5-2010

Labels for Anchors: The Good, the Bad, and the Endpoints

James Lee Hellrung II
Western Kentucky University, james.hellrung295@wku.edu

LABELS FOR ANCHORS:
THE GOOD, THE BAD, AND THE ENDPOINTS

A Thesis
Presented to
The Faculty of the Department of Psychology
Western Kentucky University
Bowling Green, Kentucky

In Partial Fulfillment
Of the Requirements for the Degree
Master of Arts

By
James Lee Hellrung II

May 2010

LABELS FOR ANCHORS:
THE GOOD, THE BAD, AND THE ENDPOINTS

Date Recommended___March 26, 2010___

__Reagan Brown_____
Director of Thesis

__Elizabeth Shoenfelt_____

__Tony Paquin _____

_____
Dean, Graduate Studies and Research Date

Acknowledgements

First, I would like to express gratitude to my thesis chair, Dr. Reagan Brown, for his patient guidance throughout the process of writing this thesis. I am sure you are as happy as I am to see this project completed. Next, I would like to thank Dr. Betsy Shoenfelt for her continued assistance and support towards the finishing of my degree. I would also like to thank Dr. Tony Paquin for being on my committee.

Table of Contents

LABELS FOR ANCHORS:

THE GOOD, THE BAD, AND THE ENDPOINTS

James Lee Hellrung II May 2010 Pages 27

Directed by: Reagan Brown, Elizabeth Shoenfelt, and Tony Paquin

Department of Psychology Western Kentucky University

This study was designed to determine if any differences in internal consistency existed between different designs of scale anchors. The three different designs explored were properly designed scales, improperly designed scales, and endpoint only scales. Two-Hundred and thirty-five participants rated the frequency of which they performed various computer activities on a survey using one of the three different designs. Contrary to expectations, internal consistency did not differ across the three designs.

LABELS FOR ANCHORS:

THE GOOD, THE BAD, AND THE ENDPOINTS

Studies in social sciences frequently use standardized response scales to obtain the attitude and opinion data from people. Thus, the results of many studies are dependent upon the psychometric soundness of a scale. Researchers repeatedly have attempted to develop scales that provide an accurate view of peoples' reactions to selected stimuli (Bass, Cascio, & O'Connor, 1974; Schriesheim & Novelli, 1989; Thurstone, 1931; Weng, 2004). Thurstone's research offered an early method for the development of scales anchors that would be effectively labeled so that people would be able to accurately respond to the items. Other studies since Thurstone have attempted to create even better response scales (e.g., Schriesheim & Novelli) that accurately and reliably measure a person's reactions to different stimuli.

Researchers who fail to develop proper scales may inadvertently sabotage their studies as these poorly developed scales may yield unreliable results. Moreover, in the cases of improperly developed scales, a simpler and more pragmatic method might be the best way to obtain more accurate results. For example, a researcher might label fewer of the response points to enable respondents to make assumptions about what is supposed to be in the unlabeled response points. One method is to simply label the endpoints of a scale, effectively denoting only the high and low points of the scale. Such a method removes potentially confusing middle labels with unclear definitions that might negatively affect the reliability of an instrument. However, this seemingly pragmatic solution might lower the reliability of the instrument, if the participants were unable to properly rate their response without a word or phrase. Therefore, it would be beneficial to

determine if differences exist in the reliability of scales that are properly developed, improperly developed, and pragmatically (i.e., anchors only on the scale endpoints) developed.

The Use of Fewer Anchors

Several studies have addressed the proper number of response points on a scale. Churchill and Peter (1984) found that scales become more reliable with the addition of more response points on a scale. Jenkins and Taber (1977), as well as Lissitz and Green (1975), ran Monte Carlo analyses to determine the point of diminishing returns. Both studies simulated a variety of conditions (e.g., item covariances) and concluded that there is little benefit to be obtained from extending a scale beyond five points.

