Journal of Media Critiques [JMC]
doi: 10.17349/jmc117103 P-ISSN: 2056-9785 E-ISSN: 2056 9793
http://www.mediacritiques.net jmc@mediacritiques.net

RESPONSE OPTIONS FOR SCALES: DOES IT MATTER WHAT WORDS YOU USE?

NARISSRA MARIA PUNYANUNT-CARTER*

ABSTRACT

Scales are a commonly used method for measuring variables in media and communication research. As researchers use scales designed in previous studies, they may find it necessary to alter the scale to fit the subjects or research conditions. The current study sought to explore the effects of these alterations by comparing scores on two scales when response options were changed. When response options were changed, significant differences occurred in the way subjects scored on the scales. This suggests that such alterations need to be carefully considered when designing a study. Findings from this study are especially beneficial for researchers collecting data for online instruments and organizational advancements.

Keywords: media, measurement, scale, mass communication.

INTRODUCTION

Scales are an integral part of social science research. Scales or indexes allow scholars to ordinally rank cases of a given variable. Kerlinger defines a scale "as a set of symbols or numerals so constructed that the symbols or numerals can be assigned by rule to the individual or their behaviors to whom the scale is applied, the assignment being indicated by the individual's possession of whatever the scale is supposed to measure" (p. 450). Scales can be used to measure a variety of areas: intelligence or aptitude, the proficiency or understanding of a specific area, personality traits, or attitudes and values.

DeCoster (2005) noted that scale construction is specific to survey design. Moreover, good scales are ones that are both valid and reliable. A valid scale is one that measures the construct that it is trying to assess and a valid scale is one that will consistently yield the same results.

DeCoster believed that researchers need to follow certain guidelines for writing survey questions. First, he argued that questions need to be simple. In other words, the items must be simple and concise so that respondents can answer items easily. It is important to note that scales that do not offer simple questions will often times cause misunderstanding and frustration for your participants. Second, make sure that the questions are understandable. When researchers offer vague or abstract concepts,

* Assoc.Prof., Department of Communication Studies, Texas Tech University, USA. n.punyanunt@ttu.edu

it may cause the participants to think of multiple interpretations. Thus, it would make the scale less reliable and valid. Third, avoid using biased language. For that reason, researchers should not sure emotional phrases or words in their questions. As a researcher, it is important to get your respondents to read the question. Fourth, it is necessary to use a common structure. Most Likert scales that are used have a common structure, meaning that each item is worth the same weight.

Trochim (2006) noted that "Constructing a survey instrument is an art in itself. There are numerous small decisions that must be made -- about content, wording, format, placement -- that can have important consequences for your entire study. While there's no one perfect way to accomplish this job, we have lots of advice to offer that might increase your chances of developing a better final product (pg.1)".

Trochim believed that survey research has changed a lot over the years. He stated, "We have automated telephone surveys that use random dialing methods. There are computerized kiosks in public places that allows people to ask for input. A whole new variation of group interview has evolved as focus group methodology. Increasingly, survey research is tightly integrated with the delivery of service. Your hotel room has a survey on the desk. Your waiter presents a short customer satisfaction survey with your check. You get a call for an interview several days after your last call to a computer company for technical assistance. You're asked to complete a short survey when you visit a web site (pg. 2)." Moreover, he noted," When most people think of questionnaires, they think of the mail survey. All of us have, at one time or another, received a questionnaire in the mail. There are many advantages to mail surveys. They are relatively inexpensive to administer. You can send the exact same instrument to a wide number of people. They allow the respondent to fill it out at their own convenience. But there are some disadvantages as well. Response rates from mail surveys are often very low. And, mail questionnaires are not the best vehicles for asking for detailed written responses. (pg. 2)." Further, A second type is the group administered questionnaire. A sample of respondents is brought together and asked to respond to a structured sequence of questions. Traditionally, questionnaires were administered in group settings for convenience. The researcher could give the questionnaire to those who were present and be fairly sure that there would be a high response rate. If the respondents were unclear about the meaning of a question they could ask for clarification. And, there were often organizational settings where it was relatively easy to assemble the group (in a company or business, for instance). (pg. 2)"

Dewes (1987) stated that "Scales can also be classified according to the source of scale score variation, following Torgerson (1958), as stimulus-centered, subject-centered, or response scales. Scale scores in stimulus-centered scales (also called judgment scales) reflect stimulus (item) differences along the measurement dimension. An example would be a life events scale, in which a respondent rates or ranks particular life events in terms of how stressful they are to the respondent. In contrast, for subject-centered scales (also called individual differences scales), scale scores reflect differences among the subjects (respondents) in terms of their standing along the scale's dimension. Personality trait scales of the inventory or questionnaire variety are common examples of subject-centered scales. Lastly, response scales are those for which scale score variation is attributed to both stimuli (items) and subjects (respondents). Scales constructed according to the Rasch scaling methodology (Wright & Masters, 1982) are examples of response scales. (pg. 220)".

