EXPLORING OPTIMAL RESPONSE LABELS FOR CONSTRUCTING AN INTERVAL TYPE 5-POINT LIKERT SCALE

by

Douglas Hutchinson

A thesis

submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Organizational Performance and Workplace Learning

Boise State University

May 2021

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Douglas Hutchinson

Thesis Title: Exploring Optimal Response Labels for Constructing an Interval Level 5-Point Likert Scale

Date of Final Oral Examination: 21 January 2021

The following individuals read and discussed the thesis submitted by Douglas Hutchinson, and they evaluated their presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Seung Youn (Yonnie) Chyung, Ed.D.	Chair, Supervisory Committee
Soo Jeoung "Crystal" Han, Ph.D.	Member, Supervisory Committee
In Gu Kang, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Seung Youn (Yonnie) Chyung, Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

ACKNOWLEDGMENTS

I've received so much support and assistance from the OPWL program, my thesis committee, and the Boise State community. Really, it seems impossible to thank and acknowledge all of the people who have helped me throughout this process.

First, I would like to thank Dr. Chyung who has served as my Graduate Assistant Supervisor, Thesis Committee Chair, and mentor. Your constant feedback, guidance, and encouragement was so crucial to the success of this thesis. I'm so grateful for the opportunity to learn from you.

Also, to my committee members Dr. Han and Dr. Kang. Thank you for all of your time and feedback. Your helpful insights were vital in guiding the direction of my research and making necessary improvements to my work.

I'd like to express my gratitude for the students, faculty, and administrative staff of the Boise State OPWL program for making my graduate experience so enjoyable. In particular, Jo Ann Fenner and Kelly Weak have provided so much support and assistance as I worked towards joining this program and earning my degree.

Thank you to Tayler Smith and Ben Quintana who introduced me to this field and to the OPWL program. Your passion and interest in serving others was such an inspiration at a time when I needed direction.

Finally, to Taylor Fennell and all of my friends and peers in the Student Union Building. You constantly pushed me to continue learning and growing even when I didn't believe in my own capabilities. None of this would have been possible without you.

ABSTRACT

Performance improvement practitioners value evidence-based practices, which include data-driven decisions. Data can be obtained through survey questionnaires designed with closed-ended questions and response scales. The Likert scale (*Strongly Disagree*, *Disagree*, *Neutral*, *Agree*, and *Strongly Agree*) is one of the most commonly used response scales. Whether the 5-point Likert scale, as a verbal descriptor scale, should be treated as an ordinal or interval scale is an on-going debate, as different types of statistical analyses are applied to ordinal and interval data. I conducted this study to examine if survey participants would perceive a 5-point Likert scale close to an interval level measurement when an adverb such as *Moderately*, *Somewhat*, or *Slightly* is added in front of *Agree* and *Disagree*. This information could be used by researchers who wish to construct an interval type Likert scale.

I conducted this study using a convenient sample of performance improvement practitioners, including master's degree and graduate certificate seeking students, recent alumni, and faculty in the Organizational Performance and Workplace Learning department at Boise State University. For this study, I developed a web-based survey instrument using a horizontal slider format. The first screen of the survey instrument contained eight partially-labeled Likert scale sliders, each of which presented three anchors in ascending order (*Strongly Disagree* on the far-left side, *Neutral* in the middle, and *Strongly Agree* on the far-right side) along with their numerical values (-2, 0, and +2, respectively). The slider bar was initially placed on *Neutral* (0). Participants were

instructed to move the slider bar to locate each of the following eight anchors on the Likert scale slider; *Disagree*, *Moderately Disagree*, *Somewhat Disagree*, *Slightly Disagree*, *Agree*, *Moderately Agree*, *Somewhat Agree*, and *Slightly Agree*. To test the response order effect, the second screen of the instrument asked the participants to repeat the above procedure using another set of eight Likert scale sliders presented in descending order. The third screen of the instrument asked for participants' gender, age group, and native English speaker status.

The data was collected in October of 2020. The web-based survey system (Qualtrics) recorded data rounding to two decimal points and provided summary report data including mean, standard deviation, variance, and minimums and maximum response scores for each item. A survey invitation was sent to 327 practitioners, and a total of 109 of them submitted the survey. However, the initial data screening detected 37 datasets with responses where any responses were incomplete or used the incorrect side of the slider continuum, which were excluded. Two additional responses from non-native English speakers were also excluded due to the linguistic aspect of the study. This left 70 responses available for analysis (51 females, 18 males, 1 "do not want to report").

The anchor being tested would be considered useful for constructing an interval measurement if its corresponding confidence interval included the value -1 or +1. To test this, 95% confidence intervals were constructed for each of the 16 items. Response order effects were investigated by performing paired sample t-tests comparing the average scores of the 8 response options when presented in ascending versus descending order.

The results showed that, *Moderately Disagree* and *Moderately Agree* were closely aligned with -1 or +1 on the continuum, respectively, regardless of the response orders

used. *Agree* was aligned with +1 only when presented in ascending order, but not when presented in descending order. Adding other adverbs *Somewhat* and *Slightly* to *Agree* and *Disagree* made the 5-point Likert scales to be clearly ordinal scales in both response orders used. Therefore, the study concluded that when one needs to collect interval data from a 5-point Likert scale, *Moderately Agree* and *Moderately Disagree* can be used in either ascending or descending order of the scale.

Although *Somewhat* would not be a good adverb to be added to *Disagree* and *Agree* when the 5-point Likert scale is expected to generate interval data, an unexpected interesting finding was that *Strongly Agree*, *Somewhat Agree*, *Somewhat Disagree*, and *Strongly Disagree* in descending order can be used as an interval-level 4-point Likert scale.

This study was conducted with several limitations including the use of a convenience sample, and the generalization of the findings may be limited.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

Performance Improvement Using Evidence Based Practice

The field of performance improvement focuses on reducing performance gaps between existing and desired performance in an organization (Aziz, 2013). Human performance improvement (HPI) practitioners use evidence-based practices to ensure that their decisions are defensible and add value to clients (Aument & Conley, 2018). Practitioners use different forms of evidence. One form is professional expertise—frameworks, models, or procedures developed by experts. For example, professionals often use the Human Performance Technology (HPT) model to help conduct performance and cause analysis, select, design, and develop appropriate interventions, implement interventions, evaluate the process and outcomes, and apply effective change management strategies (Van Tiem et al., 2012). Each phase of the HPT model should be informed with appropriate evidence. This includes reviewing existing data, such as an organization's training materials, customer feedback, or past performance evaluations. The practitioner may collect new data through methods such as interviews, focus groups, on the job observation, or surveys (Chyung, 2019). Capturing new data often requires the additional step of developing data collection instruments such as survey questionnaires during which the practitioners need to make a number of design decisions that can influence the likelihood that data collected is both appropriate and accurate. In other words, HPI practitioners also need to use evidence-based practices while developing survey questionnaires with appropriate response scales.

Appropriate Response Scale Design for Intended Data

Likert scale formats are often used in closed-ended survey questions. Likert (1932) developed the 5-point descriptive response scale to measure an individual's perceptions or opinions on a topic. The Likert scale presents response options in a bipolar format (both negative and positive options), allowing researchers to collect information regarding both the direction and the intensity of an individual's opinion. In his original research, Likert used the descriptors: *Strongly Approve, Approve, Undecided, Disapprove,* and *Strongly Disapprove.* Modern Likert scales often use *Disagree* and *Agree* anchors instead of *Disapprove* and *Approve*, but still use an odd number of response options allowing for equal number of positive and negative rating options plus a *Neutral* midpoint option (Chyung et al., 2017).

While the Likert scale possesses many practical advantages, researchers may be limited in the statistical analysis that can be performed on data collected. The type of statistical analysis that can be performed on data is dependent on whether it can be classified as nominal, ordinal, interval, or ratio (Stevens, 1946). The Likert scale produces data that is ordinal, meaning that the response options are placed in a linear order allowing for median and mode calculations (Jamieson, 2004). However, some researchers choose to treat Likert scale data as interval, in which response options are not only provided in a linear order, but also presented with an equal distance between any two consecutive points. Interval data is desirable to researchers if they wish to calculate statistics such as mean or standard deviation. When interval data is normally distributed, it also allows researchers to analyze the data with parametric tests, while ordinal data,

usually not normally distributed, are analyzed with nonparametric tests. Parametric tests are considered to be more powerful than nonparametric tests.

Some researchers believe that Likert-type scales can be treated as interval as long as data is both ordered and has somewhat equal spacing between response points. However, Borgatta and Bohrnstedt (1980) pointed out that measurements used in social sciences rarely produce data that is perfectly interval with consecutive points that are equidistant from one another. In their view, data can be considered interval as long as it is normally distributed. Carifio and Perla (2008) reason that a construct (concept) should be measured with not just one but several survey items with Likert scales, and that this design creates empirically interval data. Likewise, Joshi et al. (2015) noted that data can be categorized as interval when items are combined to create a composite score.

A Method to Make Likert Scale Data Close to Interval

Based on the assumption that the 5-point Likert scale is an ordinal scale, Worcester and Burns (1975) suggested a method to make the Likert scale data close to interval data, which is to modify the second and fourth response options (a.k.a. intermediate anchors). They performed a study in which 1,932 adults completed a survey with Likert response options. Participants completed one of four possible versions of the survey with unique response labels for each scale. Participants then indicated their interpretation of their chosen response by marking its location on a continuous line. Researchers found that participants interpreted the intermediate anchors such as *Agree* differently than *Tend to Agree* or *Agree Slightly*. Additionally, the scale with *Disagree Slightly* and *Agree Slightly* most closely resembled an interval scale.

It is worth noting that this research was performed roughly forty-five years ago using paper surveys. Today, surveys are frequently administered online. The Web-based online survey systems often include pre-set response options to assist researchers in developing their survey instruments. Different default response options, especially the differently worded intermediate anchors of the Likert scale, provided in different online survey systems could potentially impact the quality of data that researchers receive. The primary focus of this thesis is to investigate whether adding an adverb such as *Moderately*, *Slightly*, and *Somewhat* in front of the intermediate anchors, *Agree* and *Disagree*, in the Likert scale would make the data closer to interval when tested in a web-based environment. If this is the case, researchers would have the opportunity to construct Likert scales that produce data that can be analyzed using more powerful statistical methods of analysis.

Likert scale responses can be presented in either ascending or descending order, which can also impact the data produced by scales (a.k.a. response order effects). For example, research has shown that descending-ordered scales likely produce more positive data (Betts & Hartley, 2012; Chan, 1991; Maeda, 2015). To avoid the response order effect, it is recommended to use an ascending-ordered Likert scale. However, in addition to testing the effects of adding an adverb to the response labels in an ascending-ordered Likert scale, it is worthwhile to repeat the same testing using a descending-ordered Likert scales. This paired testing would allow investigating if the effects of adding an adverb to the Likert response labels are influenced by the response order.

Research Questions

The main purpose of this research is to answer the following research question:

Research Question 1: Does adding an adverb (such as *Moderately*, *Slightly*, and *Somewhat*) to *Disagree* and *Agree* in the 5-point Likert scale influence people to perceive the scale to be closer to an interval scale, when administered in a web-based environment?

This research also included a secondary research question:

Research Question 2: Does the order of response options in the Likert scale (ascending vs. descending) make a difference in people's perceptions as tested in Research Question #1?

Significance of the Problem

When conducting surveys, practitioners and researchers often use web-based survey systems, such as Qualtrics or Survey Monkey. These online survey tools often provide survey designers with a Likert scale template that automatically generates the labels for scale options. When selecting a 5-point Likert scale, Qualtrics provides *Somewhat Agree* and *Somewhat Disagree* as the default labels for intermediate response options (*Strongly Agree*, *Somewhat Agree*, *Neither Agree nor Disagree*, *Somewhat Disagree*, *Strongly Disagree*), whereas Survey Monkey provides *Agree* and *Disagree* as the default (*Strongly Agree*, *Agree*, *Neither Agree nor Disagree*, *Disagree*, *Strongly Disagree*). Practitioners would likely use the default labels provided by their preferred survey system, and it is unknown whether the use of adverbs in front of the intermediate response option would produce a meaningful difference in the data collected on these platforms. This research will help survey designers to make an evidence-based decision

regarding whether to add adverbs to agree/disagree options in the Likert scale when they hope to collect interval data.