Although much research exists discussing the proper method to create a scale with effective intervals, few research articles are available regarding the reliability of labeling only the endpoints on a scale. Moreover, findings about the differences between endpoint only and fully labeled scales are often contradictory. Churchill and Peter (1984) conducted a meta-analysis to examine whether differences exist between endpoint only and fully labeled scaling. In their meta-analysis, Churchill and Peter analyzed 108 studies with a combined sample size of over 27,000 participants. They found that the different types of scales did not exhibit a difference in reliability. It is worth noting that Churchill and Peter did not examine the types of labels that were used during the studies. As such, it is possible the studies that had the scales fully labeled might have had poorly designed labels. Thus, the differences between properly fully labeled scales, improperly fully labeled scales, and endpoint labeled scales might have been obscured by the combination of the properly and improperly fully labeled scales into one category. Moreover, a

smaller meta-analysis by Ofir, Reddy, and Bechtel (1987) confirmed the lack of differences in reliability between the different types of scales found in Churchill and Peter study. Like Churchill and Peter, Ofir et al. did not indicate whether the fully labeled scales were developed properly.

A more recent study that examined the differences in reliability between fully labeled scales and endpoint-labeled scales was conducted by Weng (2004). Weng used properly designed scale labels in order to ensure the scales were as reliable as possible. Weng found that the fully labeled scales consistently yielded responses with much higher reliability than those of the scales with labels only on the endpoints. Therefore, it appears that scales with carefully designed labels are more reliable. It should be noted that Weng's study was conducted across 13 colleges, all of which are in Taiwan. Thus, there may have been some cultural differences that accounted for the difference in results between it and the previous studies. Moreover, none of the studies determined whether endpoint only labeling was superior to poorly labeled anchors.

Churchill and Peter (1984) failed to find reliability differences between completely labeled scales and scales with only the endpoints labeled. Conversely, Weng (2004) found that the completely labeled scales had a better reliability compared to scales with only the endpoints labeled. The difference between the two studies was that Weng had better methods for proper scale labeling during the development of the scale. Therefore, Weng was able to establish that properly weighted anchors were more reliable than anchors with only the endpoints labeled. However, whether endpoint labeled anchors are more reliable than poorly labeled anchors remains unresolved.

Developing Good Intervals for Anchors

Over the years, two popular techniques for developing appropriate and reliable weights for the scale anchors have emerged. Thurstone's Case III (1931) pair comparison study was the first technique to emerge as a popular method to develop effective labels for the responses on different scales. With Thurstone's Case III method, participants assign any number above one to different words presented to them (Schriesheim & Novelli, 1989). The words, rank ordered by mean rating, were then used to develop scale labels. Thurstone's method was an effective start towards developing a magnitude estimation technique. Nevertheless, there is an inherent problem with the method due to the unlimited upper end of the rating process. This unlimited upper boundary is problematic because different raters will hold different perceptions of the highest value for the highest word. This disagreement occurs not only at the upper end, but also along the rest of the scale (i.e., all points down).

Bass et al. (1974) developed a technique that is nearly identical to Thurstone's (1931) method and has been used in many studies (e.g., Schriesheim & Novelli, 1989; Paquin, Moore, & Sanchez-Ku, 2000). Bass et al.'s technique was called the Magnitude Estimation Technique (MET). The only difference between Bass et al.'s method and Thurstone's method is a discussion of ratio scaling with the participants. Ratio scaling utilizes the rater's concept of different terms in relation to other terms. For instance, if a subject were to think of the word "often" as meaning twice as frequent as the word "seldom," he or she would assign a numerical value to the word "often" that was twice the value assigned to the word "seldom." Although the introduction of ratio concepts into the scaling process is desirable, the benefits may not be realized due to the inconsistent

upper limit to the ratings (i.e., although the ratios are the same, 20 versus 10 is not the same as 50 versus 25). Bass et al. (1974, p. 315) gave the following instructions to their participants.