In regards to scale format, Dewes mentioned that "Items in structured verbal scales typically consist of a stimulus part (the item stem) and a response part (the response choices). Item stems may consist of full sentences, phrases, or even single words. They may describe some attribute of an object (e.g., "The counselor appears trustworthy"), or the state of the object ("The counselor is passive"), or some event involving the object ("The counselor is reflecting the client's feelings"), to varying degrees of specificity or generality. Item stems ordinarily consist of single components, but may have two or more components (as in paired comparison or multiple rank-order scales). (pg. 220)".

Dewes argued that "In choosing a scale format, the general rule might be to choose the simpler format. However, there are other considerations: More complex formats might make the task of filling out the scale more interesting for the more experienced or knowledgeable respondent. When rating response formats are used, more scale points are better than fewer, because once the data are in, one can always combine scale points to reduce their number, but one cannot increase that number after the fact. Also, more scale points can generate more variability in response, a desirable scale characteristic if the response is reliable. Inordinate use of the middlemost scale point can be avoided by eliminating that scale point, that is, by using an even number of scale points. This has the further advantage of ensuring that the underlying dimension will be linear or can be made linear. At times rank ordering may be easier to do than rating, but use of ranking response formats may place limits on the statistical analysis of the data. Finally, the amount of space available for the scale (e.g., in an extended questionnaire) might preclude the use of certain formats. (pg. 221)."

A variety of different types of scales are available for social science research. A Bogardus Social Distance scale, is its simplest format, asks a series of questions, each decreasing in terms of social distance. With this type of scale, a researcher can predict a person's comfort level. Dewes wrote that "Respondents were asked whether they would be willing to admit members of a race or nationality group (a) to close kinship by marriage, (b) to membership in their club, (c) to their streets as neighbors, (d) to employment in their occupation, (e) to citizenship in their country, (f) only as visitors to their country, or (g) whether they would exclude them completely from their country. Admitting individuals at one level implies admitting them at lower levels but does not imply admitting them at higher levels." (pg. 221).

Guttman scales are unidimensional, cumulative measures that measure one variable. According to Dewes (1987), Guttman scales, "With a unidimensional scale, according to Guttman, knowledge of the respondent's scale score should permit the reproduction of the respondent's item score pattern. In a unidimensional scale, the items can be arranged in order (of endorsement or descriptiveness or whatever the underlying dimension is) in such a way that positive response to an item (e.g., agree , in an attitude scale) should imply positive response to all other items lower down the scale, and conversely, negative response to an item should imply negative response to all other items higher up the scale. To ascertain unidimensionality, Guttman developed the scalogram technique.Suppose we had a unidimensional attitude scale that was administered to a group of individuals. The scalogram technique would call for the data to be displayed as follows: Items are displayed as columns and ordered (from left to right) according to endorsement level from the most to the least endorsed item.

Individuals are displayed as rows and ordered (from top to bottom) according to total score, from highest to lowest score. If the test were perfectly unidimensional, then the scalogram would show an orderly stepwise progression of endorsement for both the individuals and the items. Any exceptions to this expectation can be easily seen in a scalogram display, and the number of exceptions can be expressed as a proportion of the total matrix (N individuals × m items). Guttman (1944) defines a coefficient of reproducibility as 1 minus the proportion of exceptions, where 1.00 means that the response pattern for any given scale score can be reproduced perfectly. When the coefficient of reproducibility is not high (e.g., below .9 or .8), the scalogram display will reveal the items that do not conform to expectation. After removing these items, the coefficient of reproducibility is recalculated, and the process repeated until the desired level of the coefficient is attained. Sometimes it may also be necessary to eliminate some aberrant individuals whose responses do not conform to the expected pattern. (This underscores the fact that response is a function not just of the scale or instrument but also of the respondent population. Aberrant individuals might be hypothesized to belong to a different population insofar as the scale is concerned.) (pg. 221-222)."

Likert scales measure the intensity of a person's response to different variables. Measures can range from "strongly agree" to "strongly disagree", or use other similar intensity labels. According to Dewes (1987, pg. 222-223), The Likert procedure can be described as follows:

1. A number of items are written to represent the content domain. Five-point anchored rating scales are typically used as response choices for each item (hence, the mistaken use of Likert to refer to the 5-point-rating item format). Scoring weights from 1 to 5 are assigned to the five rating-scale points. Direction of scoring (whether 1 or 5 is high) is immaterial provided it is consistent for all items.
2. The items are administered to a large group of respondents (N of at least 100). Each respondent's item rating choices are scored and the item scores summed to constitute the respondent's total score.
3. Items are selected according to their ability to discriminate between high and low scorers on total score. Likert used a group-difference procedure (difference in item means between high-scoring and low-scoring groups, e.g., uppermost 25% and lowermost 25%). One could also use an item-total-score correlation procedure, as is currently done in ability test construction. Maximizing item-total-score correlation will also maximize the scale's internal consistency reliability coefficient (coefficient alpha). Computer programs (e.g., the Statistical Package for the Social Sciences Reliability program) are available for use in this connection.
4. The best discriminating items are then selected to constitute the scale, and the scale score is obtained by summing the item scores for the selected items. At this point, scale scores can be treated as normative scores (i.e., transformed to standardized scores, used to determine percentile equivalents for specific populations, etc.).

