

**RATING SCALES IN ACCOUNTING RESEARCH:  
THE IMPACT OF SCALE POINTS AND LABELS**

**Jared Eutsler**  
**University of Central Florida**  
**Kenneth G. Dixon School of Accounting**  
**P. O. Box 161400**  
**Orlando, FL 32816**  
**Phone: (407) 823-5837**  
**Fax: (407) 823-3881**  
**Email: jeutsler@ucf.edu**

**Brad Lang**  
**University of Central Florida**  
**Kenneth G. Dixon School of Accounting**  
**P. O. Box 161400**  
**Orlando, FL 32816**  
**Phone: (407) 823-5149**  
**Fax: (407) 823-3881**  
**Email: blang@ucf.edu**

**Acknowledgements:** We are thankful for valuable comments from Matthew Holt, Erin Nickell, Anis Triki, Greg Trompeter, and Martin Weisner as well as the research assistance of Rosemary Mantle and Matthew Marstaller. The authors also appreciate the gracious guidance and suggestions of Vicky Arnold.

**Keywords:** Rating Scales; Scale Points; Scale Labels; Likert Scale; Semantic Differential Scale.

**Data Availability:** Please contact the authors.

# **RATING SCALES IN ACCOUNTING RESEARCH: THE IMPACT OF SCALE POINTS AND LABELS**

## **ABSTRACT**

Rating scales are one of the most widely used tools in behavioral research. Decisions regarding scale design can have a potentially profound effect on research findings. Despite this importance, an analysis of extant literature in top accounting journals reveals a wide variety of rating scale compositions. The purpose of this paper is to experimentally investigate the impact of scale characteristics on participants' responses. Two experiments are conducted that manipulate the number of scale points and the corresponding labels to study their influence on the statistical properties of the resultant data. Results suggest that scale design impacts the statistical characteristics of response data and emphasize the importance of labeling all scale points. A scale with all points labeled effectively minimizes response bias, maximizes variance, maximizes power, and minimizes error. This analysis also suggests variance may be maximized when the scale length is set at 7 points. Although researchers commonly believe using additional scale points will maximize variance, results indicate increasing scale points beyond 7 does not increase variance. Taken together, a *fully labeled 7-point scale* may provide the greatest benefits to researchers. The importance of scale labels provides a significant contribution to accounting research as only 5 percent of accounting studies reviewed have reported scales with all points labeled.

## INTRODUCTION

Rating scales are one of the most widely used tools in behavioral research (Dillman, 2009) and decisions regarding their design have a potentially profound effect on findings. Scale attributes may affect results of regression analysis, analysis of variance, confirmatory factor analysis, and structural equation models (Dawes, 2008). Consequently, decisions regarding scale construction are critical in research design. While researchers are diligent during instrument creation, spending a great deal of time and effort on most design aspects, scale design is often considered ancillary. Moreover, despite their importance, accounting researchers have yet to investigate potential effects of scale design choices and, as a community, currently adopt a wide variety of scale design alternatives in their research.

The purpose of this paper is to explore the effects of scale design in accounting research. Scale design has yet to be studied in accounting and, although examined in other fields, many important aspects of scale design have not been fully investigated. Furthermore, investigating scale design choices specifically in the context of accounting decision making is important, as context (such as the unique environment provided in the study of accounting decisions) affects optimal design decisions (e.g. Lai et al., 2010; Masters, 1974; Weathers et al., 2005). This study concentrates on two primary characteristics of rating scales: the number of scale points and the corresponding labeling of the scale points (Dillman et al., 2009).

After reviewing literature of scale format research in other fields, an investigation of studies that employ rating scales and published in the top accounting journals is conducted to determine the prevailing tendencies of scale design attributes and the consistency with which they are applied. Analysis reveals substantial variation in scale points (ranging from 3 to 101 points with a mode of 7 points), and scale labels (endpoints, all points, endpoints and midpoint,

and various other) in recent accounting research. Based on this analysis, two experiments are administered to explore the impact scale design decisions have on the statistical properties of responses from participants. Specifically, these studies manipulate the number of scale points and corresponding labels to examine their effects on the frequency of responses, measures of normality, variances, statistical power, and error.