Definition of Terms

The following terms, as defined below, are used in this thesis:

A closed-ended question consists of an inquiry statement or question and a response scale that respondents use to express their opinions.

Evidence-based practice involves making decisions that are expertise-based and data-driven, and is essential in performance improvement practices, including survey development and data collection.

The Likert scale captures respondent's opinions by asking them to express the extent to which they disagree or agree with a provided statement, typically using a 5-point scale with a midpoint, or a 4-point scale without a midpoint.

Response labels (or anchors) are words or phrases used by survey designers to communicate the specific meaning of an individual response point to survey participants.

Response-order bias occurs when the response provided by a survey participant is dependent on the order (ascending vs. descending) in which response labels are presented. This can happen in scenarios in which participants are not motivated to provide precise responses and simply select the first acceptable option they find, or when they are inclined to select an option that is deemed desired by others.

CHAPTER TWO: LITERATURE REVIEW

Evidence-Based Practice for Performance Improvement

How does this study connect to Human Performance Technology (HPT)? Pershing et al. (2006) define HPT as "a professional field of study and application, the main purpose of which is to engineer systems that allow people and organizations to perform in ways that they and all stakeholders value" (p.xiii). The International Society for Performance Improvement (ISPI) (2012), a leading organization in advancing the field of performance improvement, describe HPT as "a systematic approach to improving productivity and competence, uses a set of methods and procedures—and a strategy for solving problems—for realizing opportunities related to the performance of people." The term "technology" is essential to the HPT discipline. The Editors of Encyclopaedia Britannica (2021) define technology as "the application of scientific knowledge to the practical aims of human life, or as it is sometimes phrased, to the change and manipulation of the human environment." The HPT field is meant to be scientific—driven by research and evidence-based practice. Performance improvement practitioners are encouraged to use evidence-based practices whenever possible.

Evidence-based practice (EBP) entails making decisions that are expertise-based and data-driven (Chyung, 2019). The application of evidence-based practice is important to the HPT field in order to achieve professional status and to provide better bottom-line results to clients (Clark, 2006). HPT professionals rely on field-tested models, frameworks, or procedures in order to develop and employ professional expertise. This

expertise is complemented by reviewing or generating evidence that is specific to the project at hand (Duan, 2011). For example, even when practitioners conduct surveys to collect data, they must consciously apply existing evidence-based practices to design their survey instruments. They should be aware that a number of factors relating to a survey design with closed-ended questions and response scales will change the data they receive.

Evidence-Based Survey Design Principles

There are several important survey design principles to be applied when designing survey instruments with closed-ended questions and response scales. For example, practitioners must decide whether to include a midpoint when using a Likert-type scale (Chyung et al., 2017); whether to use positively worded items only or a mix of positively and negatively worded items (Chyung et al., 2018a); how to design response scales to avoid ceiling effects in data (Chyung et al., 2020). These design decisions likely impact the quality of survey data to be collected.

Surveys responses are also impacted by the design decision of whether to present response options in ascending or descending order (Krosnick & Alwin, 1987). An ascending-ordered scale presents response options in order from lowest to highest values (e.g., *Strongly Dissatisfied, Dissatisfied, Neutral, Satisfied, Strongly Satisfied*). Descending-ordered scales instead order response options from highest to lowest (e.g., *Strongly Satisfied, Satisfied, Neutral, Dissatisfied, Strongly Dissatisfied*). Scales that are presented in descending order tend to produce higher mean scores than those presented in ascending order (Chan, 1991; Friedman et al., 1993). But, why are participant responses

dependent on response order? Researchers have found several explanations that describe the impact response order has on survey data.

Survey data may suffer from a primacy effect when respondents show increased bias towards response options that are presented to them earlier in a scale. This is likely due to the satisficing principle, in which people simply select the first response option they see that resembles their opinion instead of going through the full process of reading through each alternative, deciding which one best represents their opinion and selecting the best option (Simon, 1957). In web-based and paper-based surveys, this effect favors response options that are presented furthest to the left when participants read from left to right (Chyung et al., 2018b). A potential consequence of this left-side selection bias when response options are presented in descending order is inflated data, as the positive options on the far left are more often selected. To avoid this inflated data phenomenon, providing response options in ascending order is recommended (Chyung et al., 2018b). Thus, the primary research question of this thesis study (Research Question 1) was tested with ascending-ordered response scales. Then, for comparison purposes, the testing was repeated using descending-ordered response scales (Research Question 2).

Measurement Scales Used in Research

Another issue that practitioners who are using surveys should be aware of is the type of data their survey questionnaires would generate and if the types of data would support the type of statistical analyses that they intend to use. Stevens (1946) developed four types of measurement scales that determine which statistical analysis methods can be applied to data—nominal, ordinal, interval, and ratio. In research, survey items should be

properly designed to generate the intended types of data and to be analyzed with the intended statistical methods.

Nominal Scales

Nominal scales (or data) contain no inherent quantitative property, and the order in which individual response options are presented are insignificant for testing purposes. Scales with only two response options (such as *Yes* or *No*, or *Pass* or *Fail*) are typically treated as measuring nominal or dichotomous variables (Morgan et al., 2013). Information such as gender, names, or office location are considered nominal. The order in which this data is presented does not represent any inherent order between individual response options. In surveys, nominal data is often presented by showing how frequently individual response options were selected and is often used to describe or profile research participants.

For instance, Wisshak and Hochholdinger (2020) performed a study with 190 survey participants in which they investigated whether hard-skill trainers and soft-skill trainers differed in terms of the knowledge and skill they believed their position required. Their classifications of participants as hard skill trainers or soft skill trainers are nominal data. The researchers also used nominal data to describe characteristics of research participants in terms of their gender; male (51%) and female (49%), and in terms of their job category; training as their main job (63%), training as a side job (30%), and other (7%).

Nominal data are also found in a study conducted by Mousa et al. (2020) who investigated if organizational learning is positively correlated with individual components of organizational resilience including robustness, agility, and integrity by asking 236

academics at universities in Cairo, Egypt, to complete a paper-based questionnaire including demographic information. Researchers used nominal variables to profile their research participants by collecting information on their work base (part-time or full-time employees), gender (male and female), marital status (single, married, other), and religion (Buddhist, Christian, Muslim). Each variable was presented by describing the frequency and percentage of the nominal data options.

Ordinal Scales

In ordinal scale data, the order in which data is presented is relevant because the response options are ranked, but the relative distance between response points across the scale is not equal (Morgan et al., 2013). This could include information such as educational attainment or tax brackets. With ordinal data, certain statistical analyses can be performed, such as frequency, median, and nonparametric testing, but not averages because the distance between each set of consecutive response options is not equal. Greater than and less than conclusions can also be drawn because the data is ordered. For example, a researcher could conclude that earned an "A" on a test performed better than a student that earned a "B."

Ordinal scales are used in research. For instance, Steil et al. (2019) investigated the relationship between technical and managerial employees' perceptions of learning opportunities within their organization and their intention to stay with the organization in technology companies in Santa Catarina, Brazil. With their survey instrument, the researchers measured the variables using an ordinal ascending 5-point scale with response labels either ranging from *Never* to *Always* or from *Totally disagree* to *Totally agree*. Normality testing rejected the normality assumption of the data, and the

researchers used non-paramedic tests such as Kruskal-Wallis test, Spearman's rho, and Mann-Whitney test.

Interval Scales

Interval data is characterized as not only data that is ordered, but also data where the intervals between all categories is equal. Any zero point in interval data is arbitrary because it does not indicate an absence of a specific characteristic (Morgan et al., 2013). Examples of interval data include standardized test scores or Fahrenheit temperature. Notice that a temperature of zero degrees Fahrenheit does not mean that no heat exists. With interval data, you can use parametric statistics and calculate means and standard deviations.

As an example of research studies where interval data were collected and analyzed, Maddy and Rosenbaum (2018) studied whether Leadership Self-Efficacy (LSE) score correlated with self-assessed leadership levels based on 124 completed survey data obtained from employees mostly at a state college. Employees participated by completing a questionnaire to generate an LSE score, which is an interval measurement of scores ranging from 203 to 436. Employees then completed a second self-assessment questionnaire, which measured their leadership skills across 18 dimensions and were averaged to create a composite score. Researchers calculated a Pearson Correlation coefficient and conducted a linear regression analysis to test their hypothesis that the two scores would be correlated.

Ratio Scales

The highest level of measurement scale is known as ratio. Ratio is similar to an interval measurement in that it is ordered and has equal intervals between any two

consecutive points. Ratio data differs from interval data in that ratio data has a true zero point, meaning that a data point of "0" refers to an absolute zero value (Morgan et al., 2013). Money is one example of ratio level data. If a person has $0, they have no money. Additional operations such as multiplying and dividing are appropriate for ratio data as well. This allows for conclusions to be drawn such as an employee with $1,000,000 in income per year makes ten times as much as an employee with a salary of $100,000 per year. It would be inaccurate to do this with an interval measure such as temperature by concluding it is twice as hot 6 months ago (90 degrees) as it is today (45 degrees).

Ratio scale data can be found in studies such as the one conducted by Park and Jacobs (2011). They performed a study investigating whether investments in workplace learning corresponded with increased financial performance of companies in South Korea. Company financial information was extracted from the Korea Information Service, providing ratio data including sales per employee, net profit per employee, gross margin, and return on assets.

A summary of the four levels of measurement and their properties are presented in Table 1. Ordinal data has its own characteristic (a certain order among data) as well as the characteristics of nominal data (distinct values); interval data has its own characteristic (equal intervals between options) as well as the characteristics of ordinal data; ratio data has its own characteristic (absolute zero value) as well as the characteristics of interval data.

Table 1 Summary of Properties of Levels of Measurement (Adapted from Chyung, 2019, Figure 16. The Relationship Among Four Types of Measurement Scales, pp. 167)

Level / Property	Nominal	Ordinal	Interval	Ratio
Distinct value of each option	Yes	Yes	Yes	Yes
Order among options	No	Yes	Yes	Yes
Equal interval between options	No	No	Yes	Yes
Absolute zero value	No	No	No	Yes

Statistical Procedures for Ordinal and Interval Data

A clear understanding of the types of data (nominal, ordinal, interval, and ratio) collected from survey questionnaires helps practitioners and researchers select appropriate statistical analyses. Conversely, when they intend to use certain types of statistical analysis, they would need to plan to collect appropriate types of data. When analyzing interval or ordinal data, generally descriptive and/or inferential statistics are calculated.

Descriptive Statistics

A descriptive statistic describes a dataset with a single value. This includes measures of central tendency, such as mean, median, and mode as well as measures of variability, such as standard deviation or interquartile range. A median, which shows the value between the upper- and lower-half of a sample, is most appropriate for data that is ordinal and is preferable over a mean calculation when data is skewed. On the other hand,

the preferred measure of central tendency for interval data (when not skewed) is the mean, which is an average of all values in a sample. Assuming normal distribution, a mean calculation is more reliable than a median and less likely to vary from sample to sample (Pershing et al., 2006).

Descriptive statistics that measure variability describe how widely data is dispersed away from the center. Data that is widely spread apart will have a higher level of variability than data with scores that are relatively close together. Interquartile range (IQR) is often used as a measure of variability for ordinal data. IQR is calculated by ordering data from least to greatest, finding the values located 25% and 75% of the way through the data set, and subtracting the two values. As is the case with median, interquartile range is useful because it is not distorted by outlier values. On the other hand, standard deviation is best suited for interval data. This statistic shows the average of how far each value is away from the sample's mean. When data is normally distributed, 68% of the sample's total values will fall within a single standard deviation to the right and left sides of the mean (Morgan et al., 2013; Pershing et al., 2006).

