

CONSTRUCTING EQUAL DISTANCE RESPONSE

ANCHORS

By

WILLIAM CAMRON CASPER

Bachelor of Science, Psychology
Brigham Young University
Provo, UT
1988

Master of Business Administration
Brigham Young University
Provo, UT
1990

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2013

CONSTRUCTING EQUAL DISTANCE RESPONSE
ANCHORS

Dissertation Approved:

Bryan D. Edwards

Dissertation Adviser

Kenneth K. Eastman

J. Craig Wallace

Robert A. Baron

ACKNOWLEDGEMENTS

A great many people have contributed to the completion of this dissertation and I owe my gratitude to all those people who have made this dissertation possible.

My personal gratitude is to my advisor Dr. Bryan D. Edwards, for always believing in me, even when I did not.

My personal gratitude is to Dr. Robert A. Baron, for reminding me that the joys of life can be found in a gritty diner, in the woodshop, and in the classroom.

My personal gratitude is to Dr. Kenneth K. Eastman, for his loyal support as the Head of the Management Department.

My personal gratitude is to Dr. J. Craig Wallace, for his quick wit, high intellect, and stellar example of the work ethic.

My personal gratitude is also extended to the above for their service on my dissertation committee. Thank you.

My personal gratitude is to Dr. James M. Pappas, for his extensive efforts as head of the Ph.D. program and to Dr. Wm. Matthew Bowler, for extending an invitation to join the OSU program.

My personal gratitude is to my cohort partner Corey Fox, for his strength and stability during many years of learning together.

My personal regards to my professors, the Spears School of Management and its employees, the other Ph.D. students, and graduate assistants that have made this dissertation possible. Thank you.

My personal regards is to my Research Methods class for work performed collecting recent data trends from JAP and AMJ.

Foremost and finally, my best and personal gratitude is to my wife, Tamra L. Casper, for all my family responsibilities she graciously accepted—thank you! Thank you for helping me through my *Via Dolorosa*.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: WILLIAM CAMRON CASPER

Date of Degree: May, 2013

Title of Study: CONSTRUCTING EQUAL DISTANCE RESPONSE ANCHORS

Major Field: MANAGEMENT

Abstract:

Summated rating scales are so ubiquitous in organizational research that they are used in most every context including agreement, similarity, frequency, amount, and judgment. Summated rating scales are defined by their usage of multiple questions or items in a “Likert-type” format. A long history of research on latent constructs has provided well-delineated guidelines for scale development (c.f., Hinkin, 1998), while there has been comparatively less research on the selection of the actual anchors used in the items. Constructing survey questions with equal distant properties is important if the researcher plans to analyze the data using common parametric statistics. As such, the primary objectives of my research are to a) determine the most common contexts for summated rating scale usage; b) determine the most commonly employed anchors selections within these contexts; c) empirically document the distance between the verbal anchors; d) suggest the optimal verbal anchors for each context based on the aforementioned research and commonly accepted number of anchors per item (5 or 7 choices). The resultant schemata will be a starting point for summated scale construction and will lead to better construction of equal distance anchors for summated rating scales in behavior science research.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION.....	1
II. REVIEW OF LITERATURE.....	11
III. METHODOLOGY	15
Gathering Initial Set of Anchors.....	16
SME Evaluation of Initial Set of Anchors	18
Pilot Study.....	18
Focus Groups	20
Final Study	22
IV. FINDINGS.....	23
V. CONCLUSION.....	37
REFERENCES	42
APPENDICES	46
A. Instructions Given to SMEs and Results	46
B. Comments from Pilot Study.....	55
C. Final Study	57
VITA.....	63

LIST OF TABLES

Table	Page
1. Number of Anchor Points Used in Scales Published in JAP and AMJ from 2005-2009	9
2. Categories in Recent Literature	16
3. Anchors Receiving Discard Votes	18
4. Standard Deviation of Means for Mean Values	19
5. Number of Anchors in Each Study	22
6. Agreement; Mean Values (graph), SD, and Overlaps	27
7. Similarity; Mean Values (graph), SD, and Overlaps	28
8. Frequency; Mean Values (graph), SD, and Overlaps	29
9. Amount; Mean Values (graph), SD, and Overlaps	30
10. Judgment; Mean Values (graph), SD, and Overlaps.....	31
11. Agreement; 5 and 7 Anchor Points	32
12. Similarity; 5 and 7 Anchor Points.....	33
13. Frequency; 5 and 7 Anchor Points.....	34
14. Amount; 5 and 7 Anchor Points.....	35
15. Judgment; 5 and 7 Anchor Points	36

CHAPTER I

INTRODUCTION

Summated rating scales are a frequently used measurement instrument in social science research to measure latent constructs (e.g., job satisfaction, commitment, psychological safety, performance, etc.). In constructing summated scales, individual items are pooled to measure the latent construct. The creation of summated rating scales intended to measure latent constructs is well documented (c.f. Hinkin, 1998; Spector 1992). However, not much research has deliberated on the word choices for the anchors (the potential answer choices are called anchors) in each item/question. Given that the anchors are the defining characteristics of summated rating scales, the lack of associated research is curious. Further, anchoring word choices are not benevolent characteristics of scales that can or should be assigned heuristically or haphazardly. To wit, equal distant intervals between anchors are important and the assumption of equal intervals between anchors is the basis for the parametric statistical analyses that are customarily applied to the data (Coombs, Dawes, and Tversky, 1970; Schriesheim & Novelli, 1989; Spector, 1976; Suppes & Zinnes, 1963; Wildt & Mazis, 1978).

Assigning anchors heuristically or haphazardly may affect the assumption of equal distance between anchors. Irregularly spaced anchors can affect the legitimacy of any analysis by adding systematic error (Spector, 1976). Further, Renesis Likert, who is widely considered the father of summated rating scales, stated that equal distance between anchors is critical (1932). In fact, a *condicio sine qua non* of parametric statistical analysis is the equal interval assumption. Without equal spacing, the ability to integrate the central limit theorem and extrapolate to the population from the sample is uncertain. Interval-level statistics, based on equal distance anchors, are desirable because they allow for the use of parametric statistics and application of central limit theorem. Social science analysis often requires the flexibility and power of interval level statistics. Although non-parametric statistics are available, parametric statistics allow for more flexibility, better understanding, and stronger conclusions regarding the data (Abelson and Tukey, 1970; Hensler & Stipak, 1979; Tufte, 1970; Labovitz, 1970). Further, in most studies, researchers regularly use parametric statistics to analyze latent constructs. Because of the clear superiority of interval-level statistics, data analysts will continue to use them even when measurements are unknown or not interval. Further, if the underlying distribution is normal, the measurement systems should reflect this underlying population distribution and interval statistics should be used. The lack of empirical data and guidelines with respect to appropriate anchor selection would seem to be a major gap in the summated rating scale development literature and any other applied research use of summated rating scales.

Given the common usage of parametric-based statistics to analyze summated scale ratings, it is curious that extant research does not fully address the issue of equally spaced anchor choices. For example, studies continue to select response categories which do not

have equal interval properties, or with unknown properties (Hanson & Rethans, 1980; Wildt & Mazis, 1978). Perhaps it is common practice for researchers to assume an interval number scale of measurement because the ordered anchors appear to be on an interval scale (e.g., anchors ordered into a Likert type scale seems to suggest equal interval spacing). However, this logic is flawed because the word choices have underlying numerical values may not be represented by application to a common number line. If the text anchors are not accurately represented on an interval scale, then the unknown distribution will not be accurately represented on the number line. For example, a common scale measuring relative frequency of behavioral expression might be represented by the anchors: never; rarely; sometimes; occasionally; always. The research question the current paper investigates involves determining the mathematical distances between these choices and presenting those choices to researchers in a useable fashion by answering questions such as, “ Is the distance between “sometimes” and “occasionally”, the same as the distance between “never” and “rarely”? Should “sometimes” and “occasionally” be used together? Analysis in the present manuscript indicates that “sometimes” and “occasionally” occupy practically the same point on the number line and should not be used together as choices in a scale, although they often are used together in practice. If the distance between anchors is unequal or uncalculated, researchers should not apply parametric distributions to these unknown distributions. If the distance between anchors is uncalculated, do researchers still apply parametric distributions to analyze the results? Often! At present, there is no source where survey developers can find or select anchors with known distances between them for summated rating scales construction. The current research study addresses this gap in the literature and provides real research assistance to those building surveys. The present research endeavors to create lists

of potential anchor choices and the distances between them, from which researchers can guide their selections in creating surveys, thus providing a basis for usage of interval-level statistics. Therefore, the practical question the present research addresses is how to best assign verbal choices that approach equidistance between anchors and thereby allow an interval scale of measurement in support of parametric statistical analysis.

To arrive at the research questions, further explanation is needed with respect to the assumption of equal distant anchor selection. Summated rating scale results are typically analyzed according to classical test theory. True score is the theoretical value of each participant on the construct. The observed score is the actual score from the measurement process. True scores cannot be directly observed but are inferred from the observed scores. If there were perfect reliability and validity, the observed score and true score would be equal. The observed score has two parts, the true score and random error. Observed score is usually given by the following formula:

$$O = T + E$$

O is the observed score; T is the true score; and E is the random error. If the observations are truly random, than E should average approximately zero because the errors are from a population where the mean is zero. In a summated rating scale, each item is intended to be an observation of the construct. So, every item represents an itemized assessment of the true score. When the mean of the individual items is calculated, the errors are assumed to average approximately zero and result in an estimate of the true score (Spector, 1992).

Errors in measurement are inverses of reliability and the larger the error component, the worse the reliability. Single items are generally not used to measure a construct because they are notoriously unreliable (Spector, 1992). With multiple items, sometimes the errors

will inflate and sometimes they will deflate the observed score. The errors will tend to average out, giving a more accurate and reliable measurement. So, in theory, increasing the number of good items will increase the reliability of a measurement. Hence, scales are created for construct measurement rather than single items. Classical test theory (with the possible exclusion of Generalizable-theory, which includes bias (c.f., Shavelson, Webb, & Rowley, 1989)) oversimplifies by combining random and systematic error (SE); however, the systematic error can be separated from random error as illustrated in the following formula:

$$O = T + E + SE$$

Systematic errors are methodological influences on observed scores that are not reflective of the true score. They do not come from populations with means of zero, and therefore do not average out with multiple items. Systemic errors come from other traits (other than the desired construct) that influence the observed score. Presenting anchors which are not equal distant creates systemic bias in responses.