Results suggest that scale labels and scale points impact the statistical characteristics of response data and emphasize the importance of labeling all scale points. A scale with all points labeled effectively minimizes response bias, maximizes variance, maximizes power, and minimizes error. The results also suggest that variance may be maximized when the scale length is set at 7 points. Although researchers commonly believe using additional scale points will increase variance, results indicate increasing scale points beyond 7 does not increase variance. Taken together, a *fully labeled*, 7-point scale may provide the greatest benefits to researchers. The importance of scale labels over scale points provides a significant contribution to accounting research as only 5 percent of accounting studies reviewed have reported scales with all points labeled.

This study contributes to the literature by bringing to light an important subject that has had little focus in accounting research. Scale design decisions can affect response bias (Weijters et al. 2010; Kulas et al. 2008), statistical power (Churchill and Peter, 1984), Type I and Type II error rates, and estimation of significance or effect sizes in multivariate testing (Osborne and Waters 2002; Bray and Maxwell 1985). Findings of this analysis are important, as they provide accounting researchers empirical evidence of the outcomes of decisions in the scale design process. This study also contributes to the literature by providing specific recommendations for proper scale design in behavioral accounting research. At a minimum, this study should stimulate

a conversation regarding scale design choices and the methodological decisions of future accounting studies that employ rating scales. If, as a community, accounting researchers adopt more consistent scale design choices, comparability of results within the field would increase and perhaps improve the quality of research using rating scales.

## **BACKGROUND**

Rating scales are a cornerstone of experimental and survey research. Scales typically require respondents to choose one response option from several arranged in order by degree (Friedman and Amoo, 1999). In general, there are two types of rating scales: Likert-type scales and semantic differential scales (Cox III, 1980). The Likert-type rating scale (Likert, 1932) has played a major role in behavioral science research and continues to be the most commonly used design. Likert-type scales are typically 5- or 7-point scales where each point is labeled to indicate some extent of *agreement* with a particular statement. For example, the Likert scale presented in Figure 1 includes labels for each scale point as follows – (1) strongly disagree, (2) disagree, (3) neither agree nor disagree, (4) agree, and (5) strongly agree. Semantic differential scales, also known as construct specific or item specific scales, are categorized by the particular construct to be measured (Osgood, 1952). The semantic differential scale in Figure 1 measures favorability. Both examples in Figure 1 are bipolar scales or scales that vary across two dimensions. Semantic differential scales can also be unipolar. Overall, research agrees that Likert-type and semantic differential scale formats may be treated as functionally equivalent (Cox III, 1980).

<Insert Figure 1 Here>

Although scale design has not been research in accounting, rating scale format has been studied in the fields of marketing, education, and statistics. Friedman and Amoo (1999) identify

10 design factors researchers need to consider when creating scales.<sup>1</sup> These include two primary characteristics of all rating scales: the number of scale points and their corresponding labels (Dillman et al., 2009). Research into the efficacy of these specific design factors have produced mixed results.

### **Scale Points**

The proper number of scale points (also referred to as response alternatives, scale length, scale width, and scale coarseness) has been contested. To date, no definitive length has been agreed upon. Some studies declare that more scale points improve the variance and reliability of the scale (Churchill and Peter, 1984; Pearse, 2011). However, increasing scale points may have unintended consequences. Providing more points than participants can discriminate between is shown to lead to increases in variability without an increase in precision (Miller, 1956). “At a general level, a scale with the optimal number of response alternatives is refined enough to be capable of transmitting most of the information available from respondents without being so refined that it simply encourages response error” (Cox III, 1980, page 408). That is, the ideal number of scale points needs to be long enough for subjects to discriminate between response alternatives, but not so long as to be cognitively taxing, resulting in an increase in measurement error (Viswanathan et al., 2004).

Extant literature displays a substantial discrepancy in the recommended number of scale points. Dillman, et al. (2009), which is largely considered an essential reference guide of scale design in accounting research, recommend 5- or 7-point scales. Additional studies have determined the “*best*” number of scale points is three (Jacoby and Matell, 1971), five (Likert, 1932; Revilla et al., 2014), six (Green and Rao, 1970), seven (Miller, 1956), nine (Preston and

---

<sup>1</sup> These include number of points, connotation of labels, frequency of response alternatives, implicit assumptions of the question, forcing a choice, unbalanced rating scales, order effects of rating scales, direction of comparison, context effects, and type of overall evaluation question (Friedman and Amoo, 1999).

Colman, 2000; Cook and Beckman 2009), ten (Cummins and Gullone, 2000; Dawes 2008) and even 21 points (Pearse, 2011). Arguments for shorter scales include no improvement of discrimination (Jacoby and Matell, 1971), little additional information gained (Green and Rao, 1970), and limits in subjects' ability to differentiate (Miller, 1956; Viswanathan et al., 2004). In contrast, proponents of more scale points contend they increase variance (Cook et al., 2001), yield more information (Dawes, 2008), are more accurate (Cook and Beckman 2009), and are preferred by respondents (Preston and Colman, 2000).