Inferential Statistics

Researchers must decide whether to perform parametric or nonparametric tests when calculating inferential statistics, which are designed to help draw conclusions from sample data. Parametric tests often require either interval or ratio data because they are based on the calculation of standard deviations and means. Data should also have a sample of at least 30 and be normally distributed. While nonparametric tests are slightly less powerful than parametric tests, they also have less rigorous requirements. Nonparametric tests can accommodate smaller sample sizes and do not require normal

distributions of data (Pershing et al., 2006). Again, researchers must A summary of commonly used tests is provided in Table 2.

Table 2 Summary of Common Inferential Statistical Tests (Summarized from Morgan et al., 2013)

Goal of Analysis	Test	When to use
To analyze the level of correlation between two variables	Pearson's r	When data is interval or ratio, linearly correlated, and normally distributed
	Kendall's Tau	When data is ordinal, a monotonic relationship exists between variables. Preferred over Spearman's Rho when data is collected from a smaller sample.
	Spearman's rho	When data is ordinal, a monotonic relationship exists between variables.
To test differences between two related observations	Paired Sample t-Test	When data is interval/ratio, data is normally distributed, the subjects in each sample are the same.
	Wilcoxon Signed Rank Test	When data is not normally distributed, but the dependent variable is at least ordinal.
To test differences between two unrelated observations	Independent Sample t-Test	When each variable is normally distributed and the dependent variable is interval/ratio
	Mann-Whitney U Test	When sample sizes are small or data is not normally distributed
To test for differences between three or more independent groups	One-Way ANOVA	The dependent variable is normally distributed, variances of the independent variable are equal across groups
	Kruskal-Wallis Test	Data is ordinal or equality of group variances does not exist

Likert Scale as an Ordinal or Interval Scale

The Likert scale is often used as a response scale in performance improvement related survey research (Cangialosi et al., 2020; Ghosh et al., 2019; Vandergoot et al., 2020; Zadeh & Azedeh, 2020). Rensis Likert (1932) developed the original Likert scale (*Strongly Approve, Approve, Undecided, Disapprove,* and *Strongly Disapprove*), which measures an individual's attitude towards a subject using a 5-point bipolar scale with descriptive anchors for each response option. The modern form of the 5-point Likert scale is *Strongly Agree, Agree, Neutral, Disagree, Strongly Disagree*. This bipolar Likert scale allows the respondents to communicate both the direction and strength of their attitude with the option to select a neutral response.

What type of measurement (nominal, ordinal, interval, or ratio) should the 5-point Likert scale be categorized under? Is Likert scale data an ordinal or interval data? Should Likert scale data be analyzed with parametric or nonparametric tests? There is little doubt that the 5-point Likert scale is at least ordinal because the five response options are arranged in a hierarchical manner. Most researchers hold the opinion that Likert scale data should be classified as not interval but ordinal data (Jakobsson, 2004; Jamieson, 2004; Kerro & Lee, 2015; Knapp, 1990; Vigderhous, 1977). After all, it is difficult to definitively state that survey respondents view the psychological distance between Likert response options, such as *Strongly Agree* and *Agree* to be equivalent to the distance between *Agree* and *Undecided* (Edmondson, 2005). Kuzon et al. (1996) supported this argument stating, "Just as it is invalid to results of a given surgical procedure as poor, fair, good, or excellent and state that the average result is 'fair and a half.' It is invalid to rate those same outcomes as 1, 2, 3, or 4 and state that the average result is 2.5" (p. 266).

If the Likert scale is an ordinal scale, the soundest practice is to limit data analysis to nonparametric testing. However, some practitioners and researchers alike continue to treat Likert scale data as interval. Jakobson (2004) conducted a review of three medical journals to see how ordinal data was handled in nursing literature. This review found that only 49% of research articles that included ordinal variables appropriately presented their data. The author found that some researchers used means and standard deviations when medians and interquartile ranges were more appropriate. Further, 57% of research articles improperly used parametric tests on ordinal variables. It appears unlikely that researchers will stop treating Likert scale data as interval in the foreseeable future.

Also, there are arguments that it is acceptable to treat the Likert scale as an interval scale, particularly under certain conditions. Some researchers argue that parametric tests can be performed as long as a series of Likert items are used to measure the same construct (Carifio & Perla, 2008). Norman (2010) showed that some parametric tests may be performed on Likert scale data without leading researchers to the wrong conclusion. This was supported by Murray (2013), who analyzed Likert scale data using Pearson's *r* (a parametric test) and Spearman's *rho* (a nonparametric test) and found that both tests would lead researchers to similar conclusions. However, these ad-hoc decisions may not always be reliable, and it is preferred to use instruments that are designed to generate the intended type of data.

To Design the Likert Scale as an Interval Scale

Compared to ordinal data, interval data has more potential to be normally distributed, which allows the use of parametric tests. Although the Likert scale is likely an ordinal scale, it is possible to modify the verbal descriptors used in the Likert scale to increase the likelihood that the 5-point Likert scale generates data close to interval data. One such possible method is to add a modifying adverb (e.g., *Moderately, Slightly, Somewhat*, etc.) to the *Agree* and *Disagree* response labels. Worcester and Burns (1975) explored this idea in a paper-based survey of 1,932 adults from Great Britain. The researchers asked participants to answer three questions regarding politics. There were four versions of the survey, each of which used unique modifiers to describe intermediate response points (options 2 and 4 on a 5-point scale) on a descending scale (see Table 3).

Table 3 Sample Question from Worcester and Burns (1975) Study

Q.1. Looking at this card, I would like you to tell me to what extent you agree with the following statement: Neither the Conservative nor the Labour Party represent the views of people like me.			
Scale A	**Scale B**	**Scale C**	**Scale D**
Agree Strongly	Agree Strongly	Agree Strongly	Agree Strongly
Tend to Agree	Agree Slightly	Tend to Agree	Agree
Neither Agree Nor Disagree	Neither Agree Nor Disagree	Tend to Disagree	Neither Agree Nor Disagree
Tend to Agree	Disagree Slightly	Disagree Strongly	Disagree
Disagree Strongly	Disagree Strongly		Disagree Strongly

After participants responded to questions, the researchers instructed them to indicate their interpretation of their chosen response by marking its location on a continuous line. Results indicated that the use of different modifying adverbs did impact the way respondents interpreted the scale. Participants interpreted intermediate anchors

such as *Agree* differently than *Tend to Agree* or *Agree Slightly*. For example, respondents perceived *Tend to Agree* and *Agree Slightly* to be less extreme responses than *Agree*. The modifier *Fairly* was considered less extreme than *Quite* or *Mainly*. Additionally, the scale used in Scale B (in Table 3) with *Agree Slightly* and *Disagree Slightly* most closely resembled an interval scale, generating appropriately equal intervals between options in data.

Since then, it seems that very little research has been performed on the effects of modifiers on Likert scales. I found little information from extensive library search using various databases such as Academic Search Premier, JSTOR, and Science Direct, except a couple of dissertations completed by Casper (2013) and Spratto (2018). The research design used by Spratto (2018) was similar to that of Worcester and Burns (1975), except that Spratto attempted to find an approximately interval 4-point agree-disagree Likert scale (rather than a 5-point scale). For this study, the researcher tested the effects of response scales with the modifiers *Completely, Very Strongly, Strongly, Mostly, Moderately, Somewhat,* and *Slightly,* and a response scale without a modifier. Participants were found from an online social media site, first-year students at a university in the United States where the study was performed, and a website that hires remote workers to perform computer-based tasks. The survey contained items on three topics (mindfulness, agreeableness, and conscientiousness), and was conducted using Qualtrics, a web-based survey tool. Though the researcher held a pessimistic view of their findings, results suggested that a 4-point scale with the anchors *Completely Disagree*, *Moderately Disagree*, *Moderately Agree*, and *Completely Agree* were close to having equal intervals.

Casper (2013) tested the distance between multiple verbal anchors used in the Likert-type scales; however, the researcher approached the research from an angle different than the way Worcester and Burns (1975) and Spratto (2018) experimented. Participants were provided with a list of thirteen response labels and instructed to rank each one from most to least. From there, distribution overlap between anchor points was calculated. Based off of this research, Casper recommended using the anchors *Strongly Disagree, Disagree, Neither Agree nor Disagree, Moderately Agree, and Very Much Agree* in order to create a scale with approximately equal distance intervals in a 5-point Likert scale. A summary of research conducted by each author is provided in Table 4.

Aside from the actual number of scale points, Casper's research contrasted Spratto's and Worcester and Burns' studies by recommending a set of anchors using the different modifiers on each side of the bipolar Likert scale. However, these studies by Worcester and Burns (1975), Casper (2013), and Spratto (2018) all had one in common—an attempt to identify verbal descriptors used in Likert scales that satisfy the interval assumption. Worcester and Burns (1975) suggested that adding the adverb *Slightly* to the *Agree* or *Disagree* anchors in the 5-point Likert scale would help produce data close to interval. Although Spratto's (2018) study supported the conclusion drawn by Worcester and Burns (1975), Spratto (2018) used 4-point Likert scales while Worcester and Burns (1975) used 5-point Likert-type scales. On the other hand, Casper (2013) found that using no adverb on the *Disagree* side of a scale and *Moderately* to the *Agree* side best represented an interval scale. The review of these studies led me to want to investigate further to find a way to design 5-point Likert scales with modifiers so that the survey data are close to interval data when administered in a web-based environment.

Table 4 Summary of Research on Verbal Descriptors in Likert Scales

Author	**Data Collection Medium**	**Response Order Applied**	**Study Recommendations**
Worcester and Burns (1975)	Paper-Based	Descending Order	Use *Slightly* to construct Interval 5-point Likert scales
Casper (2013)	Web-based	Ascending Order	Use *Disagree* and *Moderately Agree* to construct interval 5-point Likert scale
Spratto (2018)	Web-Based	Ascending Order	Use *Completely Agree/Disagree* as endpoints and *Moderately Agree/Disagree* as intermediate points to construct an interval level 4-point Likert scale

CHAPTER THREE: METHODOLOGY

Research Questions

The purpose of this research is to explore a way to create 5-point Likert scale data that is approximately interval. This research further looked at if the response order impacts the outcomes. This research was approved by Boise State University's Institutional Review Board (approval #126-SB20-160).

This research was designed to answer two questions:

1) Does adding an adverb (such as *Moderately, Slightly, or Somewhat*) to *Disagree* and *Agree* in the 5-point Likert scale influence people to perceive the scale to be closer to an interval scale, when administered in a web-based environment?

   H_0: Adding an adverb such as *Moderately, Slightly, or Somewhat* to *Disagree* and *Agree* in the 5-point Likert scale does not statistically significantly influence people to perceive the scale as being closer to interval scale.

   H_a: Adding an adverb such as *Moderately, Slightly, or Somewhat* to *Disagree* and *Agree* in the 5-point Likert scale statistically significantly influences people to perceive the scale as being closer to interval scale.

2) Does the order of response options in the Likert scale (ascending vs. descending) make a difference in people's perceptions as tested in Research Question #1?

   a) H_0: There is no statistically significant difference between mean scores of data obtained from the tested scales when presented in ascending vs. descending order.

H_a: There is a statistically significant difference in the mean scores of data obtained from the tested scales when presented in ascending vs. descending order.

Population and Sampling

The population of this study is native English-speaking adults (18 or older) with a minimum of undergraduate college education. This study used a convenience sample (n = 327) of current students, recent alumni, and faculty of Boise State University's Organizational Performance and Workplace Learning (OPWL) graduate program in the United States.

Web-Based Survey Tools

Surveys were conducted with pen and paper methods before internet access became widely available. However, web-based survey tools have become increasingly popular in recent years and present a number of advantages for both researchers and survey participants. First, online platforms remove the need for researchers to manually enter survey data, which can require a significant investment in time and carries the risk of transposition error (Touvier et al., 2010). These tools also allow researchers to easily incorporate design features such as forced responses and branch logic (Vergnaud et al., 2011). Survey participants may even be more willing to provide candid responses to sensitive topics in an online, self-administered environment (Dayan et al., 2007).