Of all the scale construction methods, the most convenient for researchers is the Likert method because it can be employed with the use of ordinary SPSS programs. To implement the Likert method requires only (a) computing total score, (b) computing item-total-score correlations, and (c) computing alpha reliability for the final set of

items. Incidentally, reliability should be computed for every research use of Likert scales, not just at scale development, because reliability is a function not only of the scale but also of the respondent sample.

Unfortunately, not all scales that are purported to be Likert-type scales are constructed according to the Likert procedure. They only look like Likert scales because of the use of the 5-point rating response format (Triandis, 1971). If, in such scales, the correlation of the items with total scale is not high, then the interpretation of the scale score is problematic."

Semantic differential scales examine the unidimensional aspect of respondents' feelings. Dewes (1987) noted, Unlike the Likert, which uses only one rating dimension for all items in a scale, the semantic differential uses several rating dimensions for rating the same item or stimulus object. Semantic differential rating dimensions are typically bipolar, anchored at both ends with contrasting adjectives, with a 7-point rating continuum. Provided that response distributions are not forced, semantic differential data can be treated like any other rating data." (pg. 223). Krosnick (1991) noted that regardless of the type of scale used, researchers must realize that any changes, such as response options or word choice, made to a scale can have significant effects on the results. The primary goal of this study was to explore the effects of changing the response options of two distinct scales.

After all, the main objective of developing a scale is to create a measure that is valid (Clark & Watson, 1995). There are many factors that are important when creating the scale, such as the items and the variables. Perhaps, one of the most important aspects of scale development is the word choice. Clark and Watson noted that the word choice is extremely important for internal consistency and construct validity.

While many researchers (e.g., Sheatsley, 1983; Tourageu, Rips, & Rasinski, 2000; Weisberg, 2005) utilize scales in their research, more studies are needed in analyzing the methodological effect of altering scales and response options. Researchers have, however, examined specific scales (e.g., Boster & Hale, 1989; Chan, 1991; Cogliser & Schriesheim, 1994). Boster, 1994 and Hale (1989) examined the response scale ambiguity as a limitation to social comparison. Cogliser and Schriesheim (1994) conceptualized a bipolarity approach in the use of semantic differential scales. Hale, Boster, and Mongeau (1991) examined the phenomena of choice shifts in response scales, critiquing the odd of success type of response scale. They suggest using Likert scales. Chan (1991) considered the order of response items in Likert-type scales as a primacy factor in respondents. Dawis (1987) examined the design and development of scales used in psychological counseling research. She discussed the various scales and evaluates their use in research.

Clark and Watson (1995) noted that scale development is very popular and the amount of scales has increased dramatically over the years. They argued that several scholars ignored the concept of construct validity. Construct validity cannot be obtain from only one set of observations or self-reports from scales because scales only provide a means for determining inferences (Cronbach & Meehl, 1955).

The utilization of various survey items, especially word choices, to measure communication behaviors needs to be analyzed. Researchers have noted that positive and negative words used in measurements have caused response and systematic bias, which lead to a variety of methodological issues (e.g., misleading interpretations)

(Devellis, 2007; Horan, Distegano, & Motl, 2003). Moreover, Loevinger (1957) notes that Likert-type scales cause response biases, because the equal-intervals used in scales are frequently not warranted. Thus, scales are constructed may cause it to be less valid.

Proper scale selection and development are tantamount in the pursuit of quantitative studies. In some cases, researchers have altered a particular scale to fit the needs of a group of subjects or research conditions without considering the impact those changes might have on the results. The objective of this study was to examine the effect of scale adaptation on the measurement process. The goal was to see if there was an effect on changing the labels used in a Likert scale on the overall measurement results.

Two scales were selected, the Television Affinity Scale and the Perceived Realism Scale, to use for this study. The Television Affinity Scale (Rubin & Rubin, 1982) is a measure that looks at how individuals rate the importance of television. The scale has been employed in many research studies (Abelman, 1988; Conway & Rubin, 1991; and Rubin & Perse, 1987) and has been shown to be both valid and reliable. Abelman found a positive relationship between television affinity and ritualistic viewing of the 700 Club. Conway and Rubin discovered that television affinity was linked to motivations for watching television. Rubin and Perse found a connection between television affinity, perceived realism, and motivations for viewing soap operas. The Perceived Realism scale was produced by Greenberg (1974) and was later expanded by Rubin (1981). Rubin (1979) proposed that perceived realism is related to television viewing motivation. This scale has also been shown to be a valid and reliable instrument.