A *measurement rating scale* (this is not analogous to summated rating scale) is a method requiring a rater to value an object. According to Stevens (1946), rating scales can be classified as categorical/nominal, ordinal, interval, or ratio scale. Stevens proscribes statistical techniques for the different measurement scales based on their attributes. Classical test theory assumes that most psychological constructs conform to the standard normal distribution and that this theoretical distribution can only be captured on an *interval* level scale of measurement

I therefore pose the following research question:

1. Which verbal anchors should be selected for summated rating scale construction to create equal distance between anchors so interval level statistics can be applied?

There are several categories or situations that are relevant to anchor selections in psychometrics. For example, the same anchors cannot be used to measure agreement and judgment. Therefore, the second research question becomes:

2. Which contextual categories represent summated scale usage in management research?

I chose to consider the most common contexts reported in the literature: agreement, similarity, frequency, amount, and judgment. According to my review of the extant literature using 5 years of research from both *Journal of Applied Psychology* and *Academy of Management Journal*, these 5 contexts cover about 95% of the usage of summated rating scales in the research.

Agreement items ask participants to indicate the extent to which they might agree or disagree with an item. Agreement response choices are typically bipolar and symmetrical around a neutral point, and include the magnitude of the belief. Agreement item stems are declarative in nature, so that the participant can indicate their level of agreement (e.g., I enjoy my job.)

Similarity (like me) asks participants to indicate the level of “likeness” or similarity to themselves. Similarity response choices are typically bipolar and symmetrical around a neutral point, and include the magnitude of the belief. The choices for these items are declarative and the participant is usually asked if the statement describes them. (e.g., to what extent does the following describe you: I think about my work when I am at home.) Although

similarity can be considered as a type of *agreement*, it is used as a separate category in the extant literature and therefore merits separate attention.

In the organizational literature we also rely heavily on judgments of relative *frequency*. Frequency asks for judgments of how often each item has, should, or will occur. For example, a supervisor may judge how often an employee engages in a specific behavior and a measure of attitudes (e.g., job satisfaction) would ask respondents to judge how often an item stem reflects his/her feelings about the attitude. Although some researchers prefer numeric anchors such as “once per day” to measure absolute frequency (Newstead & Collis, 1987), most use verbal anchors for relative frequency. Frequency response items are generally unipolar with an absence of the attribute at one end of the scale, and a fullness of the attribute at the other end of the scale. They are commonly used to measure personality where participants are asked to provide information on how often a subject engages in a certain behavior. They are also used in environmental measures to indicate how often an event might occur. An example of a frequency item would be, “You speak positively of your job with someone outside of work.”

Amount response anchors are similar to frequency, in that they measure the amount of an item, as perceived by the participant. Amount response choices are generally unipolar with an absence of the item at one end, and a fullness or completeness of the item at the other pole. Complexity, for example, can be measured from the perspective of amount or how much complexity is present (i.e. “a lot” or “none”).

Judgment response choices ask participants for an evaluative rating of each item. These choices are along a good/bad dimension. They typically range from the positive (excellent) to the negative (terrible). Judgment choices can measure attitudes, persons,

places, things, events, behaviors, and performance; however, the current research does not attempt to present an exhaustive list of anchors that could be used in performance evaluations. An evaluative scale might include, “What is your opinion of your job?”

Therefore, our research question expands to include properly selecting anchor choices based on the category of the survey. (For example, the same anchors cannot be used for situational constructs exploring “frequency” as those trying to understand “agreement” and so forth.)

As part of scale development, the number of anchor points along with the verbal anchors themselves, are constructed. The number of rating points on a scale can vary from 2 to 100 or more. According to the extant research, the optimal number of anchor points may still be unresolved (Preston & Colman, 2000) but there are several useful points in the literature that can guide our research. Therefore, the final research question will be:

3. How many anchor choices should be used in creating summated scales?

Although some research asserts that the number of anchor points is irrelevant (c.f. Schutz and Rucker, 1975) that conclusion is not in line with other studies. Research indicates that at least 2 or 3 points are necessary. Garner (1960) suggested that maximum information is obtained by using more than 20 points. Others suggest that information is maximized by using 6 or 7 points and little extra information, or reliability, is gained by using more than 7 points (Bendig, 1954; Green & Rao, 1970; Preston & Colman, 2000; Cicchetti, Showalter, & Tyrer, 1985). Symonds (1924) suggested early on that inter-rater reliability is optimized by the use of 7 points. Miller (1956) further supported the ideal of 7 points by suggesting that the human mind has a span of apprehension ability limited to 7 (plus or minus 1 or 2), on judgments of unidimensional judgments. McKelvie (1978)

suggests that the number of points be limited to 5 or 6. Although not unanimous, there is a general consensus in the literature that the optimal number of points is 5 to 7. (Champney & Marshall, 1939; Cox, 1980; Linacre, 2002).

In current practice, most rating scales use five or seven point response scales (Bearden, Netemeyer, & Mobley, 1993; Peter, 1979; Preston & Colman, 2000; Shaw & Wright, 1967). Further, I conducted a review of actual anchors used from 5 years of research and published *Journal of Applied Psychology* and *Academy of Management Journal*. The mode and median number of scale points is 5 (point scales), the mean is 5.7. Specifically, 56% of studies in those journals use 5 point scales and 33% use 7 point scales, and the remaining results are varied. The results of this research are presented in Table 1:

Table 1

Number of Anchor Points Used in Scales Published in JAP and AMJ from 2005-2009

Number of Anchor Points	Number of Scales	Percentage
2	3	0%
3	15	1%
4	60	4%
5	949	56%
6	50	3%
7	551	33%
8	8	0%
9	30	2%
10	9	1%
11	11	1%
Total	1687	

Note: Data taken from 413 articles in *Journal of Applied Psychology* and *Academy of Management Journal* for 5 years period (2005 – 2009) that contained summated rating scales.

In order to make a decision on the optimal number of anchor points, I defer to three summary points presented by Cox (1980). First, scales with two or three points are not

capable of transmitting much information and tend to frustrate participants. Second, the ongoing returns of using more than 9 points are minimal. Third, an odd number of points is preferable to an even number.

Based on the summary of the literature (Cox, 1980) and my own analysis, it seems appropriate to select anchors for 5 and 7 point questions, which represent the determination of research and the actual usage from the literature.

CHAPTER II

LITERATURE REVIEW

More than 30 years ago, there was group of studies on the mechanics of summated scale development, which presented limited and sometimes conflicting information on verbal anchor selection (Schriesheim and Schriesheim, 1974; Bass, Cascio & O'Connor, 1974; Spector, 1976). Schriesheim & Schriesheim (1974) indicated that a number of scales, including the popular Ohio State leadership scales, use response categories of *Always*, *Often*, *Occasionally*, *Seldom*, and *Never* and it is assumed that they capture interval scale data. Schriesheim & Schriesheim questioned the validity of the intended interval nature of the data, and attempted to establish a better set that was more equally spaced. They surveyed 200 students to determine if the distances between the verbal anchors were reasonably equal when the anchors were ordered by median frequency values. Schriesheim & Schriesheim indicated that the anchors were not equidistant when ordered thus, and completed an additional study to corroborate their finding. Using the Thurstone case III paired comparison of complete ranks to scale 18 frequency expressions, Schriesheim & Schriesheim found that a common list of five anchors (*Always*, *Often*, *Occasionally*, *Seldom*, and *Never*) did not achieve equal distance interval between anchors.

Using mean frequencies, Schriesheim & Schriesheim recommended a measurement scale that used *Always*, *Very often*, *About as often as not*, *Seldom* and *Never*. Although this important study provided direction for researchers, the recommended anchors are not clear and seem archaic to the modern student. For example, the midpoint of “About as often as not” can be criticized for lack of clarity and perhaps for this reason, has not been adopted by researchers in the field. Further, the authors made a call to academia, yet unanswered, to replicate the survey across different scaling methods and different samples.

A study from Bass, Cascio, and O’Connor (1974) investigated whether the relative importance of a question would affect anchor score values used in the contextual situations involving frequency and amount. They suggested that if the contextual situation changed the meaning of the anchor, then a common list of anchors could not be created. Notably, Bass et. al’s research confirmed that context did not change the meaning of the anchors because obtained invariance across contextual situations involving important (war-related) versus unimportant (rainfall in another part of world) situations. Consequently, Bass et al. proposed that a common list of anchors could be used to develop scales with equal distance properties within context. This supposition is a foundation for the current paper and is supported by others (Cliff, 1959). Bass et al. used magnitude estimation technique to find the numerical equivalents of 39 expressions of frequency (*never* to *always*) and 44 expressions of amount (*none* to *all*) across 3 age and 3 occupation levels. Bass et al. further derived reliable means and standard deviations from the data. Equal distance anchors were extrapolated for 4 to 9 point anchor scales. Bass et al. presented the findings using an overlap statistic. For a five point scale, Bass

and colleagues recommended a different (vs. Schriesheim and Schriesheim) set of five anchors: *Always, Very often, Fairly many times, Occasionally, and Never*. Bass and colleagues challenged future researchers to demonstrate that differences in context could invalidate standard lists of anchor choices. In my research, I will expand the contextual situations from the two that Bass et al. chose to use, to five (*Frequency and Amount plus Agreement, Similarity, Judgment*) which is more inclusive of the types of scales used in the extant literature. Further, the recommended phrase *Fairly many times* is archaic or perhaps merely awkward, and should be removed from the choices, as indicated by 70% of the subject matter experts of the present study.

A third study (Spector, 1976) expanded the parameters of Bass et al. (1974) by including anchors associated with “judgment” and “agreement” along with “frequency”, but did not measure “amount.” Spector selected anchors from Shaw & Wright’s (1967) classic book, *Scales for the Measurement of Attitudes*.

At least four of the anchors were different from both earlier studies and therefore preclude direct comparison. Furthermore, the study was limited in that it included only 13 anchors for frequency. Spector’s study is limited to anchors on unipolar scales. For example, frequency, similarity, and amount are unipolar in nature, but agreement and judgment are bipolar. Spector chose the normalized rank method for complete ranks suggested by Guilford (1954).

In sum, the primary objectives of the present study are to (1) calculate mean values of anchors to establish equidistant measurement properties; (2) establish contextual categories; and (3) determine optimal guidelines for number of anchor choices such that a practical list of anchors with equidistant properties in each contextual category can be evaluated and presented for research purposes.