The survey research experience is not only dependent on the distinction between web and paper-based surveys, but also which particular web-based platform that researchers select. According to IBIS World (2020), there are currently 364 online survey businesses. Each survey platform comes with its own unique settings and features, many of which have potential to impact the quality of data collected. For example, it is likely

that survey designers will accept a platform's default response labels as long as the labels are clearly relevant to the item's query. After all, not all researchers and survey designers will have the time or interest to exhaustively investigate the impact of each seemingly minute design decision. Instead, they may instinctively trust the recommended options of their chosen platform. For that reason, I researched the default settings of numerous popular online survey platforms in order to understand the common Likert scale designs employed by researchers.

Platforms were selected by reviewing six articles and industry reports for names of the largest or most popular web-based survey tools in the United States (Business Industry Reports, 2020; Business Wire, 2017; IBIS World, 2020; Rubin, 2019; The Research Insights, 2020; WiseGuyReport, 2018). Thirty-five unique platforms were identified. All platforms that were identified by more than one source were chosen for review. This narrowed the list to twelve platforms. Further, one platform named Toluna was removed from consideration because it was used to pay participants for their survey responses, not for designing and distributing surveys. Another, Inqwise, no longer had an active website. Finally, companies SurveyPlanet and SogoSurvey were examined, but are not discussed below because they did not allow users to manipulate the number of response options for each item to create a 5-point scale.

All six sources referred to Qualtrics (Qualtrics, Provo, UT) in their discussion of survey tools. IBIS World (2020) also noted that Qualtrics is currently the largest web-based survey platform in the United States. Qualtrics uses 7-point Likert scales as a default, but is configurable to allow for 5-point scales as shown in Figure 1. By default, response options were presented in descending order (agree-disagree). In the 5-point

Likert scale, the modifier *Somewhat* was used in the response labels for the second and fourth response points. In the 7-point Likert scale, *Agree* and *Somewhat agree* (and *Somewhat disagree* and *Disagree*) were used for the second and third (and fifth and sixth) response points, implying that *Agree* (or *Disagree*) is closer to *Strongly agree* (or *Strongly disagree*) than *Somewhat agree* (or *Somewhat disagree*) is.

Figure 1 Presentation of 7-Point and 5-Point Likert Scales in Qualtrics

SurveyMonkey (Survey Monkey, San Mateo, CA) was also identified by each of the six reports. This platform provided no modifier to the second and fourth response options as shown in Figure 2, but applied the same labels as Qualtrics for the center and end-point options. SurveyMonkey was also designed to present Likert scales in a descending order. The inconsistency in the 5-point Likert scale labels presented in SurveyMonkey and the 5-point and 7-point Likert scale labels presented in Qualtrics is troublesome because the survey respondents of the different systems may end up selecting different options due to the different meaning associated with the different wording presented in the response scales.

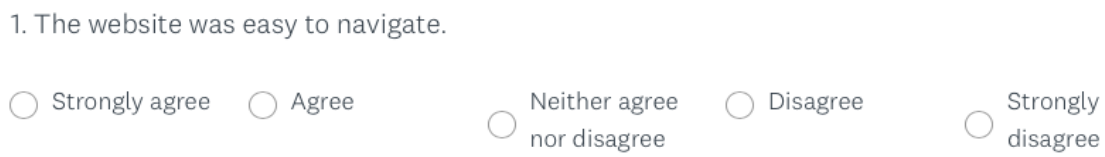


Figure 2 Presentation of Likert Scale on SurveyMonkey

Confirmit (New York, NY) was the only platform that did not offer users any default response options for Likert items and instead requires users to craft their own labels. The remaining four platforms each used an ascending response order and applied no modifying label to second and fourth response options. It is unknown which method is optimal or how much the use of different adverbs impacts participant responses.

Based on the different features available in different web-based survey tools reviewed, I decided to use Qualtrics to develop the survey instrument for this research as it provides the most appropriate web-based survey environment among all for investigating the research questions.

Survey Instrument

For the survey instrument, all items used a slider style scale. The slider bar was placed at the center of the scale and participants were asked to move the slider bar to the location that they believed best represented the response option questioned. The first screen of the instrument asked participants to perform this procedure for *Disagree* and *Agee* without adding an adverb and with the adverbs *Moderately, Slightly*, and *Somewhat* in front of *Disagree* and *Agree* (see APPENDIX A). The first four items were for testing the no-use and use of adverbs in front of the *Disagree* anchor, and the next four were for testing the no-use and use of adverbs in front of the *Agree* anchor—all presented in ascending order. End-point anchors of *Strongly Disagree* and *Strongly Agree* were used

for each item with a midpoint anchor of *Neutral*. For example, item 2 (see Figure 3) stated, “Where do you think ‘***Moderately* Disagree**’ should be placed on the continuum? Move and place the slider to indicate it.”

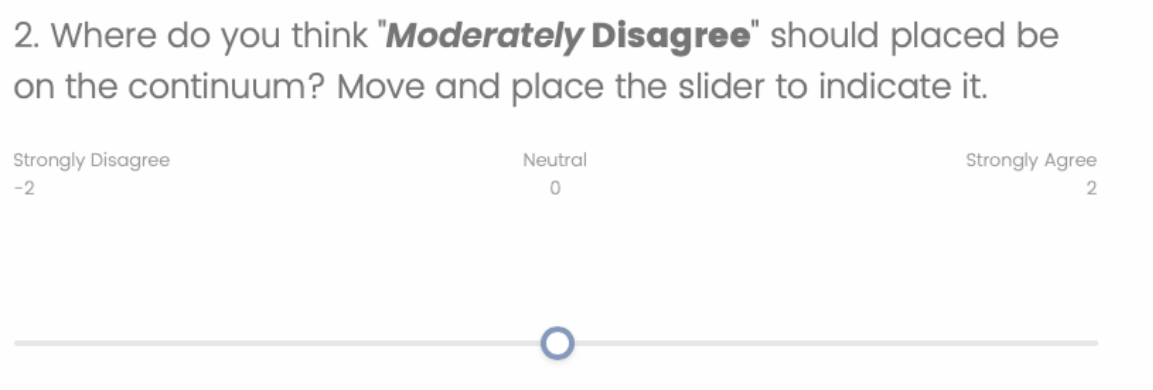


Figure 3 Presentation of Survey Item

The second screen of the instrument presented another eight items that were identical to the first eight, but with anchors presented in descending order. As discussed in Chapter 2, the response order could influence respondents’ ways to react to and select their response options. Thus, the ascending and descending orders of the response options were separately tested to assess if response order effects are present, and if so, how much the response order influences the data. The third screen of the instrument collected participants’ gender, age, and native English speaker status.

Procedure

Potential participants were recruited via email using the OPWL program's listserv type announcement system. This reached 327 people, including current graduate students and recent alumni (n = 316), as well as faculty members (n = 11). After reading the email message invitation, participants voluntarily clicked the survey link to complete an anonymous survey with 16 items (see APPENDIX A). The survey recruitment email (APPENDIX B**)** served as an informed consent form, as required by the Institutional Review Board (IRB) in order to protect human research subjects. Any respondents who were not comfortable participating in the survey were able to opt out of participation by simply not following the survey link or not completing any survey items. The survey was open to potential participants from October 8th, 2020 through October 27th, 2020.

Data Analysis

The main purpose of the study is to understand the degree to which the use of these adverbs in the intermediate anchors make the Likert scale become close to an interval scale when used in the web environment. In other words, the primary research question of this study (Research Question 1) is to analyze if adding an adverb to *Disagree* and *Agree* helped the respondents' perceptions on the second and fourth adverb-added anchors of the 5-point Likert scale be close to -1 and +1, respectively (where *Strongly Disagree* is -2, *Neutral* is 0, and *Strongly Agree* is +2). To answer Research Question 1, the following steps were taken:

1. Collected participants' marked data in two-digit numbers after the decimal point (e.g., -1.24, +0.88).

2. Analyzed the 95% confidence intervals of the marked data against the interval values for -1 and +1, as shown in Figure 4.

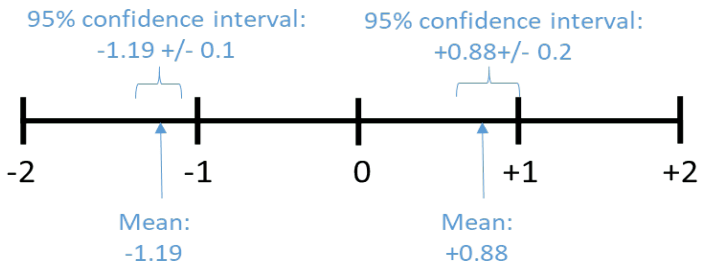


Figure 4 Illustration of Confidence Intervals

A secondary purpose of the research was to study the effects of response order on collected data when the response options were presented in ascending and descending orders (Research Question 2). This was accomplished by performing paired sample t-tests to compare the paired data for each response label when presented in ascending vs. descending order. All statistical analyses were performed with Excel.

CHAPTER FOUR: RESULTS

Survey Participants

Data Screening

The survey recruitment email was sent to a total of 327 people; 109 of them participated in this survey (a 33.3% return rate). Both Qualtrics metadata and participant responses were used to remove and disqualify participants from survey analysis. Any respondent that left an item unanswered or used the wrong side of the continuum more than once was excluded from further analysis. Ten cases were excluded because they left more than one item unanswered, and another 23 responses were excluded because they used the incorrect side of the continuum on at least one response. Three cases were removed because their responses included at least one instance in which they because they indicated that a response label should be placed at the end points of -2 or +2 where the anchors *Strongly Disagree* and *Strongly Agree* were located. One case was removed because one of the responses was on the 0 mark where *Neutral* is located (which may imply that the respondent did not move the slider). Finally, two cases were removed because they did not meet the research population criteria of being a native English-speaker. This left 70 responses available for analysis.

Characteristics of Survey Sample

The 70 participants were primarily females in their 30s and 40s (see Table 5). A series of independent samples t-tests showed no significant differences in responses between male and female participants ($p > 0.00625$, see Appendix C). Test results are

Table 5 Demographic Characteristics of Survey Participants (n = 70)

Gender	**Count**	**Percentage of Sample**
Female	51	73%
Male	18	26%
Do not want to report	1	1%
Age Group		
20s	9	13%
30s	19	27%
40s	20	29%
50s	13	18%
60 or older	9	13%

Survey Results

Results for Research Question 1

The primary research question was: Does adding an adverb (such as *Moderately, Slightly, or Somewhat*) to *Disagree* and *Agree* in the 5-point Likert scale influence people to perceive the scale to be closer to an interval scale, when administered in a web-based environment? Its null hypothesis was: Adding an adverb such as *Moderately, Slightly, or Somewhat* to *Disagree* and *Agree* in the 5-point Likert scale does not influence people to perceive the scale as being closer to interval scale.

To answer the research question 1 and test its null hypothesis, mean scores, standard deviations, and 95% confidence intervals were created using survey results for items presented in ascending order. Upon reviewing the mean scores against the 95% confidence intervals (Table 6), it was found that *Moderately Disagree* were not significantly far from -1, and that *Agree* (without an adverb) and *Moderately Agree* were not significantly far from +1, implying that the use of these labels as the 2nd and 4th anchors in the Likert scale makes it an interval-type Likert scale. On the other hand, the respondents' perception of *Disagree* was significantly far from -1 and toward *Strongly*

Disagree. Also, adding the adverbs *Somewhat* and *Slightly* to *Disagree* and *Agree* also moved the perceived values significantly far from -1 and +1, respectively, and toward *Neutral*; thus, these labels make the Likert scales ordinal scales.