The question arises as to why there is such variation in the recommendations of scale design research. One reason may be that ideal scale design is a function of various external factors (e.g. Lai et al., 2010; Masters, 1974; Weathers et al., 2005). That is, context matters. Optimal number of scale points depends on what stimulus is being evaluated (Lai et al., 2010), or how widely opinion is divided toward the content being measured (Masters, 1974). Respondent dispositional factors may also determine the number of scale points (Weathers et al. 2005). Alternatively, the number of scale points, within a realistic range, may not affect statistical attributes (Jacoby and Matell, 1971). Dawes (2008) finds no significant difference in means, variance, and measures of normality when examining 5-, 7-, and 10-point scales. Similarly, Felix (2011) studies 3-, 5-, 7-, and 9-point scales with endpoints labeled and found no effect on means, standard deviations, and skewness.

### **Scale Labels**

While many studies have examined aspects of the appropriate number of scale choices, relatively few have researched the influence of scale labels (also referred to as scale anchors). The studies that have examined labeling design generally compare either fully labeled or endpoint-only labeled scales. Dillman et al. (2009) argue against labeling only endpoints as the

meaning of the middle categories becomes open to interpretation. Thus, respondents may make different inferences leading to an increase in measurement error. Labeling all points gives the researcher more control over the signaled information and helps ensure respondents' similar interpretation. Labeling all points also leads to differences in frequency of selecting the midpoint and endpoints of the scales (O'Muircheartaigh et al., 1995). In contrast, Newstead and Arnold (1989) find that there is no difference between fully labeled and endpoint labeled scales. They conclude that no statistical basis exists for a preference of labeling, but concede that fully labeled scales may be appropriate for unfamiliar semantic differential scales. Other studies find that the types of labels do not affect variance (Huck and Jacko, 1974; Chang, 1997), means (Dixon et al., 1984) or reliability (Wyatt and Meyers, 1987).

### **Scale Design Theory**

Multiple opposing theoretical perspectives have been applied to predict the effects of scale characteristics. Cox III (1980) reviews two theoretical perspectives used to develop predictions based on response alternatives: the theory of information and the absolute judgment paradigm. The theory of information predicts that as more response categories are proposed, more information about the variable of interest can be obtained (Cox III, 1980). Conversely, the absolute judgment paradigm, which focuses on individuals' information processing capability, predicts that limited benefits are derived from increasing scale points (Cox III, 1980). Splitting the difference, Viswanathan et al. (2004) put forth a form of cognitive fit titled meaningful discrimination which purports the optimal scale points is based on the number of categories that individuals typically use in thinking about an attribute and will be more accurate than either a longer or shorter scale.



Various biases are also employed to understand effects of scale features. Response biases may introduce systematic error rather than reflect the true attitudes of the respondent and lead to misspecified data spread, correlation coefficients, and factor structure (Hui and Triandis 1989; Weijters et al. 2010). Two potential biases are extreme response (disproportionally selecting endpoints), and central tendency (over-use of the scale's midpoint). Kulas et al. (2008) refer to midpoint, when not labeled, as a dumping ground for unsure responses, which hurt the reliability and validity of scales.<sup>2</sup> While not specifically theory, these biases are advanced in the context of saliency to develop hypotheses on the effect of scale points and labels.

### **Accounting Research and Scales**

In order to examine the scale design choices of researchers in accounting literature, articles using rating scales from 2000 through April 2014 in seven leading journals that frequently publish behavioral research are examined: *The Accounting Review (TAR)*, *Journal of Accounting Research (JAR)*, *Contemporary Accounting Research (CAR)*, *Accounting, Organizations, and Society (AOS)*, *Auditing: A Journal of Practice and Theory (Audit)*, *Behavioral Research in Accounting (BRIA)*, and *Accounting Horizons (Horizons)*.<sup>3</sup> A keyword search for “point scale” and “Likert scale” identified 550 articles. Of these, 70 articles were deemed false positives; the remaining 480 are analyzed and coded.