CHAPTER III

METHODOLOGY

Gathering initial set of anchors

The objectives of the present study are essentially to create a list of equally-spaced anchors used for different situational contexts. It was necessary to begin with an initial comprehensive set of anchors that could possibly be used in survey research. Understandably, any effort to obtain a comprehensive list of anchors requires judgment calls in terms of which are most acceptable. One challenge is that there are thousands of adjectives in the English language and hundreds of constructs measured in social science. However, many adjectives would not be used as survey anchors and many surveys use the same set of anchors across constructs. Therefore, the decision was made to limit anchor selection to only those that had been used in prior scale/anchor development research and in recent published studies for the most predominant categories (i.e. agreement, similarity, etc.).

First, I obtained the anchors listed in the previous studies investigating anchor selection and equal distance between anchors (e.g., Jones & Thurstone, 1955; Wright & Shaw, 1967; Bass, Cascio, & O'Connor, 1974; Schriesheim & Schriesheim, 1974; Spector, 1976). Second, I reviewed articles in two of the leading management journals,

Academy of Management Journal and *Journal of Applied Psychology* from 2005 to 2009 and recorded both the anchors and the contextual categories (agreement, similarity, etc.) The resulting data suggested that 95% of the surveys in *Academy of Management Journal* and *Journal of Applied Psychology* from 2005 to 2009 were represented by 5 contextual categories (agreement, similarity, amount, frequency, and judgment). The data from these two journals are presented in Table 2. Other categories identified but not included due to minimal usage (i.e., in less than 1% of the studies) were accuracy, importance, likelihood, satisfaction, preference, and veracity. Anchors were categorized and limited to these 5 categories.

Table 2

Categories Analyzed in Recent Literature

	Agreement	Similarity	Frequency	Amount	Judgment
% of Studies	57%	1%	12%	18%	7%

Note: Academy of Management Journal, volumes 48-52 and *Journal of Applied Psychology*, volumes 90-94.

SME evaluation of initial list of anchors

The initial list of anchors was quite large and unmanageable for the purposes of the current study. For example, the list of agreement anchors was 34; similarity, 18; frequency, 52; amount, 46; and judgment, 55. Given limits to human information processing, these lists were too long to ask participants to hold them in memory to accurately rank them from most to least. Also, there were anchors that were rarely used (e.g., "a full amount of," fairly many times"). To help guide decisions of which anchors to omit from the survey, I asked SMEs (N=10) for their expert opinions. They were provided the long lists of anchors and were asked to "discard" those anchors they thought

should be removed as anchor selections. The remaining anchors would therefore be the most likely anchors to be used on surveys (Appendix A). Professors and PhD candidates in the management department at Oklahoma State University participated as SMEs. SMEs were presented the initial list of anchors in Qualtrics for confirmation that the list was indeed comprehensive and additionally, to assist with reducing this list length. Participants were asked to list any important anchors that may be missing. SMEs were also asked to judge if an anchor seemed to be confusing, archaic, or redundant and if so, SMEs were to recommend discarding the anchor. I omitted those anchors that 40% or more of SMEs voted to discard which resulted in 50% of the anchors being recommended for elimination from the initial list. A few additional items were removed based on comments from SMEs suggesting a strong opinion to remove items, even though the item may not have received additional votes. Some items were added based on SME input including “strongly agree” and “strongly disagree”. A few words were removed that had been inadvertently listed on more than one occasion (duplicates).

Using this process, I believed that the lists still too long for ranking purposes. For example, the list of the judgment anchors still contained 54 items. Therefore, I conducted a pilot study to guide decisions on list length and which anchors to include in the study.

Table 3

Anchors Receiving Discard Votes at Each Percentage Level

% Voting to Discard (Percentage Level)	Number of Anchors receiving votes at each Percentage Level
0	19
10	31
20	41
30	32
40	26
50	22
60	9
70	14
80	9
90	1
100	0

Note: There were 19 potential anchors that “0” participants voted to discard; there were 31 anchors that 10% voted to discard and so forth.

Pilot Study

The reduced list of anchors was then presented (via Qualtrics) as a pilot study to two undergraduate business classes at Oklahoma State University (N = 107, 81% Caucasian, 11% Native American, 6% African American, 5% Asian, and 4% Hispanic; 94% had “some college”; 44% employed part-time and 39% employed full-time). The participants were presented with 5 different lists of potential anchors one list at a time. Each of the lists had 18 to 54 words. The participants were asked to rearrange the words in the lists into a seriatim ordered format from most to least.

The purpose of the pilot study was to verify that the format was efficient and user-friendly; further, the pilot study was conducted to verify that the lists were short enough for participants to rank accurately. The results of the pilot study revealed that the two categories with longer lists (“frequency” number of anchors = 52 and “amount” number of anchors = 46), were still too long. First, comments on the survey by participants

indicated that those two lists were too long. Half (50%) of the participants registered a comment, and of those, 30% indicated that the longer word choice lists were too long to accurately assess. Example comments were “some lists were a little long for good attention span”, and “Limit the number of options to no more than 15 per questions” were two actual responses (Appendix B).

Second, the SDs for the mean values of the longer lists seemed to indicate that the participants struggled to rank them consistently. I averaged the individual SDs within each contextual category. Both of the longer lists (Frequency and Amount), had much larger SD than the shorter lists (Table 4). For “frequency”, a 7.6 SD indicated that a word selection would be ranked in a range as broad as 15 ranks wide, 68% of the time. For the purposes of this study, that range was unacceptable.

Table 4

Means of the SD of the Mean Values

	Pilot Study	Final Study
Agreement	1.9	1.2
Similarity	2.4	1.6
Frequency	7.6	1.0
Amount	6.8	1.7
Judgment	2.7	1.7

Focus Groups Next, I conducted focus groups because I wanted assistance in interpreting the data from the pilot survey (especially the comments), understanding which lists were too long, verifying which anchors to eliminate, and searching for other improvements to the method. The focus groups were the same two management classes at OSU ($N=50$) that completed the survey described above. Some of the students had participated in the survey, and all had the opportunity to have done so for extra credit. I reviewed the survey by showing it on the ELMO overhead projector. Focus group and survey feedback resulted in a few changes in the process. First, the rearrange feature used to move the words within the lists did not function well on tablet or other handheld devices. Second, the rearranging process was confusing and needed to be improved. Finally, the lists (especially Frequency and Amount) were “too long.” There were also several words in each of the lists that were identified as being redundant or archaic by the focus groups.

The rearrange feature allowed the participant to drag a word within the list and basically “rearrange” the word list. This did not work on handheld devices for technical reasons. Further, it encouraged participants to leave words in the same order in which they originated, since it was not required that any words be moved. If an individual did not understand the task, they might leave the words in the original order or close to the original order. I considered this unacceptable, especially after reviewing the data and noting that almost 10% of the results were in the same order as they were presented, with no changes. To ameliorate the problem, I created a “drag and drop” process, which required participants to move a word from the original list to a different list as part of the ranking process. As words were removed from one list, they were added to the second list until the first list became empty. This was an improvement on the first process and was evaluated by the focus groups as an improvement. (I had generated the idea and process in advance of meeting with the focus groups, based on comments received from the pilot study.)

As the feedback indicated that the lists should be shortened, I went back to the literature and to the SME data, and eliminated choices that seemed to be archaic or less useful based on SME comments and actual usage in the literature. Some lists had to be shortened more than others, because by their very nature, they were longer than their counterparts. For example, less words are used in the total population for similarity items than are used in the population of words for amount. I limited the number of choices to 15 or less. Since the lists could be segregated, into halves when ordering, a list of 15 items required that only 7-8 were in serial memory of the participant at one time. Miller (1956) supported the ideal of 7 choices by suggesting that the human mind could hold up

to 7 (plus or minus 1 or 2) words, on judgments of unidimensional judgments, in current memory. The number of items for the final survey is shown in Table 5.

Table 5

Number of Anchors in Each Study

	Agreement	Similarity	Frequency	Amount	Judgment
Potential anchors	36	18	52	46	54
Interim Pilot Study	13	12	34	32	15
Final Presentation of Anchors	13	10	13	15	10

Final Study

The final study was administered via Qualtrics to students from business courses at Oklahoma State University in exchange for extra course credit. The participants were 301 males and females drawn from master of business students ($n = 107$) and undergraduate students ($n = 194$). A little over 50% worked part-time, 20% were employed full-time, and 29% were not employed. Respondents were 69% Caucasian, 3% African American, 4% Hispanic, 18% Asian, and 5% Native American.

Participants were provided a link to a survey via an email invitation to participate. Participant responses were anonymous. Participants were given the list of potential anchors for each of the five categories (agreement, similarity, frequency, amount, and judgment). Participants seriatim ranked the anchors in each of the five categories. The survey is presented in Appendix C.

CHAPTER IV

FINDINGS

Once the rank order data was obtained, scale values were derived using normalized rank method of complete ranks set forth in Guilford (1954, Chapter 8; Spector, 1976). This method was selected because it can deal with a great number of relative stimuli easily. The method of complete ranks also forces participants to make discriminatory choices about each pair and therefore provides as much data as is possible to attain.

The ratings assigned to the anchors were converted to rank values. These rank values are denoted as R_i and for calculation purposes are listed in reverse order of the actual rank r_i . They are related by the equation where n is the number of participants:

$$R_i = n - r_i + 1$$

A rank value (reverse of actual rankings) was individually assigned to the variables in each category via spreadsheet calculations. The reverse ordering is critical because in the raw rankings, the top answer is typically given the score of “1”. However, this cardinal number does not provide information on the number of categories being ranked. If there are 15 choices, and the top choice is reset to “15,” more information is known about the results. The rank values were then transformed to mean values by way of Guilford’s common C scale value (1954, Table M, pg. 577). To calculate and standardize the results, several Excel tables were created. In Tables 6-10, the resultant mean values are presented. The C scale values are used to compute means, SDs, mean differences, SD of mean difference, and the Tilton overlap statistic. The mean values represent the relative placement of the anchor on a number line; SD is an indicator of spread or consistency for participants responses; mean differences are the distances of the spaces between chosen anchors; SD of the mean differences represents the consistency (or lack of) between the spaces and is a measure of goodness of the equal distance quality of the anchors; and the Tilton overlap shows the separation of the anchor from the contiguous anchor distribution.

