**	.27**	.06	.27**	.22**	.21**	.27**
	I3	.18*	.18**	.70**	.47**	.39**	.28**	.15*	.19**	.55**	.29**
	I4	.23**	.22**	.54**	.60**	.39**	.26**	.15*	.30**	.50**	.35**
	I5	.15*	.12	.38**	.31**	.69**	.11	.19**	.27**	.35**	.40**
	I6	.15*	.10	.28**	.23**	.23**	.67**	.05	-.02	.29**	.10
	I7	.28**	.22**	.21**	.23**	.25**	.10	.77**	.31**	.24**	.25**
	I8	.21**	.29**	.27**	.33**	.44**	.07	.31**	.70**	.29**	.67**
	I9	.25**	.26**	.51**	.45**	.39**	.37**	.23**	.31**	.66**	.36**
	I10	.23**	.26**	.36**	.45**	.46**	.11	.23**	.57**	.29**	.75**
7-point scale	I1	.76**	.57**	.27**	.36**	.26**	.29**	.24**	.37**	.32**	.37**
	I2	.65**	.77**	.21**	.26**	.29**	.18*	.32**	.24**	.28**	.31**
	I3	.29**	.24**	.81**	.61**	.48**	.39**	.25**	.22**	.67**	.40**
	I4	.29**	.23**	.62**	.64**	.49**	.39**	.24**	.42**	.52**	.55**
	I5	.21**	.18*	.53**	.40**	.82**	.14*	.37**	.43**	.47**	.54**
	I6	.27**	.19**	.29**	.30**	.24**	.76**	.18*	.05	.43**	.20**
	I7	.29**	.22**	.30**	.27**	.29**	.28**	.79**	.29**	.28**	.32**
	I8	.22**	.26**	.32**	.36**	.44**	.08	.34**	.77**	.31**	.73**
	I9	.28**	.23**	.53**	.38**	.34**	.55**	.24**	.31**	.79**	.39**
	I10	.33**	.30**	.46**	.46**	.51**	.19**	.38**	.66**	.41**	.84**

*$p \leq .05$, **$p \leq .01$

cients of factors were slightly different according to scale format and factor.

1) Scale Reliability according to the Number of Categories

Overall, the alpha coefficients of and the factors extracted from the 7-point scale were generally higher as compared to those from the 5-point scale. That is, reliability increased with increasing numbers of response categories. This is in line with the results of Alwin (1997), Andrews (1984), Hancock and Klockars (1991), and others.

Table 5. Reliability of scales and factors

Scale / Factor	5-point scale	6-point scale	7-point scale		
			Assumed factors[a]	Original factors[b]	
	α	α	α	Factor	α
Rational	.69	.75	.80	Rational	.86
Planning	.69	.78	.83	Planning	.78
Low price oriented	.76	.80	.79	Low price and convenience oriented	.71
All items	.79	.84	.88		

[a]To compare alpha coefficients among the three scale formats for identical factors, items were grouped temporarily according to factor construct in the 5- and 6-point scales.
[b]Alpha coefficients were calculated based on the original factor construct of the 7-point scale.

2) Scale Reliability according to Mid-point Inclusion

The alpha coefficient of the 6-point scale without a mid-point was in the middle of the 5- and 7-point scales. That is, the inclusion of a mid-point seems to have little influence on scale reliability. On the other hand, the alpha coefficients of each factor in the 6-point scale were also generally in the middle of the 5- and 7-point scales, although there was one exceptional case. However, the difference was small enough that it might be concluded that the reliability of scale is influenced from an increase in category number rather than mid-point inclusion. That is, reliability is independent from mid-point inclusion. This confirms the findings of Komorita (1963) and Jacoby and Matell (1971).

VI. Conclusions and Implications

This study investigated the influence of the Likert scale format on response results, validity, and reliability from a methodological point of view using instruments measuring economic shopping orientation. As for the results, construct validity, convergent validity, discriminant validity, and reliability of scales were generally good. Thus, roughly speaking, it is reasonable to use 5-, 6-, and 7-point scales in studies.

However, there were apparently several differences in research results, validity, and reliability of scales according to number of categories and mid-point inclusion. First, mean score rank for each scale, items resulting in significant differences between gender groups, and conceptual constructs (results of factor analysis) differed according to the number of categories. In addition, mean score rank for each scale and items resulting in significant difference between gender groups were different according to whether a mid-point was included or not. That is, though the differences were not statistically significant or the different cases were not very many, there are differences in the response results according to the number of categories and the inclusion of mid-point.

Second, as the number of categories increased, the responses tended to lean towards the right (agreement), and respondents tended not to select the mid-point. In addition, the conceptual construct of a scale with many categories tended to be different from that of other scales with fewer categories, and the construct validity was slightly poor compared to other scales with fewer categories. However, the variance was explained more, and observed variables were more likely to explain latent variables. Furthermore, reliability increased as the number of categories increased.

Third, when a scale included a mid-point, responses to the scale tended to lean toward the right (agreement) less. However, unlike the number of categories which influenced most of the statistical results, mid-point inclusion in a scale did not seem to influence results such as construct validity or reliability.

Putting together the results, it might be concluded that it is hard to determine which is better among scale formats varying in category numbers and mid-point inclusion. That is, each scale format has merits and defects in various statistical properties such as descriptive statistics, validity, and reliability. Thus, researchers might need to depend on empirical set-

tings or the objectives of the survey. Researchers and practitioners may need to perform a trade-off between reliability and validity in light of prevailing circumstances.

On the basis of the results and conclusion, several implications can be suggested for survey design. Specific suggestions are as follows.