Table 1 presents the raw counts and percentages comparing scale points and the labeling described in the journal article. In order of frequency, the most commonly used scales include 7 points (n=161, 34 percent), 11 points (n=105, 22 percent), 5 points (n=61, 13 percent), and 9 points (n=34, 7 percent). Lesser used scale lengths include 3, 6, 8, 10, 13, 14, 15, 21, 100, and

---

<sup>2</sup> Participants interpret midpoints to mean different things including “neither agree nor disagree”, “not applicable”, “undecided”, “don’t know”, “no opinion”, “tend to agree”, and “tend to disagree” (Raaijmakers, et al., 2000; Worcester and Burns, 1975).

<sup>3</sup> *TAR*, *JAR*, *CAR*, and *AOS* are chosen as they are the premier journals in accounting literature. *Audit*, *BRIA*, and *Horizons* are also analyzed as historically they publish a larger percentage of research utilizing response scales. *JAIE*, while also a premier journal, is omitted as it rarely publishes behavioral research.

101 points. Although there is some consensus on the standard number of points, as the top four scale lengths encompass 76 percent of the total population, there is still considerable variation among these choices.<sup>4</sup>

<Insert Table 1 Here>

Similarly, there is wide variation in how scales are labeled. Table 1 displays the frequency of labels by number of scale points. For the four most frequently used scales, 5, 7, 9, and 11 points, the overwhelming majority only label scale endpoints (70 percent, 80 percent, 68 percent, and 79 percent respectively). For 5-point scales, the second most common approach is labeling all points. The second most common approach for 7-, 9-, and 11-point scales is labeling three points (the endpoints and midpoint of the scale). More unorthodox labels include labeling quartiles, halves, and other increments that do not coincide with the scale's length. Across all scale lengths, only 5 percent label all scale points.

Similar to other fields, accounting researchers seem to suffer from a lack of consistency in scale point and label design. While these design choices may influence response patterns, almost no attention has been paid to such issues in accounting. To address this shortcoming, two experiments are conducted examining the effects of scale attributes within the context of accounting. The next sections present experiments investigating the impact of the most employed scales from accounting literature on statistical measures. Aligning with the vast majority of research in this area, this study does not propose hypotheses or research questions. Theoretical perspectives are contradictory and arguably not necessary in a methodological analysis.

---

<sup>4</sup> This variation is partially explained by the context. For example, 100- and 101-point scales (and to some extent 10 and 11-point scales) often reflect numerical questions assessing percentages (percentage of portfolio allocation, or a percent agreement).

## **EXPERIMENT 1**

Experiment 1 investigates how scale points and scale labels affect statistical properties of participant responses including frequency of response, variance, and measures of normality. A 4×2 between-subjects experiment is conducted manipulating the scale points (5, 7, 9, and 11 points) and scale labels (all points or only endpoints labeled).

### **Participants and Task**

Participants consist of undergraduate business students recruited from two large managerial accounting classes. A total of 384 participants completed the instrument. Participants, on average are 21 years old, juniors in college, and 58 percent are male. Two hundred and twenty-seven had completed an ethics course.

Participants are randomly assigned to one of the eight experimental conditions. After being thanked for their involvement, they read a realistic ethical vignette faced by accountants. Using an ethical dilemma put forth in Flory et al. (1992),<sup>5</sup> this study employs an ambiguous situation that may or may not be perceived as explicitly ethical. An uncertain scenario is selected to mitigate participants' responses from clustering. All participants receive the same ethics case.

### **Variables**

The independent variables are manipulated in the response scale of the dependent variable. The length of the scale used to measure the dependent variable (scale points) is manipulated at four levels (5, 7, 9, and 11 points). These lengths are the most common in accounting literature and consistent with optimal scale sizes in existing literature. The second independent variable manipulated is the labeling of scale points (scale labels). Either all scale points are labeled or only the endpoints are labeled. In the endpoint only manipulation, the

---

<sup>5</sup> Practical scenarios are developed by the Ethics Resource Center in Washington D.C. for the Institute of Management Accountants on videotape and converted to text by Flory et al., (1992). The most ambiguous ethical case, which involves an accountant evaluating questionable expense reporting of his manager, was selected.

scale's poles are anchored by *strongly disagree* and *strongly agree* on a horizontal continuum with disagree on the left and agree on the right. To maintain similarity across manipulations, *strongly* is also used as the descriptor for the endpoints when all points are labeled. When all scale points are labeled, the midpoint is similarly labeled *neither agree nor disagree* across all scale lengths.<sup>6</sup> Figure 2 illustrates those scales.