Using the derived mean scale values, anchors can be selected that have equal distance properties. For example, using Table 6, I might begin a 5 point agreement scale with the words “Strongly disagree” (mean value = 1.9) and continue with “Disagree” (mean value = 3.4), “Neither agree nor disagree” (mean value = 4.7), “Moderately agree” (mean value =6.1) and “Very much agree (7.5). The Mean Difference shows the mean distance between the choices, while the Mean SD indicates its consistency and is included for each of the recommendations in Tables 11-15. For each of the contextual

categories, the best fitting (based on equal distances between anchors) anchors were selected and the results are shown in Tables 11-15. These tables present 5 and 7 point anchor choices for each contextual category (agreement, similarity, frequency, amount, and judgment) respectively and show the percentage overlap in distribution between adjacent selected scale points. The expressions chosen to represent the various points on the scales are optimal in that they are the closest equal distant mathematical values. For the recommended 5 point agreement scale, the Mean Difference is 1.4 and the Mean Difference SD is .08 indicating that each point averages 1.4 units from the next and does not deviate substantially from this average.

The distribution overlap shows that choices may sometimes be very close on the number line. For example, “somewhat disagree” and “slightly disagree” are so close that they share 96% of their distributions whereas “neither agree nor disagree” shares only 46% with “slightly agree.” In choosing alternatives with equal spacing, the overlap can be considered as a secondary data point to indicate spacing between choices. Choices with unique distributions are spaced far enough (as possible) on the number line to avoid confusion from participant.

As shown in Tables 6-10, the recommended anchors approach equal spacing. One could also choose a different set of anchors and use the data in Tables 6-10 to construct a scale. For example, some constructs measured with relative frequency would never use the absolutes: never and always (as was suggested in Table 8 of the present study). Therefore, I have included the research results in their entirety (Tables 6-10) so that researchers may evaluate past, present, and future anchor constructs based on the data. The Tilton overlap can be computed between any two points using the following formula

$$\frac{M1 - M2}{\left(\frac{\sigma1 + \sigma2}{2}\right)}$$

and by looking up the corresponding per cent overlap on provided tables in Alf & Abrahams (1968).

The “Tilton” overlap statistic was computed to show the index of the amount of separation between chosen scale points. As the scale fineness or closeness increases (number of points), so does the distribution overlap (increase). With fewer points, the distribution, conversely, is coarser. Tilton’s (1937) distribution overlap statistic was calculated to provide a table of the separation between neighboring scale points (Alf & Abrahams, 1968; Bass, Cascio, & O’Connor, 1974). Typically of non Tilton overlap statistics, scores are reported as overlap only when the scores of one distribution exceed the medium score of the second. However, this type of overlap measurement is unsatisfactory as its maximum value of 50% could describe 100% overlap (Symonds, 1930). The Tilton overlap is suggested as an explicit measure to indicate the area common to two distributions. It is the percentage of scores that could belong to or match scores in the next group. Ideally, as the anchors with the equal distant separation are identified, distribution overlaps will be minimized. In selecting optimal anchor points for recommendation, I considered the overlap statistic when other factor s(for example, mean SD of mean differences) were equal. For example, when selecting the optimal anchor points, my first selection criteria was equal distant properties, and the next was the overlap statistic.

Table 6

Agreement

Ordered Mean Values, Standard Deviation, and Overlaps

Verbal Anchors	*Mean Value	SD	**Distribution Overlap
Strongly Disagree	1.9	1.5	94%
Very much disagree	2.5	1.1	50%
Disagree	3.4	1.4	83%
Moderately Disagree	3.9	0.9	60%
Somewhat Disagree	4.3	1.2	96%
Slightly Disagree	4.5	1.1	66%
Neither agree nor disagree	4.7	1.6	46%
Slightly Agree	5.5	1.2	88%
Somewhat Agree	5.7	0.9	96%
Moderately Agree	6.1	1.1	88%
Agree	6.8	1.1	95%
Very much agree	7.5	1.1	88%
Strongly Agree	8.2	1.5	

Note:

*The Mean Values are presented visually in the following graph.

**The percentage distribution overlap represents the overlap in distribution between the referent anchor and the anchor point directly below it. For example, there is a 94% overlap in distributions between “Strongly agree” and “Very much disagree”.

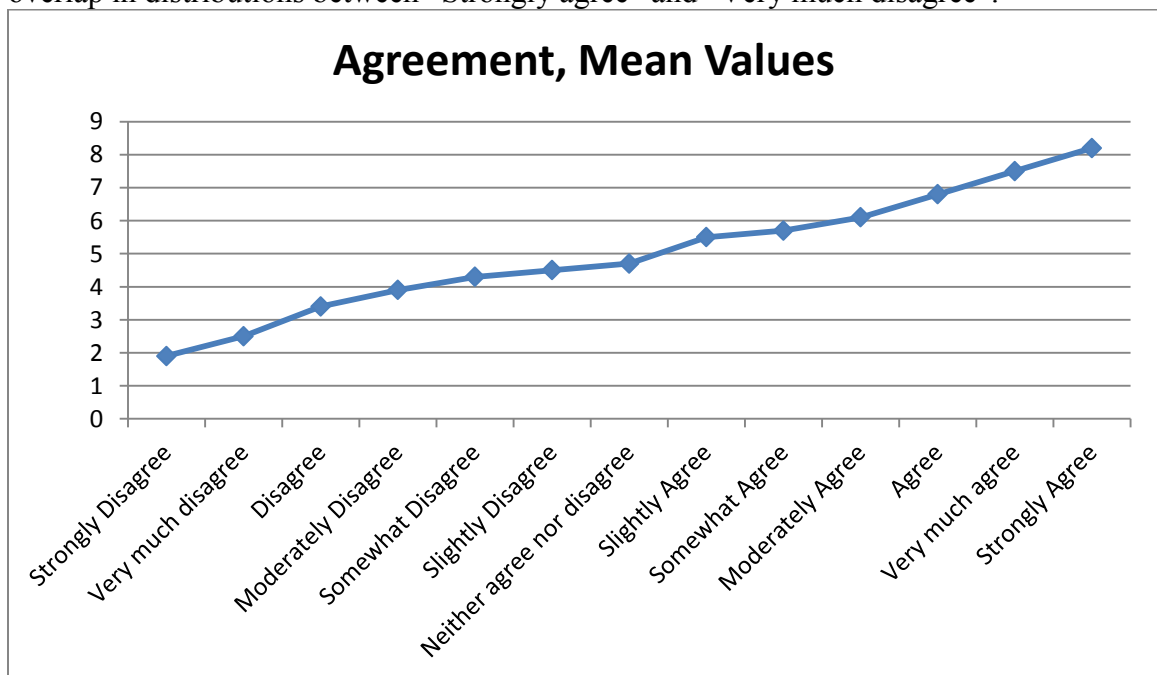


Table 7

Similarity

Ordered Mean Values, Standard Deviation, and Overlaps

Verbal Anchors	Mean Value	SD	Distribution Overlap
Not at all like me	2.4	1.4	76%
To no extent like me	2.9	1.6	70%
Neither like me nor unlike me	3.2	1.3	54%
Slightly like me	4.3	1.5	92%
A little like me	4.4	1.6	95%
Somewhat like me	4.6	1.3	60%
Moderately like me	5.0	1.1	49%
To a large extent like me	6.1	1.5	92%
Very much like me	6.3	1.7	92%
To a great extent like me	6.4	1.5	64%
Extremely like me	6.9	1.5	94%
Identical to me	7.4	2.6	

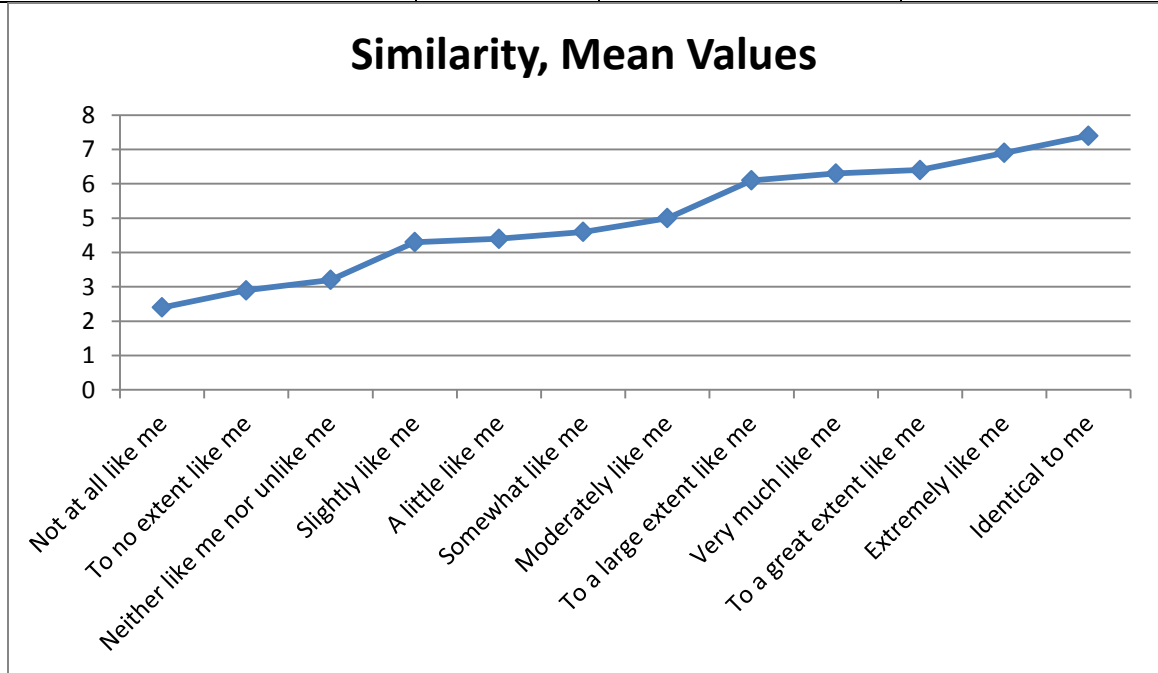


Table 8

Frequency

Ordered Mean Values, Standard Deviation, and Overlaps

Verbal Anchors	Mean Value	SD	Distribution Overlap
Never	2.1	0.7	52%
Almost never	3.1	1.0	68%
Very Infrequently	3.9	0.9	58%
Infrequently	4.0	0.8	40%
Occasionally	5.1	1.0	84%
Sometimes	5.2	1.1	50%
Frequently	6.0	1.0	84%
Most of the time	6.3	1.3	88%
Very Frequently	6.4	1.4	60%
Always	7.7	1.2	