Both scales have five items and use a five-point Likert response scale, anchored by "strongly agree" and "strongly disagree." To see if changing the labels used for the response options would significantly alter the measurement results, a study was designed to test the following two hypotheses:

H1: Scores on the Television Affinity Scale will differ significantly when alternative response options are used.

H2: Scores on the Perceived Realism Scale will differ significantly when alternative response options are used.

Study One

Method, Sample

Three hundred and twenty undergraduate students in an introductory communication class at a large Midwestern university were surveyed. Questionnaires were used in a mass testing format. In other words, participants were instructed to attend one of five sessions for extra credit. Each session was one hour long. When students arrived, they were given instructions on how to fill out their scantrons and told about human subject procedures. Students were given all the scales in a randomized manner. Of the 320 subjects, 125 (39%) were male, 195 (61%) were female; 250 (78%) were between 18 to 24, 54 (17%) were 25-30, and 16 (5%) were over 30. 112 (35%) were freshman, 112 (35%) were sophomore, 64 (20%) were juniors, 22 (7%) were seniors, and 10 classified himself/herself as "other"; 272 (85%) were Caucasian, 19 (6%) were African-American, 29 (9%) were of other ethnic origin. Students' participation was voluntary; 144 (45%) watch at least thirty minutes to an hour of television, 118 (37%) watch around an hour and a half to three hours, 42

(13%) watch three to five hours, 10 (3%) watch less than thirty minutes, and only 6 (2%) watched more than five hours of television a day. None of the students reported not watching any television. Based on that data, the students were grouped into high (three hours or more), medium (between an hour and a half to three hours), and low television viewers (one hour or less). Students, who participated in the study, received course research credit for the participation.

Measures

Two measures were used for this study: the Television Affinity Scale and the Perceived Realism Scale. Five differently labeled response scales, each with a seven point range, were used for the two scales: (1) strongly agree – agree - agree some and disagree some – disagree -strongly disagree; (2) YES! – YES - yes - ? – no - NO - NO!; (3) not at all like me - not like me – somewhat unlike me – not sure – somewhat like me – like me - very much like me; (4) always - usually – often – occasionally – often not – usually not - never; and (5) almost never true – rarely true – occasionally true – often true – almost always true. Further, the questionnaire was created with the ten versions of the scales in random order and the order of the scale items was put into random order to reduce the effect of measurement order.

Television Affinity Scale. Television Affinity Scale measures viewers' attitudes towards television or certain television programs. The most commonly used scale to measure television affinity was created by Rubin and Rubin (1982). Rubin, Palgreen, and Sypher (1994) found that the scale was very reliable having Cronbach alphas ranging from .75 to .93. Research using this scale has mainly been linked to viewers' motivations for watching television (Rubin et al.).

Perceived Realism Scale. A popular scale that has been used to investigate mass media perceptions is the perceived realism scale. The scale measures whether viewers' actually believe what they see on television as realistic or not. The scale was originally used to look at how different racial and age populations varied in their television viewing behaviors. The PRS was developed by Rubin (1981) to understand media uses and gratifications research. The scale has proven to be reliable with Cronbach alphas ranging from .74 - .93 (Rubin, Palmgreen, & Sypher, 1994).

Results

The Television Affinity Scale proved to be a reliable instrument regardless of the response options used. Reliability alphas ranged from .84 for response options strongly agree-strongly disagree, to .90 for scale options always-never. The reliability of the Perceived Realism Scale did not prove to be as strong, with alphas ranging from .53 for scale options always-never, to .67 for scale options YES!-NO!. Table 1 provides a summary of data for each of the five versions of these two scales.

Table 1 Summary data for Television Affinity Scales and Perceived Realism Scales

Scale	*Range*	*Mean*	*Median*	*SD*	*Alpha*
Affinity					
Version 1	**5-35**	**14.86**	**14**	**7.32**	**.84**
Version 2	5-35	14.86	15	6.90	.88
Version 3	5-35	13.83	13	6.99	88
Version 4	5-35	14.66	15	7.12	.90
Version 5	5-35	15.51	15	7.16	.86
Realism					
Version 1	**5-29**	**16.37**	**17**	**4.97**	**.53**
Version 2	5-29	16.70	17	5.14	.67
Version 3	5-30	17.24	18	4.85	.62
Version 4	5-28	17.10	18	4.52	.53
Version 5	5-33	17.15	18	4.75	.64

Hypothesis one was supported by the data. The current study predicted that there would be significant differences in the scores on the Television Affinity Scale if response options were varied. A repeated measures multiple analysis of variance indicated a significant difference existed ($F (4, 1208) = 8.28$, $p = .001$). Table 2 summarizes the data on the five versions of this scale.