> …On the attached form please assign a number to what you conceive 'sometimes' to mean. You may use any number greater than or equal to zero (0). Please place the number on the line next to 'sometimes.' Then, using the number you have assigned to 'sometimes' as a standard, please assign a number to each of the other words on the form indicating each word's value greater than or equal to zero (0). The number placed next to each word or phrase should reflect what you feel that word or phrase means when compared with 'sometimes.' For example, if you assign a value of 50 to 'sometimes,' you might assign the value 100 to any other word or phrase which you felt represented twice the frequency of 'sometimes.' You may use any whole or decimal number greater than or equal to zero (0), just as long as you feel it represents the numerical value of a word or phrase when that word or phrase is compared with 'sometimes.'

Participants then rated 38 expressions of frequency (such as "always," "often," and "seldom") and 43 expressions of amount (such as "hardly any," "some," and "all"). Therefore, participants rated a total of 81 words (including "sometimes"). Participants received two forms (one for frequency and one for amount), each containing one of five different random orders of all of the words (Bass et al.). Thus, participants did not rate the words in any logical order. Moreover, the entire experiment was designed to be resistant to any form of an order effect; the results were most likely not biased toward incidental word placement.

Twenty five years after Bass et al. (1974), Paquin et al. (2000) conducted a study based on the Bass et al. study. Paquin et al. offered insight into how a more modern subject pool responded to a MET response set. In their replication, Paquin et al. reported that approximately ten percent of the words had changed over five ranks since Bass et al. conducted their study. The studies have inconsistent results with regards to the ranking of specific words. Either the population values have changed since 1974, or the differences were due to sampling error. Paquin et al. also noted in their data that several of the words near the top of the ranking had a standard deviation greater than that of the mean score of the word. Therefore, the words were extremely skewed by some participants setting abnormally large maximums for their rankings.

Schriesheim and Novelli (1989) compared Thurstone's (1931) Case III pair comparison technique to the Magnitude Estimation Technique. In order to make the different scales from the studies similar to each other, Schriesheim and Novelli balanced the differences between the different scales using an endpoint equating linear transformations. Specifically, the researchers transformed each MET score with log transformations which utilized the mean and standard deviation of each word's score. Schriesheim and Novelli then transferred the Thurstone's Case III scores to their implied ranks. Once the scales were equated with one another, Schriesheim and Novelli found that the Magnitude Estimation Technique replicated better across samples when compared to Thurstone's Case III pair comparisons. Based on this result, a modified Magnitude Estimation Technique will be utilized to derive the appropriate scales for this study.

The Present Study

The current study will investigate differences in reliability among different types of labels on a common scale. Moreover, the current study hopes to address the problem previously discussed concerning the Bass et al. (1974) study. Specifically, frequency rankings derived from the modified Magnitude Estimation Technique will be used to design both a properly and improperly created set of labels.

This study will examine differences in reliability among scales with properly weighted anchors, improperly weighted anchors, and anchors on only the endpoints. The study has two hypotheses.

Hypothesis 1: Properly labeled scales will exhibit greater internal consistency than improperly labeled scales and endpoint only scales.

Hypothesis 2: Endpoint labeled scales will exhibit greater internal consistency than improperly labeled scales.

Pilot Study

Method

Participants

Eighty-nine undergraduate students (78% female, 87% White) from a large southeastern university participated in the pilot study. Participants completed the experiment in order to fulfill part of a class requirement.

Materials

For descriptive information regarding the sample, participants completed a demographic information form (Appendix A). Because the main part of this study consisted of responses to a questionnaire with various anchors of known properties, it was necessary to collect data on a variety of possible anchors. To obtain this data, the researcher asked participants to offer their quantitative interpretations of the meaning of various words or phrases used to describe frequency (Appendix B). The design of the rating form was similar to the forms used in the Bass et al. (1974) as well as Paquin et al. (2000) studies. However, the form used in the present study contained instructions on the minimum and maximum values (i.e., 0 equals never and 100 equals always). The ratio example, used in Bass et al., was changed to allow for the previously stated alteration. The participants also completed the Computer Usage form (Appendix C), in order to test the variability of the items. The Computer Usage form used the response scale developed by Bass et al. for a five point scale (5 “Always,” 4 “Very Often,” 3 “Fairly Many Times,” 2 “Occasionally,” and 1 “Never”).