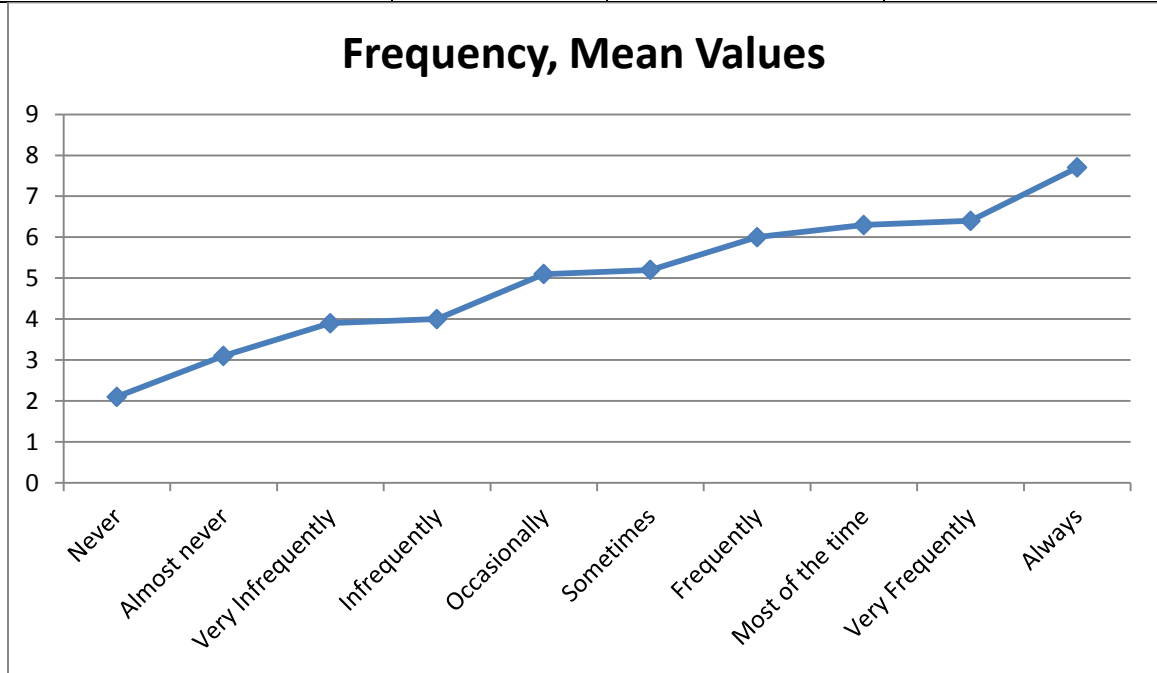


Table 9

Amount

Ordered Mean Values, Standard Deviation, and Overlaps

Verbal Anchors	Mean Value	SD	Distribution Overlap
None	2.2	2.3	100%
Not at all	2.5	1.6	58%
Hardly any	3.6	1.5	90%
Very little	3.9	1.6	86%
Not much	4.1	1.5	65%
A small amount of	4.8	1.4	80%
A little of	5.0	1.3	64%
Some	5.1	1.2	66%
Very much	6.1	2.9	99%
A large amount of	6.2	2.3	100%
To a large extent	6.2	2.7	90%
A lot of	6.4	1.4	1%
All	8.9	0.6	

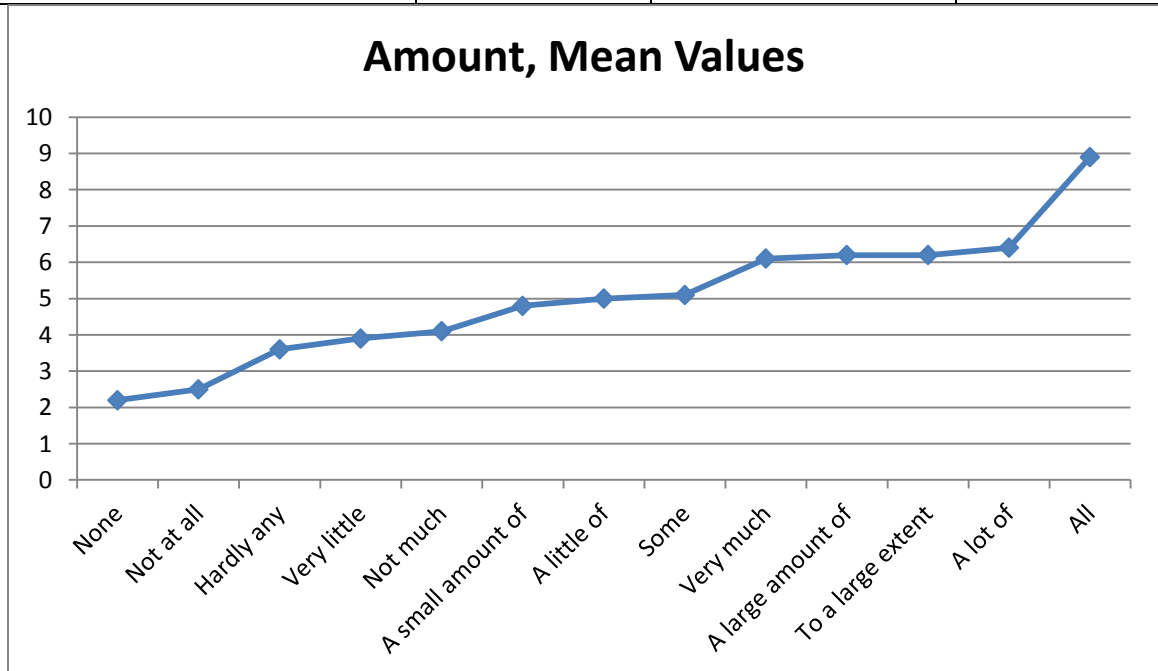


Table 10

Judgment

Ordered Mean Values, Standard Deviation, and Overlaps

Verbal Anchors	Mean Value	SD	Distribution Overlap
Horrible	1.7	1.3	88%
Terrible	2.0	1.9	55%
Bad	3.6	1.6	90%
Inferior	3.8	2.8	95%
Unsatisfactory	3.9	1.8	100%
Poor	4.0	1.5	40%
Mediocre	5.0	1.9	95%
Passable	5.1	2.0	69%
Average	5.5	1.6	100%
Decent	5.5	2.3	94%
Fair	5.6	2.0	75%
Satisfactory	6.1	1.8	69%
Good	6.6	1.4	56%
Excellent	8.2	1.6	92%
Outstanding	8.5	1.0	

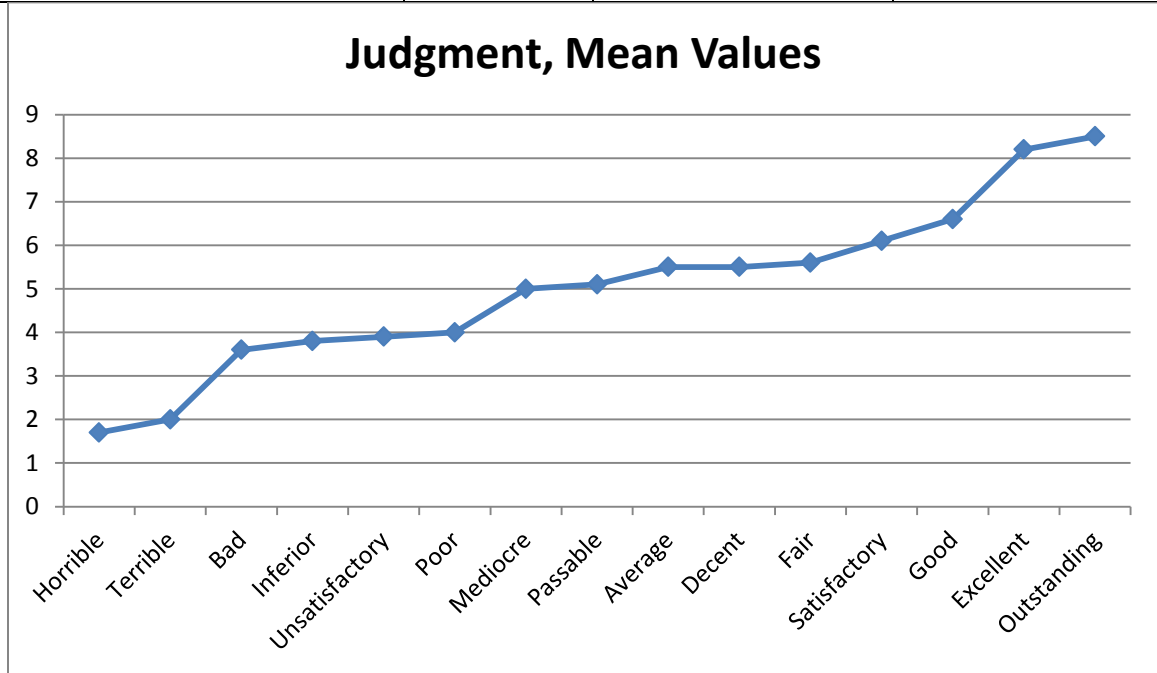


Table 11

Anchor Points of Agreement based on Equal Distance Intervals

(Overlap)

<i>5 Point Scale</i>	<i>7 Point Scale</i>
<i>*Mean Difference 1.4</i>	<i>Mean Difference 1.05</i>
<i>*Standard Deviation 0.08</i>	<i>Standard Deviation 0.43</i>
<i>Strongly disagree</i> <i>(51%)</i>	<i>Strongly disagree</i> <i>(51%)</i>
<i>Disagree</i> <i>(33%)</i>	<i>Disagree</i> <i>(33%)</i>
<i>Neither agree nor disagree</i> <i>(24%)</i>	<i>Moderately disagree</i> <i>(15%)</i>
<i>Moderately Agree</i> <i>(30%)</i>	<i>Neither agree nor disagree</i> <i>(22%)</i>
<i>Very Much Agree</i>	<i>Moderately agree</i> <i>(31%)</i>
	<i>Agree</i> <i>(58%)</i>
	<i>Strongly agree</i>

**Note: The Mean Difference represents the mean of the differences between scale mean values of the anchors shown in each list. The accompanying standard deviation is the standard deviation of the Mean Difference. If all the numbers were perfectly evenly spaced, the standard deviation would be zero. Hence, the lower the standard deviation, the closer the distances are to being evenly spaced. In these cases, smaller is better.*