Table 2 Summary of Television Affinity Scale MANOVA data

Scale	Mean	SD
Version 1	14.59	7.14
Version 2	14.79	7.00
Version 3	13.74	6.97
Version 4	14.34	7.02
Version 5	15.22	7.12

Note. $N = 303$.

Hypothesis two was also supported by the data. The current study predicted that there would be significant differences in the scores on the Perceived Realism Scale if response options were varied. A repeated measures multiple analysis of variance indicated a significant difference existed ($F (4, 1208) = 3.95$, $p = .005$). Table 3 summarizes the data on the five versions of this scale.

Table 3 Summary of Perceived Realism Scale MANOVA data

Scale	Mean	SD
Version 1	16.23	5.03
Version 2	16.68	5.16
Version 3	17.04	4.98
Version 4	16.95	4.65
Version 5	17.02	4.83

Note. $N = 303$.

In order to investigate if the response options might affect response, such as variance, reliability, selection of middle option, selection of extreme points, etc. A more thorough analysis was conducted. Because the same participants completed all five

versions of the scales, one could not use tests that assume independence. For that reason, Woodruff and Feldt's procedures for comparing dependent alphas was used in this study. The results are shown in Table 4.

Table 4 Characteristics of the Television Affinity Scales and Perceived Realism Scales

Scale	***Range of Reliabilities***			***Range of Correlations***		
Affinity						
Version 1	.80	-	.90	.40	-	.80
Version 2	.80	-	.94	.42	-	.65
Version 3	.80	-	.94	.40	-	.68
Version 4	.82	-	.96	.37	-	.66
Version 5	.77	-	.93	.44	-	.61
Realism						
Version 1	.44	-	.60	.30	-	.67
Version 2	.50	-	.76	.33	-	.77
Version 3	.54	-	.72	.40	-	.65
Version 4	.39	-	.66	.42	-	.59
Version 5	.58	-	.77	.37	-	.62

At the same time, it was necessary to test for differences in collections between each scale and some common criterion. Thus, a correlation was conducted between television affinity and television viewing. In order to compare these results, Meng, Rosenthal, and Rubin's (1992) procedures were used to compare correlated coefficients. The results are located on Table 5.

Table 5 Correlations between the TRS and PRS Scales and Television Watching

Scale	***High***	***Medium***	***Low***
Affinity			
Version 1	.88	.86	.83
Version 2	.87	.82	.81
Version 3	.88	.74	.70
Version 4	.90	.88	.86
Version 5	.93	.86	.84
Realism			
Version 1	.80	.77	.77
Version 2	.86	.82	.84
Version 3	.78	.76	.76
Version 4	.82	.81	.82
Version 5	.71	.69	.70

Note.
High (three hours or more) = 48
Medium (between one hour and a half – three hours) = 118
Low (one hour or less) = 145

Study Two

Study two was created to see if the different media from television to Internet would impact the results. Hence, the following hypotheses were constructed:

H3: Scores on the Internet Affinity Scale will differ significantly when alternative response options are used.

H4: Scores on the Perceived Realism of the Internet Scale adapted for the Internet will differ significantly when alternative response options are used.

Sample

Two hundred and two undergraduate students in an introductory communication class at a large Southwestern university were surveyed. Questionnaires were distributed in a large lecture for extra credit. Of the 202 subjects, 100 (49%) were male, 102 (51%) were female; 190 (94%) were between 18 to 24, 10 (5%) were 25-30, and 2 (1%) were over 30. 112 (55%) were freshman, 23 (11%) were sophomore, 22 (11%) were juniors, 22 (11%) were seniors; 190 (94%) were Caucasian, 10 (5%) were African-American, 2 (1%) were of other ethnic origin. Students' participation was voluntary; 100 (49%) were involved at least thirty minutes to an hour of Internet, 35 (17%) were involved on the Internet for around an hour and a half to three hours, 35 (17%) were involved with the Internet for three to five hours,30 (15%) were involved with the Internet for less than thirty minutes, and only 2 (1%) were involved with the Internet for more than five hours a day. None of the students reported not using the Internet. Based on that data, the students were grouped into high (three hours or more), medium (between an hour and a half to three hours), and low Internet viewers (one hour or less). Students, who participated in the study, received course research credit for the participation.

Measures

Two measures were used for this study: The Internet Affinity Scale and the Perceived Realism Scale (Adapted for Internet). Five differently labeled response scales, each with a seven-point range, were used for the two scales: (1) strongly agree – agree - agree some and disagree some – disagree -strongly disagree; (2) YES! – YES - yes -? – no - NO -NO!; (3) not at all like me - not like me – somewhat unlike me – not sure – somewhat like me – like me - very much like me; (4) always - usually – often – occasionally – often not – usually not - never; and (5) almost never true – rarely true – occasionally true – often true – almost always true. Further, the questionnaire was created with the ten versions of the scales in random order and the order of the scale items was put into random order to reduce the effect of measurement order.