Table 6 Conceptual Meaning of Response Labels Presented in Ascending Order

	Mean	**Standard Deviation**	**95% Confidence Interval**	**Significantly Different from -1 or +1?**
Disagree	-1.08	0.31	[-1.01] - [-1.15]	Yes
Moderately Disagree	-0.95	0.34	[-0.87] - [-1.03]	No
Somewhat Disagree	-0.58	0.31	[-0.51] - [-0.65]	Yes
Slightly Disagree	-0.40	0.16	[-0.36] – [-0.44]	Yes
Slightly Agree	0.46	0.19	[0.42] – [0.51]	Yes
Somewhat Agree	0.63	0.30	[0.56] – [0.70]	Yes
Moderately Agree	0.96	0.32	[0.88] – [1.03]	No
Agree	1.08	0.37	[0.99] – [1.17]	No

Based on the findings, the null hypothesis 1 was partially retained and rejected:

- Rejected: "H_0: Adding an adverb such as *Moderately* to *Disagree* and *Agree* in the 5-point Likert scale does not statistically significantly influence people to perceive the scale as being closer to interval scale." Adding *Moderately* to *Disagree* and *Agree* did influence people to perceive the scale as being closer to interval scale.
- Retained: "H_0: Adding an adverb such as *Somewhat* to *Disagree* and *Agree* in the 5-point Likert scale does not statistically significantly influence people to perceive the scale as being closer to interval scale."

- Retained: "H_0: Adding an adverb such as *Slightly* to *Disagree* and *Agree* in the 5-point Likert scale does not statistically significantly influence people to perceive the scale as being closer to interval scale."

Also, interestingly, the assumption I had about the *Disagree* and *Agree* anchors being ordinal anchors was not fully supported by the data. While *Disagree* was shown to be significantly different than -1 (i.e., characterized as an ordinal-type anchor), *Agree* was not significantly different than +1 (i.e., satisfied as an interval-type anchor).

Results for Research Question 2

Research question 2 stated, "Does the order of response options in the Likert scale (ascending vs. descending) make a difference in people's perceptions as tested in Research Question #1?"

First, the second eight items presented in descending order generated results similar to the first eight items presented in ascending order (see Table 7); both *Moderately Agree* and *Moderately Disagree* were not significantly far from +1 and -1, respectively. All other anchors were significantly far from +1 or -1. One difference between the two response orders was that when presented in ascending order, *Agree* was not significantly different from +1; however, when presented in descending order, *Agree* was significantly far from +1, moving toward *Strongly Agree* (which was placed at the fat left side). This could be related to the left-side selection bias observed in response scales presented in descending order; perhaps, respondents tend to perceive *Agree* to be more positive when it is presented on the left side.

Table 7 Conceptual Meaning of Response Labels Presented in Descending Order

	Mean	Standard Deviation	Confidence Interval (+/-)	Significantly Different from (+/-1)?
Agree	1.17	0.39	[1.07] – [1.26]	Yes
Moderately Agree	0.99	0.34	[0.92] – [1.07]	No
Somewhat Agree	0.63	0.30	[0.56] - [0.70]	Yes
Slightly Agree	0.40	0.20	[0.35] – [0.45]	Yes
Slightly Disagree	-0.46	0.16	[-0.42] – [-0.50]	Yes
Somewhat Disagree	-0.66	0.29	[-0.59] – [-0.73]	Yes
Moderately Disagree	-0.98	0.33	[-0.91] – [-1.06]	No
Disagree	-1.09	0.37	[-1.01] – [-1.18]	Yes

Next, paired sample t-tests were used to compare the means of each response label presented in ascending order against the same label presented in descending order. T-tests (parametric tests) were used because the data were approximately normally distributed; all 8 sets of data were not highly skewed ($-1 < \text{Skewness} < +1$) (Morgan et al., 2013). The risk of a type 1 error of incorrectly rejecting a null hypothesis increases any time multiple hypothesis are being compared; thus, it is suggested that researchers adjust for this risk using a Bonferroni Correction (Armstrong, 2014) in which the proposed alpha of .05 is divided by the number of comparisons being made. Using this method, conducting eight paired sample t-tests created a need for using the alpha value of 0.00625 instead of 0.05. Any test results that have a *p*-value of less than 0.00625 will indicate that response order does make a difference in people's perceptions of the meaning of that particular response anchor. Alternatively, a *p*-value of 0.00625 or above

indicates that response order did not change their perceptions of that anchor. A summary of the paired sample t-test results is shown in Table 8.

Table 8 Paired Sample T-Test Results

	Mean (Ascending Order)	**Mean (Descending Order)**	**df**	**t-stat**	**P-value (2-tailed)**	**Significantly different?**
Disagree	-1.08	-1.09	69	0.69	0.4950	No
Moderately Disagree	-0.95	-0.98	69	1.18	0.2421	No
Somewhat Disagree	-0.58	-0.66	69	3.53	0.0007	Yes
Slightly Disagree	-0.40	-0.46	69	2.83	0.0060	Yes
Slightly Agree	0.46	0.40	69	2.62	0.0108	No
Somewhat Agree	0.63	0.63	69	0.14	0.8890	No
Moderately Agree	0.96	0.99	69	-1.34	0.1846	No
Agree	1.08	1.17	69	-3.33	0.0001	Yes

Five of the eight response labels tested showed no statistical difference in the means of response labels presented in ascending versus descending order. *Somewhat Disagree, Slightly Disagree,* and *Agree* each had mean scores in which responses presented in ascending order were significantly different from those presented in descending order. As noted above, the outcome associated with *Agree* presented in descending order seems to support the left-side selection bias, resulting in a higher mean score. The reason for the outcomes associated with *Somewhat Disagree* and *Slightly Disagree* is unclear and I was unable to find any theoretical support for this finding. It is possible that this outcome was the result of sampling error or technology related issues.

Other Findings

While reviewing the data, an unexpected interesting finding emerged. As reported above, the response labels *Somewhat Disagree* and *Somewhat Agree* were close to *Neutral*. It was noticed that when *Neutral* was ignored, the two labels *Somewhat Disagree* and *Somewhat Agree* were placed in locations that are approximately equally distanced from *Strongly Disagree* and *Strongly Agree*, respectively. This was especially true when they were used in descending-ordered scales. The intervals between two consecutive response points in this descending-ordered 4-point Likert scale (*Strongly Agree, Somewhat Agree, Somewhat Disagree, Strongly Disagree*) were, on average, 1.37, 1.29, and 1.34 (see Figure 5). To confirm this observation, a single factor ANOVA was performed, and it revealed a non-significant difference among these three sets of distance [$F(2, 207) = 0.86$, $p = 0.43$)], suggesting this descending-ordered 4-point Likert scale as an interval scale. On the other hand, the ascending-ordered 4-point Likert scale (*Strongly Disagree, Somewhat Disagree, Somewhat Agree, Strongly Agree*) revealed unequal distances between anchors [$F(2, 207) = 4.72$, $p < .01$].

Figure 5 Somewhat Agree and Somewhat Disagree Presented in Descending Order

CHAPTER FIVE: DISCUSSION OF FINDINGS AND CONCLUSIONS

Summary

The primary purpose of this study was to answer the following research question, "Does adding an adverb (such as *Moderately, Slightly, or Somewhat*) to *Disagree* or *Agree* in the 5-point Likert scale influence people to perceive the scale to be closer to an interval scale, when administered in a web-based environment?" Exploring this subject is worthwhile because Likert scales are considered to be an ordinal measurement, but in practice these scales are often tested and analyzed as an interval measurement. Identifying which sets of modified anchors make the Likert scale close to interval could be helpful to those wishing to perform parametric analysis on Likert scale data.

This thesis studied the use of different adverbs in Likert scales using a survey with 16-items via the Qualtrics online survey platform. Each item used a slider bar, which was placed at the center of the scale (the *Neutral* position). Participants were asked to move the slider bar to the location where that they believed best represented the response option questioned. Participants were asked to perform this procedure with the adverbs *Moderately*, *Somewhat*, and *Slightly* in front of *Agree* and *Disagree*, as well as no adverb, first presented in ascending order, and then again in descending order. Confidence intervals were then calculated to indicate whether any anchors were conceptually close to ±1 on a bipolar 5-point Likert scale. Based on the data, it was found that applying the adverbs *Somewhat* or *Slightly* to *Disagree* and *Agree* made the anchors significantly far from -1 and +1, respectively, meaning that these adverbs are not appropriate for

generating interval data. However, *Moderately Disagree* and *Moderately Agree* were not significantly far from -1 and +1, respectively, and could be used to create an interval scale. *Disagree* and *Agree* without using any adverb also have potential to be suitable for an interval-type Likert scale (see Figure 6); however, only *Agree* presented in ascending order aligned with +1, and *Disagree* in ascending order and *Disagree* and *Agree* in descending order were not aligned with -1 or +1, respectively. Therefore, it would be safe to use *Moderately Disagree* and *Moderately Agree* when attempting to construct an interval level 5-point Likert scale, regardless of the response order used.

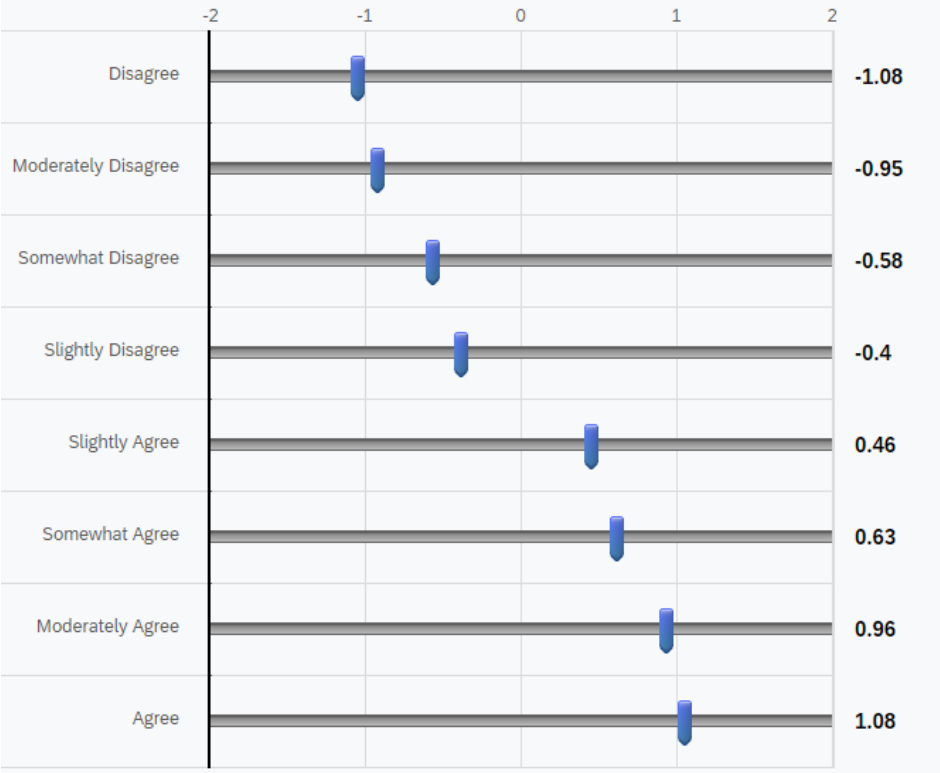


Figure 6 Mean Scores of Items Presented in Ascending Order

A secondary goal of this research was to investigate whether response order influences participants' perception of each response label. This was tested using paired sample t-tests to compare the scores of each response option when presented in ascending

versus descending order. Results showed that response order significantly impacted participants' perceptions of the response labels *Somewhat Disagree, Slightly Disagree,* and *Agree.* Each of the three had mean scores that were more extreme when presented in descending order (Figure 7). Especially, *Agree* was significantly more toward *Strongly Agree* when presented in descending order, suggesting an association with the left-side selection bias. There seem no clear explanations for the significant differences in participants' perceptions of *Somewhat Disagree* and *Slightly Disagree* when presented in different orders.