Table 12

*Anchor Points of **Similarity** based on Equal Distance Intervals*

(Overlap)

<i>5 Point Scale</i> <i>*Mean Difference 1.25</i> <i>*Standard Deviation 0.42</i>	<i>7 Point Scale</i> <i>Mean Difference .83</i> <i>Standard Deviation .27</i>
<i>Not at all like me</i> <i>(55%)</i>	<i>Not at all like me</i> <i>(55%)</i>
<i>Neither like me nor unlike me</i> <i>(14%)</i>	<i>Neither like me nor unlike me</i> <i>(14%)</i>
<i>Moderately like me</i> <i>(46%)</i>	<i>A little like me</i> <i>(32%)</i>
<i>Very much like me</i> <i>(92%)</i>	<i>Moderately like me</i> <i>(49%)</i>
<i>Identical to me</i>	<i>To a large extent like me</i> <i>(86%)</i>
	<i>Extremely Like me</i> <i>(96%)</i>
	<i>Identical to me</i>

Table 13

*Anchor Points of **Frequency** based on Equal Distance Intervals*

(Overlap)

<i>5 Point Scale</i>	<i>7 Point Scale</i>
<i>*Mean Difference 1.40</i>	<i>Mean Difference 0.91</i>
<i>*Standard Deviation 0.28</i>	<i>Standard Deviation 0.37</i>
<i>Never</i> <i>(18%)</i>	<i>Never</i> <i>(52%)</i>
<i>Very infrequently</i> <i>(7%)</i>	<i>Almost never</i> <i>(32%)</i>
<i>Occasionally</i> <i>(28%)</i>	<i>Infrequently</i> <i>(39%)</i>
<i>Most of the time</i> <i>(50%)</i>	<i>Occasionally</i> <i>(74%)</i>
<i>Always</i>	<i>Frequently</i> <i>(84%)</i>
	<i>Very frequently</i> <i>(81%)</i>
	<i>Always</i>

Table 14

*Anchor Points of **Amount** based on Equal Distance Intervals*

(Overlap)

<p><i>5 Point Scale</i></p> <p><i>*Mean Difference 1.68</i></p> <p><i>*Standard Deviation 0.56</i></p>	<p><i>7 Point Scale</i></p> <p><i>Mean Difference 1.11</i></p> <p><i>Standard Deviation 0.80</i></p>
<p><i>None</i></p> <p><i>(22%)</i></p>	<p><i>None</i></p> <p><i>(22%)</i></p>
<p><i>Hardly any</i></p> <p><i>(12%)</i></p>	<p><i>Hardly any</i></p> <p><i>(12%)</i></p>
<p><i>Some</i></p> <p><i>(42%)</i></p>	<p><i>Not much</i></p> <p><i>(24%)</i></p>
<p><i>A lot of</i></p> <p><i>(17%)</i></p>	<p><i>Some</i></p> <p><i>(65%)</i></p>
<p><i>All</i></p>	<p><i>To a large extent</i></p> <p><i>(94%)</i></p>
	<p><i>A lot of</i></p> <p><i>(17%)</i></p>
	<p><i>All</i></p>

Table 15

*Anchor Points of **Judgment** based on Equal Distance Intervals*

(Overlap)

<p><i>5 Point Scale</i></p> <p><i>*Mean Difference 1.55</i></p> <p><i>*Standard Deviation 0.10</i></p>	<p><i>7Point Scale</i></p> <p><i>Mean Difference 1.03</i></p> <p><i>Standard Deviation 0.51</i></p>
<p><i>Terrible</i></p> <p><i>(57%)</i></p>	<p><i>Terrible</i></p> <p><i>(57%)</i></p>
<p><i>Bad</i></p> <p><i>(29%)</i></p>	<p><i>Bad</i></p> <p><i>(79%)</i></p>
<p><i>Mediocre</i></p> <p><i>(16%)</i></p>	<p><i>Poor</i></p> <p><i>(40%)</i></p>
<p><i>Good</i></p> <p><i>(58%)</i></p>	<p><i>Mediocre</i></p> <p><i>(37%)</i></p>
<p><i>Excellent</i></p>	<p><i>Satisfactory</i></p> <p><i>(69%)</i></p>
	<p><i>Good</i></p> <p><i>(56%)</i></p>
	<p><i>Excellent</i></p>

CHAPTER V

CONCLUSIONS

The present study demonstrated a method by which anchors could be selected based on equal distant properties. In addition, for 95% of summated scales categories (those using agreement, similarity, frequency, amount, and judgment), actual examples were presented to assist researchers in constructing scales with 5 or 7 points that satisfy the assumption of equidistant points. These suggestions satisfy Spector's (1976) contention that scaling construction procedures include anchor distance calculations because the anchors presented in this research could be used across populations to satisfy the interval assumption needed for statistical level analysis.

The present study revealed 5 different categories in which 95% of research using summated scales can be classified (agreement, similarity, frequency, amount, and judgment). From hundreds of potential anchors, for each of the categories, reasonably exhaustive lists of fewer than 15 per category, were ultimately presented for ranking and analysis. The initial list was gathered from extant research and seminal literature regarding anchors. The lists were narrowed by SMEs who indicated which anchor choices could be eliminated and which might be added. The lists were further narrowed based on pilot studies and focus groups as well as empirical data (high SDs). A final list of no more than 15 anchors was presented to survey participants for ranking. Tables 6-10 show the raw data results for every anchor, and Tables 11-15 respectively show the results of the survey by indicating potential equidistant anchors and their overlaps, which can be used for the most popular (5 or 7 point) scales.

Some of the categories (Tables 11-15) have scales with better equidistant properties than others which is the nature of the calculations. Equal distant choices that are absolutely equal will be rare. The mean difference and mean SD show the mean distance between the choices, whereas the mean SD indicates its consistency. The Tilton overlap distribution indicates that some words are basically synonymous. Interestingly, some scales in existing literature show these empirical synonyms as potential choices. Use of indistinguishable anchors violate the assumption of equal distances and would cause the greatest systematic error.

Further, using the data presented in the present study, researchers can identify anchors that are not equally spaced and should not be used as anchor choices if parametric statistical analysis is desired. For example, I was recently asked to evaluate a

survey that incorporated a common latent survey construct that listed “sometimes” and “occasionally” as two contiguous anchor points. The mean values of these choices, taken from the present study are “sometimes (5.2)” and “occasionally (5.1)”. The distance between “sometimes” and “occasionally” is $-.1$. “Sometimes” and “occasionally” are not only quasi synonyms, but the researcher may have them out of order. As it is, they have a 100% distribution overlap and would be completely variant in analysis. If the researchers had chosen to begin the construction using the scales presented in the current study (i.e., Never, Very Infrequently, Occasionally, Most of the Time, and Always), they would have had anchors that satisfied the assumption for equal distant anchors and would be justified in using parametric statistics.

If a researcher desires an anchored survey that is not based on the recommended 5 or 7 points structures, they can use the research provided in the present study to construct one with equal distant anchor points. For example, if a researcher desired a 6 point anchor scale based on similarity, the data is presented to construct one with equal distant properties (e.g., Not At All (2.4); Neither Like Me Nor Unlike Me (3.2); A Little Like Me (4.4); Moderately Like Me (5.0); To A Large Extent Like Me (6.1); Identical to Me (7.4). Mean Difference = 1, SD = .29).

Limitations and directions for future research

As is true of all empirical studies, the present research has important limitations. Caution should be taken when using these scale values with participants from another population. Normalization of complete ranks (Guilford, 1954) could be completed for each population in which the summated scales are being used as part of the scale validation process (Dobson & Mothersill, 1979; Spector, 1976). In most cases, we are

providing a starting point from which equal distant anchors can be constructed. In each population, scaling procedures should be undertaken include with which scales are used. The optimal response mechanics presented in the paper will develop a sturdy foundation on which to construct scales that adhere to the equal distance assumptions.

Further, in order to shrink the list to a reasonable length for complete seriatim ranking, some words that may be important to researcher, may have been omitted (such as “often”). An additional study using a magnitude estimation technique or side by side word comparison could be completed. The initial lists from the current study could be used as a basis (i.e. greater than 50) (Bass, Cascio, & O’Conner, 1974). These larger list rankings will produce many words that seem to be at the same place on the number line, with large SD’s which will create less exact rule of measurement. However, the analysis of such a study could be combined with the findings of the current study and prove useful in scale construction.

Errors of labeling alternatives from previously used scales can be located such as the example presented in the discussion of number line synonyms and misordering of words such as “sometimes” and “occasionally”. So, commonly used 5 points scales include “sometimes” and “occasionally” as choices whereas based on their synonymic basis from the data in the present study, they should not be listed together as scale points. Are there other anchors choices that should be absolutely avoided as part of a survey response set?

More precision could be made for translations by studying the nature of words such as “always” translated into romance languages: “toujours” (French), “siempre” (Spanish), “sempre” (Portuguese and Italian). A researcher would be wise to implement

a study to native languages to ascertain exact words placement on the number line for each language. Does “never” translate perfectly on the number line with “nunca?” Perhaps within another language (or culture) words have slightly different meanings and need to be studied together to examine their equal distant properties in the native language before assuming translated words maintain equal distant properties from the original tongue. Further, does the numeric exactness of the translation also hold for other Indo-European languages (i.e. German, Polish, etc.) or the other language families such as Sino-Tibetan or Niger-Congo?

Finally, recreating the study using different popular word choices and somehow linking them to the current choices would enable researchers to map additional lists of potential anchor choices to aid in evaluating summated rating scales based upon equal distant properties.

REFERENCES

- Abelson, R.P., & Tukey, J. (1970). Efficient conversion of non-metric information into metric information. In E. R. Tufté (Ed.), *The quantitative analysis of social problems* (pp. 407-415). Reading, MA: Addison-Wesley.
- Alf, E. & Abrahams, N.M. (1968). Relationship between per cent overlap and measures of correlation. *Educational and Psychological Measurement*, 28, 779-792.
- Bass, B. M., Cascio, W. F., & O'Connor, E. J. (1974). Magnitude estimations of expressions of frequency and amount. *Journal of Applied Psychology*, 59, 313-320. doi: 10.1037/h0036653
- Bearden, W. O., Netemeyer, R. G., & Mobley, M.F. (1993). *Handbook of Marketing Scales: Multi-Item Measures for Marketing and Consumer Behavior Research*. Newbury Park, CA: Sage Publications.
- Bendig, A. W. (1954). Reliability and the number of rating-scale categories. *Journal of Applied Psychology*, 38, 38-40. doi: 10.1037/h0055647
- Champney, H., & Marshall, H. (1939). Optimal refinement of the rating scale. *Journal of Applied Psychology*, 23, 323-331. doi: 10.1037/h0054522
- Cicchetti, D.V., Showalter, D., & Tyrer, P.J. (1985). The effect of number of rating scale categories on levels of inter-rater reliability: A Monte-Carlo investigation. *Applied Psychological Measurement*, 9, 31-36.
- Cliff, N. (1959). Adverbs as multipliers. *Psychological Review*, 66, 27.
- Cox, E. P. (1980). The optimal number of response alternatives for a scale: A review.