Internet Affinity Scale. Internet Affinity Scale measures viewers' attitudes towards the Internet. Papacharissi and Rubin (2000) found that the scale was very reliable having Cronbach alphas ranging from .92. Research using this scale has mainly been linked to viewers' motivations for using the Internet.

Perceived Realism of the Internet Scale. A popular scale that has been used to investigate mass media perceptions is the perceived realism scale. The scale measures whether viewers actually believe what they see on television as realistic or not. The scale was originally used to look at how different racial and age populations varied in their television viewing behaviors. The PRS was adapted and used in Anderson's 2005 study to understand media uses and gratifications research concerning the Internet.

Results

The Internet Affinity Scale proved to be a reliable instrument regardless of the response options used. Reliability alphas ranged from .81 for response options strongly agree-strongly disagree, to .91 for scale options always-never. The reliability of the Perceived Realism of Internet Scale did not prove to be as strong, with alphas ranging from .55 for scale options always-never, to .68 for scale options YES!-NO!. Table 6 provides a summary of data for each of the five versions of these two scales.

Table 6 Summary data for Internet Affinity Scales and Perceived Realism of Internet Scales

Scale	***Range***	***Mean***	***Median***	***SD***	***Alpha***
Affinity					
Version 1	5-35	14.86	14	7.32	.81
Version 2	5-35	14.86	15	6.90	.88
Version 3	5-35	13.83	13	6.99	88
Version 4	5-35	14.66	15	7.12	.91
Version 5	5-35	15.51	15	7.16	.86
Realism					
Version 1	5-29	16.37	17	4.97	.55
Version 2	5-29	16.70	17	5.14	.68
Version 3	5-30	17.24	18	4.85	.62
Version 4	5-28	17.10	18	4.52	.55
Version 5	5-33	17.15	18	4.75	.64

Hypothesis three was supported by the data. The current study predicted that there would be significant differences in the scores on the Television Affinity Scale if response options were varied. A repeated measures multiple analysis of variance indicated a significant difference existed (F (4, 804) = 8.28, p = .001). Table 7 summarizes the data on the five versions of this scale.

Table 7 Summary of Internet Affinity Scale MANOVA data

Scale	Mean	SD
Version 1	14.51	7.14
Version 2	14.33	7.00
Version 3	13.21	6.97
Version 4	14.54	7.22
Version 5	15.21	7.12

Note. N = 202.

Hypothesis two was also supported by the data. The current study predicted that there would be significant differences in the scores on the Perceived Realism Scale if response options were varied. A repeated measures multiple analysis of variance indicated a significant difference existed (F (4, 804) = 3.95, p = .005). Table 8 summarizes the data on the five versions of this scale.

Table 8 Summary of Perceived Realism of Internet Scale MANOVA data

Scale	Mean	SD
Version 1	16.56	5.13
Version 2	16.88	5.16
Version 3	17.12	4.99
Version 4	16.99	4.65
Version 5	17.22	4.83

Note. N = 202.

In order to investigate if the response options might affect response, such as variance, reliability, selection of middle option, selection of extreme points, etc. A more thorough analysis was conducted. Because the same participants completed all five versions of the scales, one could not use tests that assume independence. For that reason, Woodruff and Feldt's procedures for comparing dependent alphas was used in this study. The results are shown in Table 9.

Table 9 Characteristics of the Internet Affinity Scales and Perceived Realism of Internet Scales

Scale	***Range of Reliabilities***			***Range of Correlations***		
Affinity						
Version 1	.80	-	.90	.40	-	.80
Version 2	.80	-	.94	.42	-	.65
Version 3	.80	-	.94	.40	-	.68
Version 4	.82	-	.96	.37	-	.66
Version 5	.77	-	.93	.44	-	.61
Realism						
Version 1	.44	-	.60	.30	-	.67
Version 2	.50	-	.76	.33	-	.77
Version 3	.54	-	.72	.40	-	.65
Version 4	.39	-	.66	.42	-	.59
Version 5	.58	-	.77	.37	-	.62

At the same time, it was necessary to test for differences in collections between each scale and some common criterion. Thus, a correlation was conducted between television affinity and television viewing. In order to compare these results, Meng, Rosenthal, and Rubin's (1992) procedures were used to compare correlated coefficients. The results are located on Table 10.

Table 10 Correlations between the TRS and PRS Scales and Internet Involvement

Scale	***High***	***Medium***	***Low***
Affinity			
Version 1	.88	.86	.83
Version 2	.87	.82	.81
Version 3	.88	.74	.70
Version 4	.90	.88	.86
Version 5	.93	.86	.84
Realism			
Version 1	.80	.77	.77
Version 2	.86	.82	.84
Version 3	.78	.76	.76
Version 4	.82	.81	.82
Version 5	.71	.69	.70

Note.
High (three hours or more) = 32
Medium (between one hour and a half – three hours) = 100
Low (one hour or less) = 70

Discussion

All in all, scales are a commonly used method for measuring variables in social science research. As researchers use scales designed in previous studies, they may find it necessary to alter the scale to fit the subjects or research conditions. Krosnick (1991) mentioned that minor changes to a scale may significantly impact the results. The current study sought to explore the effects of these alterations by comparing scores on two scales when response options were changed.