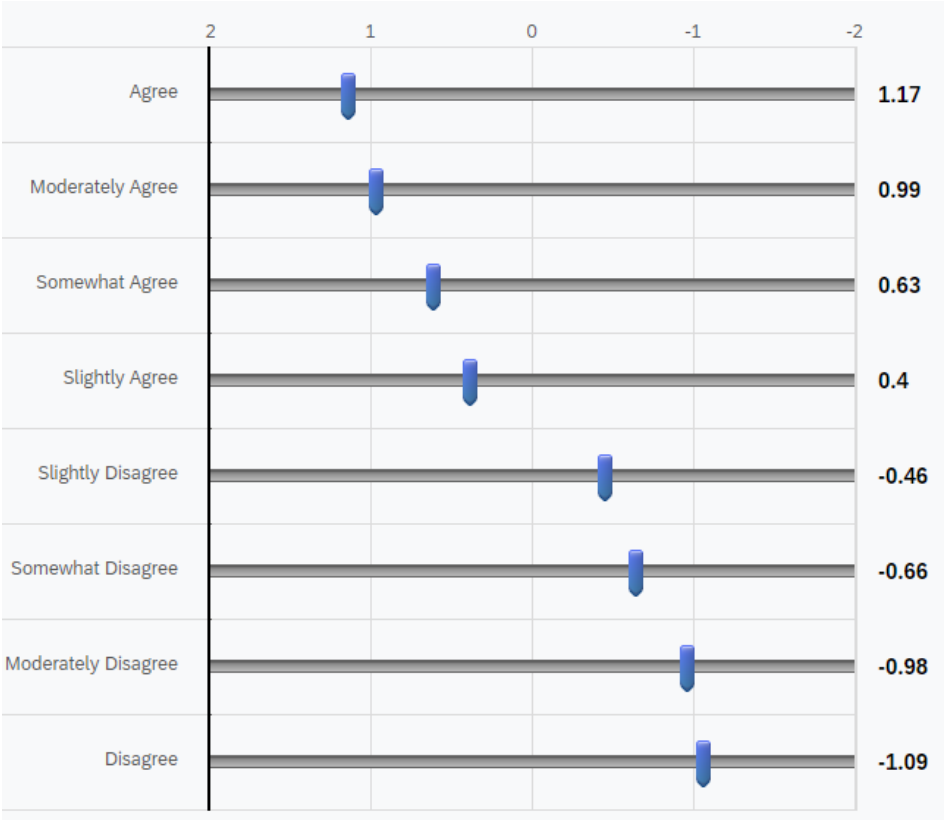


Figure 7 Mean Scores of Items Presented in Descending Order

Discussion of Findings

There were three primary findings from this thesis research. First, the results of this study indicated that using the adverb *Moderately* as a modifier for intermediate response options in 5-point Likert scales may help researchers create an interval level scale. Results indicated that the response label *Agree* could also be used for the 4th label when presented in ascending order. However, it would not be practical to use asymmetric labels such as: *Strongly disagree*, *Moderately disagree*, *Neutral*, *Agree*, *Strongly agree*. Consequently, it would be a better decision to add the modifier *Moderately* to both *Disagree* and *Agree* because it allows them to construct a symmetrically labeled scale. This finding is largely compatible with existing literature relating to the use of verbal descriptors in Likert scales. Spratto (2018) also recommended the use of the anchor *Moderately*, though this research was regarding 4-point scales. Casper's (2013) research supported the use of *Moderately* on the *Agree* side of a Likert scale, but no modifying adverb on the *Disagree* side. However, again, the value of using an asymmetrically labeled scale is questionable. Worcester and Burns (1975) did not study the use of the anchor *Moderately.*

Survey results found in this thesis showed that the adverbs *Somewhat* and *Slightly* convey an attitude that is rather close to neutral, and that both are better suited for an ordinal, not interval, 5-point scale. These findings were consistent with Spratto (2018) and Casper (2013) who also found that the adverb *Slightly* conveys the least extreme attitude of any adverb tested followed by *Somewhat*. However, this is different from Worcester and Burns (1975), who found that *Agree Slightly* and *Disagree Slightly* were appropriate for a 5-point interval scale. In fact, each response option tested by Worcester

and Burns had an average score that was significantly further from the neutral point than those tested by Spratto (2018), Casper (2013), or this thesis. For example, the mean score for *Disagree Slightly* presented in descending order was -1.03 in the Worcester and Burns study. The average score of *Slightly Disagree* presented in descending order was -0.46 in this thesis research. The reason for this discrepancy between is unclear, but could be partially explained by differences in methods of measurement. The survey by Worcester and Burns was conducted on paper and did not provide respondents with response labels for either the end-points or neutral point on the continuum, but instead asked the respondent to place the location of a response option on an unlabeled continuum.

Finally, while this study was intended to exclusively research 5-point Likert scales, it revealed some data that may also be useful to researchers wishing to design 4-point Likert scales. It was observed that the response labels *Somewhat Disagree* and *Somewhat Agree* could be used to construct a 4-point Likert scale (without *Neutral*) when presented in descending order. The intervals between two consecutive response points in this descending-ordered 4-point Likert scale (*Strongly Agree, Somewhat Agree, Somewhat Disagree, Strongly Disagree*) were approximately equal, suggesting it as an interval scale. On the other hand, the ascending-ordered 4-point Likert scale (*Strongly Disagree, Somewhat Disagree, Somewhat Agree, Strongly Agree*) revealed unequal distances between anchors, remaining as an ordinal scale. This finding is partially supported by Spratto (2018)'s research on 4-point Likert scales. While the research primarily recommended use of the *Completely Disagree/Agree* and *Moderately Disagree/Agree,* median data also showed equal intervals between *Strongly Disagree, Somewhat Disagree, Somewhat Agree,* and *Strongly Agree*.

Conclusions and Implications

The purpose of this research was to investigate the impact of adding an adverb to *Disagree* and *Agree* in a 5-point Likert scale on whether respondents would perceive the scale close to an interval scale. The following main conclusions and practical implications (recommendations) are drawn based on the data analyzed in this study.

First, the study started with an assumption that the 5-point Likert scale (without an adverb added to *Agree* and *Disagree*) would be an ordinal scale. Analysis of data supports this belief that using *Agree* and *Disagree* to create a 5-point creates an ordinal level scale because only *Agree* when presented in ascending order produced a confidence interval that included +1 or -1. Thus, it is not recommended to use *Agree* and *Disagree* when wishing to use the Likert scale as an interval scale.

Second, adding the adverb *Moderately* to *Agree* and *Disagree* was found to be appropriate for crafting an interval level Likert scale when presented in either ascending order (*Strongly Disagree, Moderately Disagree, Neutral, Moderately Agree, Strongly Agree*) or descending order (*Strongly Agree, Moderately Agree, Neutral, Moderately Disagree, Strongly Disagree*). This makes the use of *Moderately Agree* and *Moderately Disagree* preferable over *Agree* and *Disagree* when survey designers wish to present the scale response options in descending order.

Third, survey results showed that *Somewhat* and *Slightly* each portray attitudes that are considerably more neutral than *Moderately* as well as *Agree* and *Disagree* with no adverb. This means that survey designers should keep in mind that adding *Somewhat* or *Slightly* to *Agree* and *Disagree* will provide them with an ordinal level 5-point scale and that they want to treat the data as such during their statistical analyses.

Fourth, when presented in descending order and after removing the *Neutral* anchor, adding *Somewhat* to *Agree* and *Disagree* that were presented in descending order resulted in the two anchors placed in positions on the continuum which made the scale an interval-level 4-point scale. Thus, when survey designers wish to create an approximately interval level 4-point scale, they may add *Somewhat* to *Agree* and *Disagree* and present the 4-point Likert scale in descending order.

Limitations

There were several limitations to this study. First, I was constrained in my ability to recruit survey participants. The initial plan for recruiting survey participants included in-person recruitment of potential respondents on a university campus. However, due to the spread of the COVID-19 virus, I decided to exclusively solicit survey participation online, which likely reduced the total sample size examined in this study. As an alternative, I used a convenience sample for data collection, composed of master's degree or certificate seeking graduate students or recent alumni from an Organizational Performance and Workplace Learning program, which does include a course on Survey Design topics. For this reason, this research should be considered exploratory in nature and generalization of this study's findings is limited.

Correspondingly, the responses of over 25% of participants were screened out. Over half of these responses were filtered out because they used the wrong side of the continuum to provide their responses. Again, this problem could have been partially mitigated if surveys were conduct in-person so the participants were able to ask questions and seek clarification regarding the design of the survey.

Also, because surveys were completed on a remote basis, there author of this thesis was unable to control which devices were used. It is possible that responses would have varied on the basis of whether they were using their phone, tablet, or computer to respond. Further differences may also exist between employees who answered using a mouse versus those who dragged responses along the continuum using their fingers.

The survey instrument only contained items with *Agree* and *Disagree* anchors and only included the adverbs *Moderately, Somewhat,* and *Slightly.* The survey was limited to only 16-items in order to manage the risk of survey fatigue from respondents, which could have led to more incomplete or less precise responses. It is unclear if these results can be generalized to other Likert-type scales such as a satisfaction scale.

Recommendations for Future Research

Based on existing limitations of this study, future research on the subject should use a larger and perhaps more diverse sample. Perhaps studying a different sample that included participants with a greater range of educational experience and exposure to survey design principles would yield different results and findings.

Further, future research could also include a greater range of response options. This study only focused on a small number of common intermediate response descriptions used in *agree-disagree* scales and used *Strongly* as a modifier for all end-points. Spratto's (2018) research indicated that extreme modifiers such as *Completely* and *Very Strongly* convey a more extreme attitude than *Strongly*, but it is unclear if participants would truly interpret these anchors differently when responding to Likert items in surveys.

Finally, this research presented questions to participants in the context of a 5-point scale where both the end- and mid-points were provided. Findings from this research may not be generalizable to scales with a different number of response options such as a 7-point scale or a scale without a midpoint. As the number of scale points increases, it becomes increasingly difficult to avoid conceptual overlap between response descriptions. It would be worthwhile to build on Casper's (2013) research, which included findings on 7-point scales that include multiple intermediate response options on either side of each scale.

REFERENCES

Armstrong, R. A. (2014). When to use the bonferroni correction. *The Journal of the College of Optometrists*, *34*(5), 502-508. https://doi.org/10.1111/opo.12131

Aument, K., & Conley, Q. (2018). Transforming theory to practice: Integrating evidence-based practices in human performance improvement. *Performance Improvement, 57*(3), 26-33. https://doi.org/10.1002/pfi.21724

Aziz, D. M. (2013). What's in a name? A comparison of instructional systems design, organizational development, and human performance technology/improvement and their contributions to performance improvement. *Performance Imrovement, 52*(6), 28-35. https://doi.org/10.1002/pfi.21355

Betts, L., & Hartley, J. (2012, April). The effects of changes in the order of verbal labels and numerical values on children's scores on attitude and rating scales. *British Educational Research Journal, 38*(2), 319-331. https://doi.org/10.1080/01411926.2010.544712

Borgatta, E. F., & Bohrnstedt, G. W. (1980, November). Level of measurement: Once over again. *Sociological Research & Methods, 9*(2), 147-160. https://doi.org/10.1177/004912418000900202

Business Industry Reports. (2020). *Global online survey software market report 2020.* https://www.businessindustryreports.com/report/148428/global-online-survey-software-market-report-2019

Business Wire. (2017). *Global online survey software market - forecasts from 2017 to 2022. https://www.businesswire.com/news/home/20170927005571/en/Global-Online-Survey-Software-Market-Report-2017-2022---Research-and-Markets*

Cangialosi, N., Odoardi, C., & Battistelli, A. (2020). A three-way interaction model of innovative behavior, task-related learning, and job characteristics. *Performance Improvement Quarterly, 33*(2), 153-172. https://doi.org/10.1002/piq.21322

Carifio, J., & Perla, R. (2008). Resolving the 50-year debate around using and misusing likert scales. *Medical Education, 42*, 1150-1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

Casper, W.C. (2013). *Constructing equal distance response anchors.* [Doctoral dissertation, Oklahoma State University]. https://shareok.org/bitstream/handle/11244/10962/Casper_okstate_0664D_12785.pdf?sequence=1

Chan, J. C. (1991). Response-order effects in likert-type scales. *Educational and Psychology Measurement*, *51*(3), 531-540. https://doi.org/10.1177/0013164491513002

Chyung, S. Y. (2019). *10-step evaluation for training and performance improvement.* Sage Publications.