- Journal of Marketing Research*, 17, 407-422.
- Dobson, K. S., & Mothersill, K. J. (1979). Equidistant categorical labels for construction of Likert-Type scales. *Perceptual and Motor Skills*, 49, 575-580. doi: 10.2466/pms.1979.49.2.575
- Garner, W.R., (1970). Rating scales discriminability and information transmission. *Psychological Review*, 67, 343-352.
- Green, P. E., & Rao, V. R. (1970). Rating scales and information recovery. How many scales and response categories to use? *The Journal of Marketing*, 34, 33-39.
- Guilford, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.
- Hanson, J. W., & Rethans, A. J. (1980). Developing interval scale values using the normalized rank method: A multiple context, multiple group methodology. *Advances in Consumer Research*, 7, 672-675.
- Hensler, C., & Stipak, B. (1979). Estimating interval scale values for survey item response categories. *American Journal of Political Science*, 23, 627-649.
- Hinkin, T.R. (1998). A brief tutorial on the development of measures for use in survey questionnaires. *Organizational Research Methods*, 1, 104-121.
doi: 10.1177/109442819800100106
- Jones, L. V., & Thurstone, L. L. (1955). The psychophysics of semantics. *Journal of Applied Psychology*. 39, 31-36. doi 10.1037/h0042184
- Labovitz, S.. (1970). The assignment of numbers to rank order categories. *American Sociological Review*, 35, 515-524.
- Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 22, 55.
- Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied*

- Measurement*, 3, 85-106.
- McKelvie, S. J. (1978). Graphic rating scales—How many categories? *British Journal of Psychology*, 69, 185-202.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63, 81.
- Newstead, S. E. , & Collis, J. M. (1987). Context and the interpretation of quantifiers of frequency. *Ergonomics*, 30, 1447-1462. doi: 10.1080/00140138708966038
- Oppenheim, A. N. (1992). *Questionnaire Design, Interviewing, and Attitude Measurement*. New York: Pinter Publications.
- Peter, J. P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research*, 16, 6-17.
- 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.
- Schmidt, J. J. (1979). *Understanding and using statistics: Basic concepts*. Lexington, MA: Heath.
- Schriesheim, C., & Schriesheim, J. (1974). Development and empirical verification of new response categories to increase validity of multiple response alternative questionnaires. *Educational and Psychological Measurement*, 34, 877-884. doi: 10.1177/001316447403400416
- Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. *American Psychologist*, 44, 922-932. doi: 10.1037/0003-066x.44.6.922
- Shaw, M. E., & Wright, J. M. (1967). *Scales for the measurement of attitudes*. New York:

McGraw-Hill.

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.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

Spector, P. E. (1976). Choosing response categories for summated rating scales. *Journal of Applied Psychology*, 61, 374-375. doi: 10.1037/0021-9010.61.3.374

Spector, P. E. (1992) Summated rating scale construction: An introduction. *Quantitative Applications in the Social Sciences*, 7, 64. Newbury Park, CA: SAGE University Papers.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103, 677-680. doi: 10.1126/science.103.2684.677

Symonds, P. M. (1924). On the loss of reliability in ratings due to coarseness of the scale. *Journal of Experimental Psychology*, 7, 456.

Symonds, P. M. (1930). *Tests and interest questionnaires in the guidance of high school boys*. New York: Teachers College, Columbia University.

Tilton, J. W. (1937). The measurement of overlapping. *Journal of Educational Psychology*, 28, 656.

Tufte, E. R. (1970). *The quantitative analysis of social problems* . Boston, MA: Addison Wesley Publishing Company.

APPENDICES

Appendix A

Survey instructions given to SMEs and results.

Dear Participant,

I am asking for your assistance in reducing the number of anchor words in the following lists.

You will be presented with several lists of anchor words that have been used as anchors for "Likert type scales." Help me to narrow these comprehensive lists by identifying any of these anchors that could be discarded from our research because they are either anachronistic, vague, or otherwise confusing. In the next phase of my research, I will distribute another survey utilizing a shortened list compiled from the results of this survey. From the Phase 2 survey, I will be able to calculate scale values for the anchors. The end result will present a standardized list of anchors and their associated scale values so that researchers can choose anchors that are essentially equidistant from each other.*

For this survey, there is no presumption as per the number of items that should be in any group. Further, the rank order of the items that you choose to discard is not important. As an end result, I suspect that many of you will discard a similar list of anchors allowing me to aptly dismiss them from the next phase in my analysis.

Some of the decisions might be clear and others not. The task should be done quickly but carefully and a great deal of thought is not necessary.

*Thank you for your assistance,
Cam Casper*

** I use the phrase "Likert type scales" because it is commonly used in this manner to describe a type of response format containing anchors.*

Study # 1 Agreement Anchors

Verbal Anchors	% Recommending Discard
Slightly agree*	10
Agree a little	60
Mildly agree	80
Somewhat agree*	30
Agree in Part	80
Halfway agree	80
Tend to agree	90
Inclined to agree	80
Moderately agree*	40
Generally agree	50
Pretty much agree	80
Agree on the whole	70
Very much agree*	40
Agree*	0
Do not agree	10
Neither agree nor disagree*	0
Neutral	20
Slightly disagree*	20
Disagree a little	70
Mildly disagree	70
Somewhat disagree*	30
Disagree in Part	80
Halfway disagree	70
Tend to disagree	70
Inclined to disagree	60
Moderately disagree*	30
Generally disagree	70
Pretty much disagree	80
Disagree on the whole	80
Very much disagree*	50
Disagree*	0
Do not disagree	20

Note: Added “strongly agree” and “strongly disagree”

*Items that were selected for final list

Study # 1 Similarity Anchors

Verbal Anchors	% Recommending Discard
not at all like me*	0
A little like me*	10
Slightly like me*	20
Mildly like me	80
Somewhat like me*	0
Rather like me	70
Moderately like me*	30
Pretty much like me	70
Very like me	40
Extremely like me*	30
Identical to me*	30
Very much like me*	60
Absolutely very representative of me	60
Absolutely not representative of me	70
To no extent like me*	20
To a great extent like me*	0
To a large extent like me*	30
Neither like me nor unlike me*	20

*Items that were selected for final list

Study # 1 Frequency Anchors

Verbal Anchors	% Recommending Discard
Rarely	10
Seldom	10
Once or twice	30
Once in a while	40
Now and then	60
Sometimes	20
Occasionally*	0
Pretty often	50
Often	0
Frequently*	10
Repeatedly	60
Most of the time*	20
Always*	10
Continually	60
Constantly	40
Frequently if not always	30
Very often	0
A great deal of the time	50
Very frequently*	10
A great many times	50
Usually	20
Quite often	20
Rather infrequently	40
Commonly	40
Fairly often	20
Fairly many times	70
Some of the time	20
To some degree	30
Now and then	50
Once in a while	20
Not often	10
Not very often	10
Fairly infrequently	40
Infrequently*	10
Rather seldom	50
Very seldom	20
Rarely	0
Very infrequently*	10
Seldom if ever	40
Hardly at all	30

Hardly ever	40
Very rarely	30
Almost never*	0
None of the time	20
Not at all	0
Never*	0
Commonly	50
Occasionally	20
At no time	20
Not once	10
Under no circumstance	20
On no occasion	20

*Items that were selected for final list

Study # 1 Amount Anchors

Verbal Anchors	% Recommending Discard
All*	40
An exhaustive amount	50
Almost entirely	50
Completely	30
An extraordinary amount of	40
Almost completely	20
An extremely abundant amount of	60
An extreme amount of	30
a great amount of	20
A great deal of	10
Very much*	10
A full amount of	40
A lot of*	30
Much	40
Quite a bit of	60
A good bit of	50
A considerable amount of	30
Pretty much	50
Fairly much	50
An ample amount of	50
An adequate amount of	40
A moderate amount of	10
Some*	0
To some extent	30
To some degree	30
Somewhat	20
A limited amount of	40
A little of*	10
A small amount of*	30
Comparatively little	40
A little bit of	30
Not much*	20
A small degree of	10
Very little*	10
A slight amount of	30
A meager amount of	70
A scanty amount of	70
A minimum amount of	20
A trifling amount of	70
Scarcely any	40

A trivial amount of	30
Scarcely any	50
A trivial amount of	50
An insignificant amount of	20
Hardly any*	10
None*	0

*Items that were selected for final list

Study # 1 Judgment Anchors

Verbal Anchors	% Recommending Discard
I like it best of all	30
My Favorite	40
Like extremely	40
Like intensely	50
Excellent*	20
Wonderful	30
Strongly like	20
Like very much	30
Might fine	70
Especially good	50
Highly favorable	10
Like very well	50
Very good	10
Like quite a bit	20
Enjoy	20
Preferred	10
Good*	10
Welcome	40
Tasty	40
Pleasing	30
Like fairly well	50
Like	20
Like moderately	30
OK	30
Average*	10
Mildly like	40
Fair*	20
Acceptable	0
Only fair	40
Like slightly	20
Neutral	0
Like not so well	50
Like not so much	50
Dislike slightly	20
Mildly dislike	40
Not pleasing	30
Don't care for it	20
Dislike moderately	40
Poor*	20
Dislike	10

Don't like	20
Bad*	10
Highly unfavorable	10
Strongly dislike	20
Dislike very much	10
Very bad	10
Terrible*	20
Dislike intensely	30
Loath	20
Dislike extremely	30
Despise	30
Would definitely recommend	20
Would not recommend	20
Neither like nor dislike	0
Neutral	10

*Items that were selected for final list

“Horrible”, “inferior”, “unsatisfactory” “mediocre” “passable” “satisfactory”, & “outstanding” were added based on feedback to include words that would be more useable in performance appraisal scales.