<Insert Figure 2 Here>

After reading the case scenario, participants were asked to answer the following question: *Based on the information described in the scenario, please indicate the extent to which you agree or disagree with the following statement: I believe Tom's actions are acceptable.* The manipulated scale, which was used to measure the dependent variable, immediately followed. The dependent variables analyzed are the statistical attributes of participant responses across the eight conditions for the frequency of responses, variance, and measures of normality.

## **Results**

Data is rescaled to possess equal upper limits in order to facilitate comparisons (Dawes 2008). All responses are rescaled to an 11-point length using the formula  $(\text{rating} - 1) \div (\text{number of response categories} - 1) \times 11$ .<sup>7</sup> With the exception of frequency of response, rescaled values are analyzed.

### ***Frequency of Response***

Response biases may introduce systematic error rather than reflect the true attitudes of the respondent and lead to misinterpretation of the data, including biasing data spread, correlation coefficients, and factor structure (Hui and Triandis 1989; Weijters et al. 2010). Table

---

<sup>6</sup> For this study, a focus group of researchers was formed to assist in deciding the proper degree descriptors.

<sup>7</sup> Dawes (2008) presents a formula to rescale data to a common score out of 100. We adjust this equation to scale 5-, 7-, and 9-points scales to 11-point scales by simply multiplying by 11 rather than 100. The results (untabulated) are similar using other rescaling methods.

2 illustrates the frequency of responses at the endpoints and midpoints by cell. Smaller percentages of responses at endpoints and midpoints may indicate less response bias. Results indicate labeling all points reduces the central tendency bias of participants to selecting the midpoint (14 percent for endpoint labeling, vs. 4 percent for all points labeled). Further, the number of responses that were not either on the endpoint or midpoint went from 61 percent to 72 percent indicating that responses were more evenly distributed across the scale when all points were labeled. When considering only the number of points on the scale, no clear relationship between number of scale points and responses on endpoints or the midpoint exists, suggesting no improvement in bias reduction when only the number of scale points is considered.

<Insert Table 2 Here>

### ***Difference in Variance***

Descriptive statistics, including variance, across the scale points and scale labels conditions are presented in Table 3 Panel A. A visual inspection reveals means for the 7-point scale appear to be highest in both of the labeling conditions. A common concern of researchers is variance. Researchers often defend the use of longer scales by arguing that increasing the number of scale points increases variance and power (Churchill and Peter, 1984). To test for any statistical differences in variance, a series of one-way ANOVAs are run using the case dependent variable and either the scale points (restricted to the points being compared) or scale labels condition as the independent variable. Levene's Test for Homogeneity of Variance (Levene, 1960) and Bartlett's Test (Bartlett, 1937) are used to assess differences in variances, with the lowest statistical p-value (Table 3 Panel B). Results of both pairwise tests of difference indicate that regardless of scale points, labeling all points of the scale significantly increases variance ( $p < 0.01$ ).

Regarding scale points, variance is maximized at 7 points, irrespective of scale labels (Table 3 Panel A). As shown in Panel B of Table 3, the results of difference in variance (across combined labeling conditions) indicate a significant increase in variance between 5- and 7-point scales (7.26 to 10.68;  $p = 0.04$ ). As the scale length increases from 7 to 9 points, the variance decreases, but this decline only approaches significance (10.68 to 8.05;  $p = 0.13$ ).<sup>8</sup> The variance tests do not indicate a significant increase in variance as the number of scale points increases from 9 to 11 points (8.05 to 9.79;  $p = 0.32$ ).<sup>9</sup> These results are intriguing as they are in direct contravention to prevailing beliefs of many researchers. The results of the tests of variance suggest that variance is maximized with a 7-point scale and additional scale points do not increase variance. In combination, the highest variance was obtained with the 7-point scale with all points labeled.

### *Measures of Normality*

Most statistical analyses are built on the assumption of a normal distribution of the data. While multivariate tests are usually robust to departures of normality, departures from normality are still linked to increased Type I and Type II error rates, especially with kurtosis (Olson, 1972; Bray and Maxwell 1985; Osborne and Waters 2002). Further, non-normal distributions can lead to incorrect estimation of significance or effect sizes in multivariate testing (Osborne and Waters 2002; Bray and Maxwell 1985), affecting the power of statistical tests.

The normality of the data produced in Experiment 1 is assessed by examining skewness and kurtosis for each of the manipulations of labels and scale points (cells).<sup>10</sup> Table 3 presents

---

<sup>8</sup> Although, both endpoint labeled and all point labeled conditions show a decrease in variance from 7 to 9 points (8.72 to 7.51 and 13.05 to 8.75, respectively), this decrease is only significant in the all-points labeled condition ( $p = 0.025$ ; untabulated).