Procedure

Participants were able to access the study at any time via the online study board software program for registration. Data were obtained with a proprietary computer program. After participants registered for the online study, they were presented with the informed consent information. Participants who consented to participate in the study then rated the frequency descriptors displayed by the computer software. Once participants finished with the online form, they read a paragraph in which they were thanked for their time and dismissed from the website.

Results

Of the original 89 participants who enrolled in the study, the responses of only 56 participants were included in the data analysis; the other 33 participants were eliminated from the study because these participants did not follow directions. Specifically, the 33 excluded participants either failed to rate "always" as 100 or failed to rate "never" as 0 (despite the directions stating that all participants should rate "always" as 100 or "never" as 0). For the remaining 56 participants' responses, the researcher calculated the means and standard deviations of the ratings of the words and phrases. Please refer to Appendix D for these means and standard deviations.

The computer usage survey used in the main study was pilot tested using the five scale labels, as recommended by Bass et al. (1974). The piloted computer usage survey had a Cronbach's alpha of 0.76. A split half reliability analysis also yielded a Pearson's correlation of 0.76 ($p < 0.01$).

A division of the available space (i.e., from 0 to 100 points) into evenly spaced increments to allow for five scale points, resulted in the following targets for the anchor

means: 0, 25, 50, 75, and 100. These increments represent the ideal scale point values. "Often," "sometimes," and "fairly infrequently" were selected for the points of 4, 3, 2, respectively, due to their means (which differed from their target means by 0.16, 3.3, and 1.14 respectively) and low standard deviations.

For the poorly labeled scale condition, three words or phrases that had the highest standard deviations were selected. The selected words or phrases were also selected because of their uneven spacing of their means. The words or phrases in the poorly labeled scale condition were: "rather frequently," "fairly many times", and "to some degree"; they differed from their target means by 3.2, 17.5, and 11.34 respectively.

Main Study

Method

Participants

Two hundred and thirty five undergraduate students (80% females, 84% White) from a large southeastern university participated in the study. The mean age was 19.14 (SD= 2.34). The participants were randomly assigned to one of the three conditions. Participants completed the experiment in order to fulfill part of a class requirement.

Materials

The frequency statements from the pilot study were used to create three sets of anchors for the study. The three sets consisted of a good set of anchors, a poor set of anchors, and a set of endpoint anchors. For all three conditions the endpoints were "always" and "never". Therefore, the only difference between the three conditions was the middle three labels. Five anchors were used because Monte Carlo studies done by both Jenkins and Taber (1977) and Lissitz and Green (1975) indicated that the point of diminishing returns for the reliability of a scale begins at five anchors. The good anchors had five words with means that were relatively equal distance from one another and exhibited high agreement (i.e., low standard deviations). Equally spaced anchors allowed for interval level data. The poor anchors had "always" and "never" as the endpoints with middle words that were not equidistant from one another. The three middle words also had a high standard deviation, indicative of poor agreement. The endpoint anchors consisted of "never" and "always" with three unanchored points in between. The Computer Use form (Appendix C) served as the common stimulus from which participants rated the frequency based actions on the statements.

After each item, a response scale was provided for the participants to select what they felt were the most appropriate response of frequency to correspond to the scenario. Each response scale had five numerical integer values (numbered one through five) equally spaced on a line. The properly designed anchor condition had the properly designed labels under each response point. The poorly designed anchor condition had the improperly designed poor label markers under each response point. Finally, the endpoint anchor group had only the endpoints labeled on the scale with three unlabeled points in between.

Procedure

Participants were able to access the study at any time during the day via the online study board software program. The study board then directed them to proprietary computer program in which the study was actually conducted. When the participants registered for the study, they were randomly assigned to one of the three conditions. After participants had registered for the online study, they first read the online consent form describing the study. Participants who consented to participate in the study then completed the online questionnaire. Once participants were finished with the online form, they read a paragraph in which they were thanked for their time and dismissed from the website.