Appendix B

Comments from original pilot study:

We would appreciate any comments you have about the survey. After leaving a comment (not mandatory), you will exit the survey....
<i>A little more clarification on the rating system, when doing these, I didn't realize you had to rank all the words, just those applicable</i>
<i>I like that you have more than 4 options to chose from but 20 is too many!</i>
<i>I think so many words make it almost impossible to put them in order, because different words mean different things to different people.</i>
<i>The list of words were sometimes extensive, becoming a little overwhelming to organize.</i>
<i>no comment</i>
<i>this was a cool survey!</i>
<i>the list in light grey is a bit hard to read through. A darker color would be better for reading text.</i>
<i>The first 3 questions should be split into 2 questins apiece they are too large to not be frustrating.</i>
<i>cool survey, would like to know what you end up doing with the results.</i>
<i>the lists were too long and too many choices</i>
<i>I believe that surveys that use these words for ranking without corresponding numbers creates very subjective data. Words mean different things to different people - I'm sure that is part of the point of this survey.</i>
<i>It was quick but definitely made you answer questions.</i>
<i>I feel the survey was very clear and recognized alot of the selection choices from previous surveys I have taken.</i>
<i>It was interesting to see my ranks and viewpoints</i>
<i>I found some of the answers that i chose have been on previous surveys that I have taken.</i>
<i>I have never been asked about how I like to be surveyed. I appreciated this survey.</i>
<i>It was interesting to learn how you rank and view certain words</i>
<i>The questions never showed up in the Question area</i>
<i>very interesting to see how people rank certain words as having more meaning or less meaning than other words of similarity</i>
<i>it was different</i>
<i>Cool survey. It really made me realize that each person's answers are going to depend on the way they talk. Some people don't use the word "fair" when judging or giving their opinion, so they may not rank it as important. Interesting stuff.</i>
<i>Very Different from other surveys I've taken</i>
<i>Good luck on your research!</i>
<i>I liked the survey. It made me think in depth about the different degrees in word meaning when judging something.</i>
<i>i didn't know i could rearrange the list before putting them in the box until the last page just tried it a hint would've made it go faster</i>
<i>i think you clustered too many items to rank in one question.</i>
<i>Very thought provoking, interesting survey. Unlike any other I have ever taken</i>
<i>some questions were too much</i>
<i>Curious...</i>
<i>Survey was a little long, but interesting.</i>
<i>i would never like to do this survey again...just saying because it was super confusing and alot of dragging...although thank you for your time</i>
<i>This survey was much different than I had expected, but in a good way.</i>

<i>made you think</i>
<i>It was just ok, could be the life of some people</i>
<i>Perhaps the options were to numerous.</i>
<i>Limit the number of options to no more than 15 per question.</i>
<i>The instructions were a little confusing at first.</i>
<i>The third list was confusing</i>
<i>enjoyed</i>
<i>why was I asked to compare a bunch of synonyms?</i>
<i>I do not have a comment, survey was good.</i>
<i>had fun thanks</i>
<i>well thought out. thank you!</i>
<i>too many modifiers to rank on the survey</i>
<i>SOME LISTS WERE A LITTLE LONG FOR GOOD ATTENTION SPAN. SOME WORDING O THE DIRECTIONS COULD BE A BIT MORE CLEAR ON WHICH WAY YOU ARE DOING THE LISTS.</i>
<i>Maybe have better instructions for the rankings and less words that has the same meanings.</i>
<i>Too many options.</i>
<i>There are too many items to rank in some survey questions.</i>
<i>middle two to long</i>
<i>Although, it did take some thinking I liked the survey that was given.</i>

Appendix C

Survey distributed to participants (N=307)

PARTICIPANT INFORMATION
Oklahoma State University

Title:
Developing Better Summated Rating Scales by Simplifying and Standardizing Choices for Response Category Anchors

Investigators
Wm. Camron Casper, Ph.D. Candidate, Spears School of Business, Oklahoma State University
Bryan D. Edwards, Ph.D., Associate Professor, Spears School of Business, Oklahoma State University

Purpose:
The primary objective of my research is to determine the optimal number of anchor categories and the best verbal anchors for creating survey scales.
You must be 18 years or older to participate.

What to Expect:
This research study is administered online. Participation in this research will involve completion of two questionnaires. The first questionnaire will ask for you to rank and make related comments to several lists of potential rating scale choices. The second questionnaire will ask for your personal identification information so that we can provide you extra credit for your participation. They should take you about 15 minutes to complete.

Risks:
There are no risks associated with this project which are expected to be greater than those ordinarily encountered in daily life.

Benefits:
You may gain an appreciation and understanding of how research is conducted.

Compensation:
You will receive one unit of course credit for your participation. Other alternatives for course credit are available – please check with your instructor for details.

Your Rights and Confidentiality:
Your participation in this research is voluntary. There is no penalty for refusal to participate, and you are free to withdraw your consent and participation in this project at any time, without penalty. To receive the extra credit, you need to complete the survey in its entirety. If you do not wish to do so, an equivalent assignment is provided—please check with your instructor for details.

Confidentiality:
All information about you will be kept confidential and will not be released. You will not be identified individually; we will be looking at the group as a whole. The second survey will be completely separated from the first and after your professor has been given your name for extra credit recording purposes, the researchers will immediately delete those results.
All data will be stored on a password-protected computer in an encrypted file under the possession of the researchers for the life of the project or at least seven years.

Contacts:
You may contact any of the researchers at the following addresses and phone numbers, should you desire to discuss your participation in the study and/or request information about the results of the study:
William Camron Casper, Ph.D. Candidate, 372a North Hall Tulsa Campus, Spears School of Business, Oklahoma State University, camron.casper@okstate.edu 210.317.3147.
Bryan Edwards, Ph.D., Associate Professor and Joe Synar Chair, 310 Business Building, Spears School of Business, Stillwater, OK 74078, bryan.edwards@okstate.edu 405.744.8331.

If you have questions about your rights as a research volunteer, you may contact Dr. Sheila Kennison, IRB Chair, 219 Cordell North, Stillwater, OK 74078, 405-744-3377 or irb@okstate.edu

If you choose to participate:
Please, click YES if you choose to participate. By clicking YES, you are indicating that you freely and voluntarily and agree to participate in this study and you also acknowledge that you are at least 18 years of age. By clicking YES, you are consenting to participate.

It is recommended that you print a copy of this consent page for your records before you begin the study by clicking below.
(If you click yes, you will be taken to the survey. By selecting no, you will immediately exit this survey.)

The primary objective of this research is to determine the best words to use in survey questions. When you rank these potential survey question answers, you will help determine the optimal word choices.

Please rank all of the words in order of agreement (top) to disagreement (bottom). The word that indicates the highest level of agreement should be ranked 1 and the word that indicates the most disagreement should be listed last (# 13). The remaining words should be ranked in order between these two extreme points. To reorder the list, click and drag the word to its correct order.

Agree	1
Slightly agree	2
Somewhat disagree	3
Somewhat agree	4
Moderately agree	5
Moderately disagree	6
Very much agree	7
Disagree	8
Slightly disagree	9
Strongly disagree	10
Strongly agree	11
Very much disagree	12
Neither agree nor disagree	13

Please rank all of the words in order of similarity (top) to dissimilar (bottom). The word that indicates the highest level of similarity to you should be ranked 1 and the word that indicates the most dissimilar should be listed last (# 12). The remaining words should be ranked in order between these two extreme points. To reorder the list, click and drag the words into the correct order.

Not at all like me (not like me at all)	1
A little like me	2
Slightly like me	3
Moderately like me	4
Extremely like me	5
To a great extent like me	6
Neither like me nor unlike me	7
To a large extent like me	8
Somewhat like me	9
Identical to me	10
To no extent like me	11
Very much like me	12

Please rank all of the words in order of frequency (top) to lack of frequency (bottom). The word that indicates the highest level of frequency to you should be ranked 1 and the word that indicates the most dissimilar should be listed last (# 10). The remaining words should be ranked in order between these two extreme points. To reorder the list, click and drag the words into the correct order.

Sometimes	1
Occasionally	2
Frequently	3
Most of the time	4
Always	5
Very frequently	6
Infrequently	7
Very infrequently	8
Almost never	9
Never	10

Please rank all of the words in order of amount by placing the words that indicate the most at the top and the words that indicate the least at the bottom. The word that indicates the highest amount to you should be ranked 1 and the word that indicates the least amount should be listed last (# 13). The remaining words should be ranked in order between these two extreme points. To reorder the list, click and drag the words into the correct order.

All	1
A lot of	2
Some	3
A little	4
A small amount of	5
Not much	6
Hardly any	7
None	8
A large amount of	9
Not at all	10
Very little	11
To a large extent	12
Very much	13

Please rank all of the words in order of amount by placing the words that indicate the best performance at the top and the words that indicate the worst performance at the bottom. The word that indicates the highest performance to you should be ranked 1 and the word that indicates the worst performance should be listed last (# 15). The remaining words should be ranked in order between these two extreme points. To reorder the list, click and drag the words into the correct order.

Terrible	1
Fair	2
Unsatisfactory	3
Bad	4
Good	5
Mediocre	6
Poor	7
Inferior	8
Average	9
Excellent	10
Outstanding	11
Horrible	12
Satisfactory	13
Passable	14
Decent	15

VITA

William Camron Casper

Candidate for the Degree of

Doctor of Philosophy/Education

Dissertation: CONSTRUCTING EQUAL DISTANCE RESPONSE ANCHORS

Major Field: Management

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Management at Oklahoma State University, Stillwater, Oklahoma in May, 2013.

Completed the requirements for the Master of Business Administration at Brigham Young University, Provo, UT in 1990.

Completed the requirements for the Bachelor of Science in Psychology at Brigham Young University, Provo, UT in 1988. Graduated Summa Cum Laude/ Valedictorian

Experience:

Chevron (1989, 1990-1995)

Human Resource Development Program (1990-1992)

Human Resource Generalist (1992-1994)

Sr. Human Resource Generalist (1994-1995)

AlliedSignal (1995-1997)

Human Resources Plant Manager (1995-1996)

Global Human Resources Director (1996-1997)

Citi (1997-2006)

Vice-President Human Resources Bank Operations (various)

Professional Memberships:

Society for Human Resources Management

Academy of Management

Society for Industrial and Organizational Psychologists

Certification:

Senior Professional in Human Resources (Lifetime Certified) by Human Resources Certification Institute (HRCI)