Data from the current study reaffirmed the reliability of the five-item version Television Affinity Scale. However, reliability alphas on the five-item Perceived realism Scale were only moderate. Further analysis of the data may suggest that removing one of the items could improve the reliability. For both scales, the data suggests that the reliability of the instrument is not dramatically affected when response options are changed. In addition, when the scales were adapted to the Internet, the scales were still valid. However, evidence did support all hypotheses proposed in the current study. When response options were changed, significant differences occurred in the way subjects scored on the scales. These changes may also affect research results and construct validity. Findings from this study suggest that such alterations need to be carefully considered when designing a study.

Specifically, for the TRS, results revealed that using version 4 (always to never response option) yielded the highest Cronbach alpha. It also correlated higher than the other versions among television viewing time. For the PRS, results revealed that version 2 (YES! – NO!) yielded the highest Cronbach alpha. In addition, when the versions for the PRS scale were correlated with television viewing time, version 2 had highest correlations compared to the other versions. Regarding both scales, version 5 (almost never true – rarely true – occasionally true – often true – almost always true)

had the highest means. This indicates that version 5 would yield significantly higher results than the other response option versions.

Some of the cognitive /psychological processes that might have accounted for the effects in the response option wording that were reported in this paper could be due to the fact that some of the options did not make sense with the survey questions/statements. In addition, these scales were given to undergraduate college students, who may have enjoyed other types of media, such as movies or the Internet, rather than television. It is quite possible these response options have an effect on the participants' comprehension, retrival, judgement/response (Toranguea, Rips, & Rasinski, 2000). Because the participants were college students, the results of this study may not be generalized to the general public.

McFadden and Winter (2001) noted that "There is overwhelming empirical evidence that cognitive limitations and social interactions lead to biases in responses to survey questions. A well-documented example is the "unfolding brackets" technique. While this question design can successfully reduce survey non-response, it introduces anchoring effects that might distort responses significantly (see, Hurd, et al., 1998; Hurd, 1999). Similarly, in questions on subjective expectations about future income, health conditions, or survival probabilities, the format and the wording of questions can influence responses heavily (i.e., Hurd, McFadden, and Gan, 1998) (pg. 1)".

Moreover, Dewes (1987) found that "Although reliability and validity concerns are of the essence, there are other less important (but nonetheless, important) considerations. Some of these have been mentioned, for example, administrative concerns. Another concern is the character of the score distribution generated by the scale—in part, a function of the respondent sample. Most users would prefer a scale that ordinarily produces a reasonably normally distributed set of scores. However, if the scale were to be used for diagnostic purposes, a user might prefer one that generates a skewed distribution, the direction of skew depending on whether low scores or high scores are diagnostic. Scales, like ability tests, can be so constructed as to produce the shape of score distribution that is desired, by selecting the appropriate items. Another concern is that the scale produce sufficient score variation to be useful, that is, produce unattenuated correlations." (pg. 224).

Again it is important to note that there are several rules to remember when constructing scale items. First, use simple language so that our participants will understand. Second, it is also important to be clear and concise. Third, use response formats that are consistent with the survey questionnaire. In other words, it should answer what the question is asking. Fourth, use be consistent in the survey. If there are discrepancies in the scale it will make it very confusing for the respondent. If researchers follow these simple steps, then their scale will be more likely to be valid and reliable.

Future studies exploring the nature of scales should examine other instruments to see if these findings are consistent across various scales. Additionally, the effects of other changes that are commonly made to instruments, such as rewording questions, should also be explored. Results from this study prove that response options are important and may affect the reliability and the validity of research findings.

REFERENCES

Abelman, R. (1988). Motivations for viewing "The 700 Club." Journalism Quarterly, 65, 112-118.

Anderson, T. L. (2005). Relationships among Internet attitudes, internet use, romantic beliefs, and perceptions of online romantic relationships. CyberPsychology & Behavior, 8(6), 521-531.

Babbie, E. (1995). The practice of social research. (7th ed.). Belmont, CA: Wadsworth Publishing.

Bishop, G. F. (2004). The illusion of public opinion. Lanham, MD: Rowman & Littlefield.

Bishop, G.F. (1987). Experiments with the middle response alternative in survey questions. Public Opinion Quarterly, 51, 220-232.

Boster, F. J., & Hale, J. L. (1989). Response scale ambiguity as a moderator of the choice shift. Communication Research, 16, 532-548.

Chan, J. C. (1991). Response-order effects in Likert-type scales. Educational and Psychological Measurement, 51, 531-542.

Clark, L. A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. Psychological Assessment, 7(3), 309-319.