Analysis

The anchor type (properly labeled, poorly labeled, and endpoints) served as the independent variable. The responses to the computer usage questionnaire served as the dependent variable. The specific data calculated were coefficient alpha as well as simple split half correlations (based on a random split of the items). Both the coefficient alphas,

as well as the split half correlations are needed. The split half correlations can be analyzed. However, the coefficient alpha is needed to ensure that the split half correlations are representative of all of the possible split half correlations. Significance was determined with the z-test for differences between correlations from independent samples.

Results

The endpoint only condition, properly designed label condition, and poorly designed label condition had 81 participants, 80 participants, and 74 participants, respectively. The Cronbach alphas for the three conditions were 0.75, 0.81, and 0.81 for the endpoint only, properly designed label, and poorly designed label conditions, respectively. A split half reliability analysis (odd/even split) yielded Pearson correlations of 0.66, 0.76, and 0.73 for the endpoint only, properly designed label, and poorly designed label conditions, respectively.

A z-test of the difference between independent sample correlations was employed to examine whether the endpoint only and poorly designed label conditions' coefficient alphas significantly differed from the properly designed label condition's coefficient alpha. The reliabilities of the conditions were not significantly different from one another. The same analysis was performed on the split half reliability coefficients, and none of the reliabilities of the conditions differed significantly from one another. The correlation comparisons can be found in Table 1 below.

Table 1

z-test for Correlation Comparisons

Comparison	z score
Split half Proper vs. Improper	.37
Split half Proper vs. Endpoint	1.18
Cronbach's Proper vs. Improper	0
Cronbach's Proper vs. Endpoint	0.87

Note. All comparisons non-significant ($p > .05$)

Discussion

The differences between the endpoint-only, improperly labeled, and properly labeled scales were not statistically significant. Therefore, none of the hypotheses were supported. No inferences can be reliably drawn from the results. The lack of statistically significant results could mean that there are not any actual differences among the reliabilities of different scale label designs. The results of this study support Churchill and Peter (1984) and Ofir et al.'s (1987) meta-analyses. Although Churchill and Peter as well as Ofir et al. compared endpoint-only scales to fully labeled scales, properly and improperly developed scale labels were differentiated from one another in this study. However, no significant differences in reliability were found between endpoint-only, improperly labeled, and properly labeled scales in this study.

As with all null results, the lack of statistically significant differences makes it rather difficult to conclude whether there truly are no differences in reliability among the different scale formats. It is always possible that the null results obtained in the present study are due to Type II errors or factors associated with the design of the study, such as poor participant attention.

Despite the lack of statistically significant differences across the reliabilities, this study offers some insight into the development and use of different scale labels. The lack of significant differences between the reliabilities offers the tentative conclusion that scale developers have flexibility in the way they label their scales. This study supports the concept that the diverse range of scales used in a variety of social science disciplines does not necessarily impact the reliability of one study.

One of the main limitations with this study is the use of college students, a sample of people who may not be representative of the general population. The students might have ignored the anchors, and/or the directions. Future studies could use manipulation screens and checks to distill the more conscientious respondents from the less conscientious respondents. It is also possible that the use of a stimulus (i.e., questionnaire) with greater internal consistency would increase the likelihood of detecting differences between conditions. Finally, it might be rewarding to investigate whether the results found here are also found for scales composed of seven or nine anchors. It is possible that null results were obtained here simply because there were only three scale points available between the two extremes. In other words, there simply were not many opportunities for ambiguity between the positions of always and never on a five point scale.

References

Bass, B. M., Cascio, W. J., & O'Connor E. J. (1974). Magnitude estimation of expressions of frequency and amount. *Journal of Applied Psychology, 59,* 313-320.

Churchill, G. A., & Peter, J. P. (1984). Research design effects on the reliability of rating scales: A meta analysis. *Journal of Marketing Research, 21,* 360-375.

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

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability. *Journal of Applied Psychology, 60,* 10-13.

Ofir, C., Reddy, S. K., & Bechtel, G. G. (1987). Are semantic response scales equivalent? *Multivariate Behavior Research, 22,* 21-38.