Chyung, S.Y., Barkin, J.R., Shamsy, J.A. (2018a). Evidence-based survey design: The use of negatively worded items in surveys. *Performance Improvement, 57*(3). https://doi.org/10.1002/pfi.21749

Chyung, S.Y., Hutchinson, D., Shamsy, J.A. (2020). Evidence-based survey design: Ceiling effects associated with response scales. *Performance Improvement, 59*(6). https://doi.org/10.1002/pfi

Chyung, S. Y., Kennedy, M., & Campbell, I. (2018b). Evidence-based survey design: The use of ascending or descending order of likert-type response options. *Performance Improvement, 57*(9), 9-16. https://doi.org/10.1002/pfi.21800

Chyung, S. Y., Roberts, K., Swanson, I., & Hankinson, A. (2017). Evidence-based survey design: The use of a midpoint on the likert scale. *Performance Improvement, 56*(10), 15-23. https://doi.org/10.1002/pfi.21727

Clark, R. (2006). Evidence-based practice and professionalization of human performance technology. In J. A. Pershing, H. D. Stolovitch, & E. J. Keeps, *Handbook of Human Performance Technology* (pp. 873-898). Pfeiffer.

Dayan, Y., Paine, C. S., & Johnson, A. (2007). Responding to sensitive questions in surveys: A comparison of results from online panels, face to face, and self-completion interviews. *World Association for Public Opinon Research*, 1-16.

Duan, M. (2011). Application of data collection techniques by human performance technology practitioners. *Performance Improvement Quarterly, 24*(3), 77-100. https://doi.org/10.1002/piq.20118

Edmondson, D. R. (2005). Likert scales: A history. *Proceedings of the Biennial Conference on Historical Analysis and Research in Marketing (CHARM): The Future of Marketing's Past*, *12*(2005), 127-133. https://ojs.library.carleton.ca/index.php/pcharm/article/view/1613

Friedman, H. H., Herskovitz, P. J., & Pollack, S. (1993). The biasing effects of scale-checking styles on response to a Likert scale. *Proceedings of the American Statistical Association Annual Conference: Survey Research Methods: Volume 2,* 792-795.

Ghosh, R., Shuck, B., Cumberland, D., & D'Mello, J. (2019). Building psychological capital and employe engagement: Is formal mentoring a useful strategic human resource development intervention? *Performance Improvement Quarterly, 32*(1), 37-54. https://doi.org/10.1002/piq.21285

IBIS World. (2020, September 29). *Online survey software industry in the US - Market research report*. https://www.ibisworld.com/industry-statistics/market-size/online-survey-software-united-states/

International Society for Performance Improvement (ISPI). (2012). What is human performance technology? https://ispi.org/404.aspx?404;/content.aspx?id=54

Jakobsson, U. (2004). Statistical presentation and analysis of ordinal data in nursing research. *Scandinavian Journal of Caring Sciences, 18*, 437-440. https://doi.org/10.1111/j.1471-6712.2004.00305.x

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education, 38*, 1217-1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology, 7*(4), 396-403. https://doi.org/10.9734/BJAST/2015/14975

Kerro, P., & Lee, D. (2015). Slider scales and web-based surveys: A cautionary note. *Journal of Research Practice, 11*(1), 1-4.

Knapp, T. (1990, March/April). Treating ordinal scales as interval: An attempt to resolve the controversy. *Nursing Research, 39*(2), 121-123. https://doi.org/10.1097/00006199-199003000-00019

Krosnick, J. A., & Alwin, D. F. (1987). An evaluation of a cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51*(2), 201-219. https://doi.org/10.1086/269029

Kuzon, W., Urbanchek, M., & McCabe, S. (1996). The seven deadly sins of statistical analysis. *Annals of Plastic Surgery,* 37, 265-272.

Likert, R. (1932, June). A technique for the measurement of attitudes. *Archives of Psychology*, *22*(1932-33), 5-55.

Maddy, L., & Rosenbaum, L. (2018, October 5). Determining leadership levels within the dreyfus model. *Journal of Workplace Learning, 30*(8), 626-639. https://doi.org/10.1108/JWL-11-2017-0100

Maeda, H. (2015). Response option configuration of online administered Likert scales. *International Journal of Social Research Methodology, 18*(1), 15-26. https://doi.org/10.1080/13645579.2014.885159

Morgan, G. A., Leech, N. L., Gloeckner, G. W., & Barrett, K. C. (2013). *IBM SPSS for introductory statistics.* Routledge.

Mousa, M., Abdelgaffar, H. A., Chaouali, W., & Aboramadan, M. (2020, January 22). Organizational learning, organizational resilience, and the mediating role of

multi-stakeholder networks. *Journal of Workplace Learning, 32*(3), 161-181. https://doi.org/10.1108/JWL-05-2019-0057

Murray, J. (2013, September). Likert data: What to use, parametric or non-parametric? *International Journal of Business and Social Science, 4*(11), 258-264.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education, 15*, 625-632. https://doi.org/10.1007/s10459-010-9222-y

Park, Y., & Jacobs, R. L. (2011). The influence of investment in workplace learning on learning outcomes and organizational performance. *Human Resource Development Quarterly, 22*(4), 437-458. https://doi.org/10.1002/hrdq.20085

Pershing, J. A., Stolovitch, H. D., & Keeps, E. J. (2006). *Handbook of human performance technology.* Pfeiffer.

Rubin, R. (2019, June 8). *Best online survey tools for 2020*. PC Magazine. https://www.pcmag.com/picks/the-best-online-survey-tools

Simon, H. A. (1957). Models of a man: Social and rational. *Wiley*.

Spratto, E. M. (2018). *In search of equality: Developing an equal interval Likert response scale.* JMU Scholarly Commons. https://commons.lib.jmu.edu/diss201019/172

Steil, A. V., Cuffa, D. D., Iwaya, G. H., & Santos Pacheco, R. D. (2019, November). Percieved learning opportunities, behavioral intentions and employee retention in technology organizations. *Journal of Workplace Learning, 32*(2), 147-159. https://doi.org/10.1108/JWL-04-2019-0045

Stevens, S. S. (1946, June 7). On the theory of scales of measurement. *Science, 103*(2684), 677-680. http://www.jstor.org/stable/1671815

Britannica, T. Editors of Encyclopaedia (2021, February 2). Technology. Encyclopedia Britannica. https://www.britannica.com/technology/technology

The Research Insights. (2020). *Online survey market research: global status & forecast by geography, type & application (2016-2026).*

https://www.theresearchinsights.com/service-industries/Online-Survey-Software-Market-Research-Global-Status--Forecast-by-Geography-Type--Application-2016-2026-77139

Touvier, M., Mejean, C., Kesse-Guyot, E., Pollet, C., Malon, A., Castebon, K., & Hercberg, S. (2010). Comparison between web-based and paper versions of a self-administered anthropometric questionnaire. *European Journal of Epidemiology, 25*, 287-296. https://doi.org/10.1007/s10654-010-9433-9

Van Tiem, D. M., Moseley, J. L., & Dessinger, J. C. (2012). *Fundamentals of performance technology: A guide to improving people, process, and performance.* John Wiley & Sons.

Vandergoot, S., Sarris, A., Kirby, N., & Harries, J. (2020). Individual or organizational factors that influence transfer generalization and maintenance of managerial-leadership programs. *Performance Improvement Quarterly, 33*(2), 207-246. https://doi.org/10.1002/piq.21323

Vergnaud, A.-C., Touvier, M., Mejean, C., Kesse-Guyot, E., Pollet, C., Malon, A., . . . Hercberg, S. (2011). Agreement between web-based and paper versions of a socio-demographic questionnaire in the NutriNet-Sante study. *International Journal of Public Health, 56*, 407-417. https://doi.org/10.1007/s00038-011-0257-5

Vigderhous, G. (1977, January). The level of measurement "permissible" statistical analsysis in social research. *The Pacific Sociological Review, 20*(1), 61-72. Retrieved from https://www.jstor.org/stable/1388904

WiseGuyReports. (2018). *Global online survey software market research report 2018.* https://www.wiseguyreports.com/reports/3128898-global-online-survey-software-market-report-2018

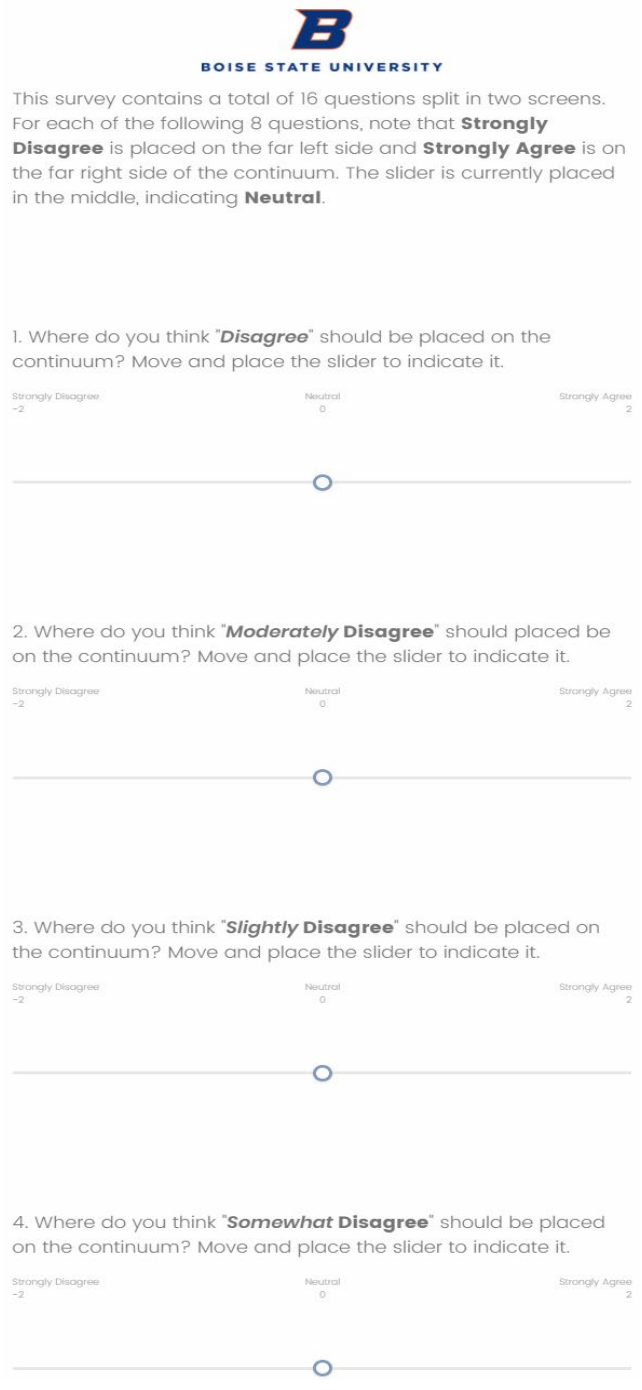
Wisshak, S., & Hochholdinger, S. (2020, June 4). Percieved instructional requirements of soft-skills trainers and hard-skills trainers. *Journal of Workplace Learning, 32*(6), 405-416. https://doi.org/10.1108/JWL-02-2020-0029

Worcester, R. M., & Burns, T. R. (1975). Statistical examination of the relative precision of verbal scales. *Journal of Market Research Society, 17*(3), 181-197.