Cogliser, C. C., & Schriesheim, C. A. (1994). Development and application of a new approach to testing the bipolarity of semantic differential items. Educational and Psychological Measurement, 54, 594-606.

Conway, J. C., & Rubin, A. M. (1991). Psychological predictors of television viewing motivation. Communication Research, 18, 443-463.

Converse, P. (1964). The nature of belief systems in mass publics. In D Apter (Ed.), Ideology and discontent (pp. 206-261). New York: Free Press.

Converse, P. (1970). Attitudes and non-attitudes: Continuation of a dialogue. In E Tufte (Ed.), The quantitative analysis of social problems (pp. 168-189). Reading, MA: Addison-Welsley.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological test. Psychological Bulletin, 52, 281-302.

Dawis, R. V. (1987). Scale construction. Journal of Counseling Psychology, 34, 481-494.

DeCoster, J. (2005). Scale Construction Notes. Retrieved June 18, 2007 from http://www.stat-help.com/notes.html

Devellis, R. F. (2007). Scale development: Theory and Applications (2nd. Edition). Newbury Park, CA: Sage Publications.

Fowler, F. J., Jr. (1995). Improving survey questions: Design and evaluation. Thousand Oaks, CA: Sage.

Greenberg, B. S. (1974). Gratifications of television viewing and their correlates for British children. In J. G. Blumler & E. Katz (Eds.), The uses of mass communication: Current perspectives on gratifications research (pp. 71-92). Beverly Hills, CA: Sage.

Hale, J. L., Boster, F. J., & Mongeau, P. A. (1991). The validity of choice dilemma response scales. Communication Reports, 4, 30-34.

Horan, P. M., DiStefano, C., & Motl, Robert, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? Structural Equation Modeling, 10(3), 435-455.

Hurd, M. D. (1999): Anchoring and acquiescence bias in measuring assets in household surveys. Journal of Risk and Uncertainty, 19, 111−136.

Hurd, M. D., D. L. McFadden, H. Chand, L. Gan, A. Merrill, and M. Roberts (1998): Consumption and savings balances of the elderly: Experimental evidence on survey response bias. In D. Wise(ed.), Frontiers in the Economics of Aging, 353−387. Chicago, IL: University of Chicago Press.

Hurd, M. D., D. L. McFadden, and L. Gan (1998): Subjective survival curves and life-cycle behav-ior. In D. A. Wise (ed.), Inquiries in the Economics of Aging, 259−305. Chicago, IL: University of Chicago Press.

Hurd, M. D., D. L. McFadden, and A. Merrill (2001): Predictors of mortality among the elderly. Forthcoming in D. A. Wise (ed.), Themes in the Economics of Aging. Chicago, IL: University of Chicago Press.

Kerlinger, F. N. (1992). Foundations of behavioral research. (3rd Ed.). New York: Harcourt Brace.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. Applied Cognitive Psychology, 5, 213-126.

Loevinger, J. (1954). The attenuation paradox in test theory. Psychological Bulletin, 51, 493-504.

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. Psychological Bulletin, 111, 172-175.

McFadden, D., & Winter, J. (2001). Experimental analysis of survey response bias over the internet : Some results from the Returement Perspectives Survey. Retrieved June 18, 2009 from http://74.125.95.132/search?q=cache%3AwrIFpkc0maAJ%3Ahrsonline.isr.umich.edu%2Fsitedocs%2Fconference%2F200111%2Fpaper8.pdf+constructing+Internet+survey&hl=en&gl=us

Papacharissi, Z., & Rubin, A. M. (2000). Predictors of internet use. Journal of Broadcasting and Electronic Media, 44(2), 175-196.

Rubin, A. M. (1979). Television use by children and adolescents. Human Communication Research, 5, 109-120.

Rubin, A. M. (1981). An examination of television viewing motivations. Communication Research, 8, 141-165.

Rubin, A. M., & Perse, E. M. (1987). Audience activity and soap opera involvement: A uses and effects investigation. Human Communication Research, 14, 246-268.

Rubin, A. M., & Rubin, R. B. (1982). Contextual age and television use. Human Communication Research, 8, 228-244.

Rubin, R. B., Palmgreen, P., & Sypher, H. E. (1994). Communication research measures: A sourcebook. New York: Guilford Press.

Sheatsley, P. F. (1983). Questionnaire construction and item writing. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.) Handbook of survey research (pp. 195-230). New York: Academic Press.

Trochim, W. M. K. (2006). Research Methods Knowledge Base. Retrieved June 18, 2009 from: http://www.socialresearchmethods.net/kb/survtype.php

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge: Cambridge University Press.

Weisberg, H. F. (2005). The total survey error approach: A guide to the new science of survey research. Chicago: University of Chicago.

Woodruff, D. J., & Feldt, L. S. (1986). Tests for equality of several alpha coefficients when their sample estimates are dependent. Psychometrika, 51, 393-413.