Paquin, A. R., Moore, D., & Sanchez-Ku, M. L. (April, 2000). Establishing anchors: An evaluation of the magnitude estimation technique. Paper presented at the 2000 meeting of the Society for Industrial Organizational Psychology. New Orleans, LA.

Schriesheim, C. A., & Novelli, L. (1989). A comparative test of the interval-scale properties of magnitude estimation and case III scaling and recommendations for equal-interval frequency response anchors. *Educational and Psychological Measurement, 49,* 59-74.

Survey Monkey. (n.d.) Retrieved from www.SurveyMonkey.com

Thurstone, L. L. (1931). The measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26,* 249-269.

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement, 64,* 956-972.

Appendix A

Demographic Form

Age (In years)_____

Gender_____

Ethnicity_____

Year in school (select one)

Freshman Sophomore Junior Senior Other

Appendix B

Frequency Rating Form

Some words are used to describe actions which happen all of the time. Other words are used to describe actions which almost never occur. Still other words are used to convey the frequency of an event that happens about half of the time. This study concerns quantifying what these words mean. Using a scale of 0 to 100 where 0 means ‘never’ and 100 means ‘always’, please rate what the following words mean to you.

Never_____	Always_____	Seldom_____
Often_____	Now and Then_____	Continually_____
Commonly_____	Not at all_____	Not often_____
Once in a while_____	Rarely_____	Almost never_____
Constantly_____	Fairly often_____	Occasionally_____
Sometimes_____	Very rarely_____	Usually_____
Very often_____	Hardly at all_____	Infrequently_____
To some degree_____	Quite often_____	Some of the time_____
Rather frequently_____	Frequently_____	A great deal of the time_____
Rather seldom_____	Fairly many times_____	None of the time_____
Not very often_____	Very seldom_____	Fairly infrequently_____
Very infrequently_____	Very frequently_____	Hardly ever_____
A great many times_____		

Appendix C

Rating Form Computer Usage

Using the following scale, please rate the following actions with respect to how frequently you perform the activity.

1 Never 2 Occasionally 3 Fairly Many Times 4 Very Often 5 Always

Scanning documents

Saving a document

Downloading songs

Tweeting on Twitter

Checking email

Checking Facebook or MySpace

Downloading videos

Uploading pictures

Typing documents for school

Watching videos on you tube

Watching videos on hulu

Reading the news online

Chatting online

Using the calculator function on a computer

Using the calendar function on a computer

Checking for viruses

Making spreadsheets on excel

Online banking

Using skype/webcam

Adjusting Monitor Brightness

Defragmenting the hard drive

Scanning the hard drive disk for errors

Saving data to a portable storage device (such as a flash drive)

Playing solitaire

Shopping online

Browsing the internet

Appendix D

Table of Means and Standard Deviations of the Phrases from the Pilot Study

Word or Phrase	Mean	Standard. Deviation
Always	100.00	.000
Constantly	96.71	5.200
Continually	90.14	13.111
Very frequently	87.86	9.356
Very often	86.25	9.364
A great deal of the time	84.71	7.981
A great many times	83.70	10.200
Frequently	83.16	10.137
Quite often	79.98	9.284
Rather frequently	78.20	17.076
Usually	78.02	13.046
Often	74.84	9.661
Commonly	74.57	12.488
Fairly often	74.30	12.574
Fairly many times	67.50	17.366
Sometimes	46.70	11.881
Occasionally	45.80	14.882
Some of the time	43.95	13.915
Now and Then	39.95	15.449
To some degree	36.34	17.106
Once in a while	29.55	14.257
Fairly infrequently	23.86	14.659
Rather seldom	21.84	14.244
Not often	20.52	11.475
Infrequently	20.09	11.612
Seldom	19.95	14.344
Not very often	18.82	11.126
Very seldom	16.95	13.981
Very infrequently	13.96	15.758
Hardly ever	13.18	10.779
Rarely	12.30	9.433
Very rarely	11.82	16.527
Hardly at all	11.71	14.816
Almost never	8.27	6.770
Not at all	1.50	6.918
None of the time	.18	1.336
Never	.00	.000