Zadeh, S. A., & Azedeh, A. (2020). Customer-relationship management: performance assessment and improvement by an intelligent algorithm. *Performance Improvement Quarterly, 33*(2), 119-152. https://doi.org/10.1002/piq.21320

APPENDIX A

Survey Instrument


BOISE STATE UNIVERSITY
This survey contains a total of 16 questions split in two screens. For each of the following 8 questions, note that Strongly Disagree is placed on the far left side and Strongly Agree is on the far right side of the continuum. The slider is currently placed in the middle, indicating Neutral.
1. Where do you think "Disagree" should be placed on the continuum? Move and place the slider to indicate it.
Strongly Disagree -2
Neutral 0
Strongly Agree 2
2. Where do you think "Moderately Disagree" should placed be on the continuum? Move and place the slider to indicate it.
Strongly Disagree -2
Neutral 0
Strongly Agree 2
3. Where do you think "Slightly Disagree" should be placed on the continuum? Move and place the slider to indicate it.
Strongly Disagree -2
Neutral 0
Strongly Agree 2
4. Where do you think "Somewhat Disagree" should be placed on the continuum? Move and place the slider to indicate it.
Strongly Disagree -2
Neutral 0
Strongly Agree 2

5. Where do you think "***Agree***" should be placed on the continuum? Move and place the slider to indicate it.

Strongly Disagree -2 | Neutral 0 | Strongly Agree 2

6. Where do you think "***Moderately* Agree**" should be placed on the continuum? Move and place the slider to indicate it.

Strongly Disagree -2 | Neutral 0 | Strongly Agree 2

7. Where do you think "***Slightly* Agree**" should be placed on the continuum? Move and place the slider to indicate it.

Strongly Disagree -2 | Neutral 0 | Strongly Agree 2

8. Where do you think "***Somewhat* Agree**" should be placed on the continuum? Move and place the slider to indicate it.

Strongly Disagree -2 | Neutral 0 | Strongly Agree 2

→

For each of the remaining 8 questions below, note that **Strongly Agree** is placed on the far left side and **Strongly Disagree** is on the far right side of the continuum. The slider is currently placed in the middle, indicating **Neutral**.

9. Where do you think "***Disagree***" should be placed on the continuum? Move and place the slider to indicate it.

Strongly Agree 2 | Neutral 0 | Strongly Disagree -2

10. Where do you think "***Moderately* Disagree**" should be placed on the continuum? Move and place the slider to indicate it.

Strongly Agree 2 | Neutral 0 | Strongly Disagree -2

11. Where do you think "***Slightly* Disagree**" should be placed on the continuum? Move and place the slider to indicate it.

Strongly Agree 2 | Neutral 0 | Strongly Disagree -2

12. Where do you think "***Somewhat* Disagree**" should be placed on the continuum? Move and place the slider to indicate it.

Strongly Agree 2 | Neutral 0 | Strongly Disagree -2

13. Where do you think "***Agree***" should be placed on the continuum? Move and place the slider to indicate it.

14. Where do you think "***Moderately* Agree**" should be placed on the continuum? Move and place the slider to indicate it.

15. Where do you think "***Slightly* Agree**" should be placed on the continuum? Move and place the slider to indicate it.

16. Where do you think "***Somewhat* Agree**" should be placed on the continuum? Move and place the slider to indicate it.

What is your age group?

- ◯ the 20s
- ◯ the 30s
- ◯ the 40s
- ◯ the 50s
- ◯ 60 or older
- ◯ I do not want to report

What is your gender group?

- ◯ Male
- ◯ Female
- ◯ Other
- ◯ I do not want to report

Are you a native English speaker?

- ◯ Yes
- ◯ No
- ◯ I do not want to report

APPENDIX B

Survey Recruitment Script

Hello [name],

I am Douglas Hutchinson, a student with the Organizational Performance and Workplace Learning program at Boise State University.

I'm sure you have used the Likert scale in surveys. I am working on a thesis studying the impact of using different adverbs to describe Likert scale response options (e.g., *Moderately* agree, *Slightly* agree, etc.).

I am conducting an anonymous survey as a part of this project, and I hope you will complete the survey to help me complete my thesis project! Your responses will help OPWL practitioners, like you and me, learn more about optimal methods for using the Likert scale.

[**survey link**]

Your participation in this survey is voluntary. It will take only five minutes to complete the survey. This study involves no foreseeable serious risks on you. I ask you to answer all questions. However, if you don't feel comfortable answering any questions presented in the survey, you can skip those questions or stop completing the survey at any time.

If you have any questions or concerns about participation in this research, please contact me at XXXXX@u.boisestate.edu or (XXX) XXX-XXXX or my thesis advisor, Dr. Yonnie Chyung at XXXXX@boisestate.edu or (XXX) XXX-XXXX.

If you have questions about your rights as a research participant, you may contact the Boise State University Institutional Review Board (IRB), which is concerned with the protection of volunteers in research projects. You may reach the board office between 8:00 AM and 5:00 PM, Monday through Friday, by calling (XXX) XXX-XXXX or by writing: Institutional Review Board, Office of Research Compliance, Boise State University, 1910 University Dr., Boise, ID 83725-1138.

Thanks for your participation!
Douglas Hutchinson
OPWL Graduate Student
Boise State University

APPENDIX C

Differences in Responses Between Sexes

t-Test: Two-Sample Assuming Unequal Variances

Q1

	Male	*Female*
	-	
Mean	1.221666667	-1.02961
Variance	0.059285294	0.100848
Observations	18	51
Hypothesized Mean Difference	0	
df	39	
	-	
t Stat	2.645369045	
P(T<=t) one-tail	0.005850743	
t Critical one-tail	2.618863732	
P(T<=t) two-tail	0.011701485	
t Critical two-tail	2.890780195	

t-Test: Two-Sample Assuming Unequal Variances

Q2

	Male	*Female*
Mean	-0.87056	-0.96824
Variance	0.100288	0.124839
Observations	18	51
Hypothesized Mean Difference	0	
df	33	
t Stat	1.090773	
P(T<=t) one-tail	0.141637	
t Critical one-tail	2.642069	
P(T<=t) two-tail	0.283273	
t Critical two-tail	2.92098	

t-Test: Two-Sample Assuming Unequal Variances

	Q3	
	Male	*Female*
Mean	0.428333333	-0.39255
Variance	0.018391176	0.029207
Observations	18	51
Hypothesized Mean Difference	0	
df	37	
t Stat	-0.896170332	
P(T<=t) one-tail	0.187977806	
t Critical one-tail	2.625723458	
P(T<=t) two-tail	0.375955613	
t Critical two-tail	2.899699563	

t-Test: Two-Sample Assuming Unequal Variances

	Q4	
	Male	*Female*
Mean	-0.54	-0.59569
Variance	0.083988	0.100461
Observations	18	51
Hypothesized Mean Difference	0	
df	32	
t Stat	0.683597	
P(T<=t) one-tail	0.249575	
t Critical one-tail	2.646829	
P(T<=t) two-tail	0.49915	
t Critical two-tail	2.927184	

t-Test: Two-Sample Assuming Unequal Variances

	Q5	
	Male	*Female*
Mean	1.203333333	1.043333
Variance	0.114188235	0.146263
Observations	18	51
Hypothesized Mean Difference	0	
df	34	
t Stat	1.667056861	
P(T<=t) one-tail	0.05234509	
t Critical one-tail	2.637603686	
P(T<=t) two-tail	0.104690179	
t Critical two-tail	2.915162676	

t-Test: Two-Sample Assuming Unequal Variances

	Q6	
	Male	*Female*
Mean	0.961667	0.943529
Variance	0.112768	0.100235
Observations	18	51
Hypothesized Mean Difference	0	
df	28	
t Stat	0.199924	
P(T<=t) one-tail	0.421493	
t Critical one-tail	2.669479	
P(T<=t) two-tail	0.842985	
t Critical two-tail	2.95675	

t-Test: Two-Sample Assuming Unequal Variances

Q7

	Male	*Female*
Mean	0.505	0.445882
Variance	0.03965	0.036845
Observations	18	51
Hypothesized Mean Difference	0	
df	29	
t Stat	1.093044017	
P(T<=t) one-tail	0.141685737	
t Critical one-tail	2.663195723	
P(T<=t) two-tail	0.283371474	
t Critical two-tail	2.948540623	

t-Test: Two-Sample Assuming Unequal Variances

Q8

	Male	*Female*
Mean	0.629444	0.635686
Variance	0.086382	0.095361
Observations	18	51
Hypothesized Mean Difference	0	
df	31	
t Stat	-0.07643	
P(T<=t) one-tail	0.469782	
t Critical one-tail	2.651913	
P(T<=t) two-tail	0.939565	
t Critical two-tail	2.933814	

t-Test: Two-Sample Assuming Unequal Variances

	Q9	
	Male	*Female*
Mean	-1.21	-1.05627
Variance	0.132211765	0.13888
Observations	18	51
Hypothesized Mean Difference	0	
df	31	
t Stat	-1.532037073	
P(T<=t) one-tail	0.067828002	
t Critical one-tail	2.651912689	
P(T<=t) two-tail	0.135656003	
t Critical two-tail	2.933813566	

t-Test: Two-Sample Assuming Unequal Variances

	Q10	
	Male	*Female*
Mean	-0.89444	-1.00804
Variance	0.112061	0.101684
Observations	18	51
Hypothesized Mean Difference	0	
df	29	
t Stat	1.25296	
P(T<=t) one-tail	0.110115	
t Critical one-tail	2.663196	
P(T<=t) two-tail	0.220229	
t Critical two-tail	2.948541	

t-Test: Two-Sample Assuming Unequal Variances

	Q11	
	Male	*Female*
Mean	0.491666667	-0.4449
Variance	0.038167647	0.022569
Observations	18	51
Hypothesized Mean Difference	0	
df	24	
t Stat	0.923734079	
P(T<=t) one-tail	0.182409719	
t Critical one-tail	2.700233126	
P(T<=t) two-tail	0.364819438	
t Critical two-tail	2.997008192	

t-Test: Two-Sample Assuming Unequal Variances

	Q12	
	Male	*Female*
Mean	-0.62889	-0.66647
Variance	0.085834	0.088575
Observations	18	51
Hypothesized Mean Difference	0	
df	30	
t Stat	0.465953	
P(T<=t) one-tail	0.322307	
t Critical one-tail	2.657355	
P(T<=t) two-tail	0.644615	
t Critical two-tail	2.940915	

t-Test: Two-Sample Assuming Unequal Variances

	Q13	
	Male	*Female*
Mean	1.285555556	1.124902
Variance	0.11359085	0.164589
Observations	18	51
Hypothesized Mean Difference	0	
df	36	
t Stat	1.644997555	
P(T<=t) one-tail	0.05433778	
t Critical one-tail	2.629452932	
P(T<=t) two-tail	0.10867556	
t Critical two-tail	2.904551629	

t-Test: Two-Sample Assuming Unequal Variances

	Q14	
	Male	*Female*
Mean	0.936667	1.007647
Variance	0.106588	0.114554
Observations	18	51
Hypothesized Mean Difference	0	
df	31	
t Stat	-0.78539	
P(T<=t) one-tail	0.219092	
t Critical one-tail	2.651913	
P(T<=t) two-tail	0.438184	
t Critical two-tail	2.933814	

t-Test: Two-Sample Assuming Unequal Variances

	Q15	
	Male	*Female*
Mean	0.453333333	0.377451
Variance	0.044529412	0.037427
Observations	18	51
Hypothesized Mean Difference	0	
df	28	
t Stat	1.339806681	
P(T<=t) one-tail	0.095540643	
t Critical one-tail	2.669479342	
P(T<=t) two-tail	0.191081287	
t Critical two-tail	2.956749986	

t-Test: Two-Sample Assuming Unequal Variances

	Q16	
	Male	*Female*
Mean	0.653889	0.62
Variance	0.077543	0.098724
Observations	18	51
Hypothesized Mean Difference	0	
df	33	
t Stat	0.428881	
P(T<=t) one-tail	0.335399	
t Critical one-tail	2.642069	
P(T<=t) two-tail	0.670797	
t Critical two-tail	2.92098	