First, if researchers prefer more definite non-neutral rather than neutral responses, exclusion of a mid-point or if included, many response categories would be preferable. Second, if the aim of the research is to measure overall conceptual construct with greater validity and specificity, a smaller number of categories might be effective. However, if the researcher wants a scale that accounts for more variance, a scale with more categories might be recommended. Third, if researchers need a more reliable scale that is internally consistent, a scale with more categories might be preferable. Fourth, the number of categories seemed to be a more crucial element when considering instrument design than the inclusion of a mid-point.

Most of the previous studies investigated the influence of either the number of categories or the inclusion of mid-point; however, this study investigated and compared the influence of both. Furthermore, the results suggested ways to design more appropriate scales that can improve the reliability and validity of measurements. This study was an initial and exploratory investigation of the influence of scale formats in the field of fashion marketing survey, and it was intended to stimulate and evoke interest in scale format in this area.

The question of whether the results of this study can be generalized to different topics, different scale formats, and different subject populations remains to be addressed. Thus, further studies are needed to explore the influence of scale format using instruments that measure other consumer behaviors or psychologies in the fashion marketing survey field.

References

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research, 25*(3), 318–340.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly, 48*, 409–442.

Bae, J. (2002). *An analysis of the validity and reliability about the utility of the neutral point response category on a likert scale*. Unpublished master's thesis, Ewha Womans University, Seoul.

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

Brown, G., Wilding, R. E. II., & Coulter, R. L. (1991). Customer evaluation of retail salespeople using the SOCO scale: A replication, extension, and application. *Journal of the Academy of Marketing Science, 9*, 347–351.

Chae, S. (2005). *Social research method and analysis* (3rd ed.). Seoul: B&M Books.

Chang, L. (1994). A psychometric evaluation of 4-point and 6-point likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(3), 205–215.

Chen, C., Lee, S., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6*(3), 170–175.

Clarke, III. I. (2000). Extreme response style in cross-cultural research: An empirical investigation. *Journal of Social Behavior & Personality, 15*(1), 137–152.

Converse, J., & Presser, S. (1986). *Survey questions*. Beverly Hills, CA: Sage Publications.

Cox, E. P. III. (1980). The optimal number of response alternatives for a scales: A review. *Journal of Marketing Research, 17*, 407–422.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point, and 10-point scales. *International Journal of Market Research, 50*(1), 61–77.

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

Friedman, H. H., Wilamowsky, Y., & Friedman, L. W. (1981). A comparison of balanced and unbalanced rating scales. *The Mid-Atlantic Journal of Business, 19*(2), 1–7.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? *Marketing Bulletine, 2*, 66–70.

Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *The Psychological Review, 67*, 342–352.

Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery: How many scales and response categories to use? *Journal of Marketing, 34*, 33–39.

Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

Hancock, G. R., & Klockars, A. J. (1991). The effect of scale manipulations on validity: Targeting frequency rating scales for anticipated performance levels. *Applied Ergonomics, 22*, 147–154.

Hofacker, C. F. (1984). Categorical judgment scaling with ordinal assumptions. *Multivariate Behavioral Research, 19*, 91–106.

Jacoby, J., & Matell, M. S. (1971). Three-point scales always good enough. *Journal of Marketing Research, 8*, 495–500.

Jenkins, G., & Taber, T. (1977). A Monte Carlo study of factors affecting three indices of composite scale reliability. *Journal of Applied Psychology, 62*, 392–398.

Kim, N. (2001). *A comparative analysis of items election methods for developing the likert scale*. Unpublished master's thesis, Yonsei University, Seoul.

Kim, S., & Rhee, E. (2004). Development of measurement scale for clothing shopping orientation (Part I). *Journal of the Korean Society of Clothing and Textiles, 28*(9/10), 1253–1264.

Komorita, S. S. (1963). Attitude content, intensity, and the neutral point on a likert scale. *Journal of Social Psychology, 61*(December), 327–334.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 25*, 987–995.

Kulas, J. T., Stachowski, A. A., & Haynes, B. A. (2008). Middle response functioning in likert-responses to personality items. *Journal of Business and Psychology, 22*(3), 251–259.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60*, 10–13.

Loken, B., Pirie, P., Virnig, K. A., Hinkle, R. L., & Salmon, C. T. (1987). The use of 0-10 scales in telephone surveys. *Journal of the Market Research Society, 29*(3), 353–362.

Lozano, L. M., Garcia-Cueto, E., & Muniz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *European Journal of Research Methods for the Behavioral and Social Sciences, 4*(2), 73–79.

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for likert scales items? Study I: Reliability and validity. *Educational and Psychological Measurement, 31*, 657–674.

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*(6), 506–509.

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

Nunnally, J. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Peabody, D. (1962). Two components in bipolar scales: Direction and extremeness. *Psychological Review, 69*, 65–73.

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*, 1–15.

Ramsay, J. O. (1973). The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrika, 38*, 513–533.

Remmers, H. H., & Ewart, E. (1941). Reliability of multiple-choice measuring instruments as a function of the Spearman-Brown prophecy formula. *Journal of Educational Psychology, 32*, 61–66.

Saris, W. E. (1988). *Variation in response functions: A source of measurement error in attitude research*. Amsterdam: Sociometric Research Foundation.

Schutz, H. G., & Rucker, M. H. (1975). A comparison of variable configurations across scale lengths: An empirical study. *Educational and Psychological Measurement, 35*, 319–324.

Son, Y., & Chae, S. (2008). *Systematic questionnaire design* (2nd ed.). Seoul: B&M Books.

Stone, M. H. (2004). Substantive scale construction. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to rasch measurement* (pp. 201–225). Maple Grove, MN: JAM Press.

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

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364.