<sup>9</sup> Beyond the proximate pairwise tests conducted on variance, all other pairwise combinations were tested as well. The variance for the 11-point scale was not significantly different than the 5-, 7-, or 9-point scales; all  $p$ -values are greater than 0.14.

<sup>10</sup> Another option to evaluate normality is to assess the joint impact of skewness and kurtosis via the Shapiro-Wilk Test (Razali and Wah 2011). Accordingly, Shapiro-Wilk's  $W$  is calculated for each cell, which indicates that no single cell deviates from a normal distribution.

statistics related to skewness and kurtosis. The skewness scores are all positive suggesting that all cells are positively skewed (i.e. skewed to the right). For the most part, the kurtosis scores are negative, indicating a platykurtic distribution (i.e. flat). This is of consequence as platykurtic distributions attenuate power (Stevens, 2012). Accordingly, Z-scores<sup>11</sup> for skewness and kurtosis are calculated in order to evaluate significant deviations from normal distributions. Z-scores greater than a critical value of 1.96 suggest a significant departure from normality. Only two cells do not have significant departures from normality for skewness. As shown with the lack of significance for the skewness Z-Scores, 5-point (endpoints labeled) and 7-point (all points labeled) scales produce data that is normally distributed. As the results are ambiguous, no conclusions are provided regarding the effect of scale points or labels on the normality of distributions.

<Insert Table 3 Here>

## **Conclusions**

The results of Experiment 1 suggest that scale design matters inasmuch that it has an impact on distributional properties of the data. Analysis of frequency counts suggest that labeling all points appears to produce data that is less affected by extreme response and central tendency biases. Additionally, labeling all points significantly increases variance regardless of scale points. Variance also appears maximized with 7-point scales; additional points do not increase variance. These findings have important implications for researchers designing behavioral instruments and the veracity of the underlying data.

## **EXPERIMENT 2**

Although the findings regarding variance in Experiment 1 begin to speak to the ability of researchers to detect results (power), more evidence is necessary. Accordingly, Experiment 2

---

<sup>11</sup> Z-scores are calculated by dividing the skewness or kurtosis score, by its respective standard error of skewness or kurtosis.

investigates the impact of scale labels and scale points on power and error. Four simple 2×1 experiments are conducted across four different versions of scale points and labels. Statistical output from four regressions, each a separate analysis of the experiment with a unique combination of scale points and labels, are compared to determine their effect on power and error.

### **Participants and Task**

Participants again consist of undergraduate business students recruited from managerial accounting classes. A total of 383 participants completed the instrument. Demographics are similar to Experiment 1.

The experimental materials are adapted from Riley et al. (2014). Participants evaluate the earnings potential of an investment after receiving an earnings release with abridged financial statements and explanations of key fluctuations. The case manipulates the company's earnings [EARNINGS] as positive or negative in relation to prior year's earnings.<sup>12</sup> The dependent variable captured in this experiment is the participant's assessment of how favorable (FAVORABLE) earnings will be next year. This 2×1 experiment is assessed across four different conditions of scale points (5 and 9 point) and labels (all or endpoints).

### **Variables**

The intent of this experiment is to assess the effect of scale points and labels on error and power, while minimizing through design any differences in variance (and effects that artificial variance has on power and error). Therefore, as the 7-point scale produces data with statistically significant higher variance in the first experiment, it is not used as an alternative. Rather, the 5-point scale is used as the lower end of the manipulation. For the second manipulation of scale

---

<sup>12</sup> The positive (negative) earnings results condition presents the current year financials as experiencing increases (decreases) in revenues, net income, and cash balances.



points, 9- and 11-point scales are considered. Although fully-labeled versions of these scales are present in academic literature, their use is rare. Ultimately, the 9-point scale is chosen as the second scale point manipulation as it is more practical and more likely to be used under both fully labeled and endpoint labeled conditions. Statistical tests of difference in variance from Experiment 1 yield no differences between 5- and 9-point scales ( $p = 0.61$ ); therefore, using these two lengths provide the benefit of examining the effects of scale points, while not confounding scale length with increased variance. As in Experiment 1, scale labels are manipulated between all points labeled or endpoints labeled (see Figure 3).

<Insert Figure 3 Here>

## Results

Experiment 2 investigates how scale design factors affect power and error by examining the ability of a simple regression model to detect how the experimental manipulation (EARNINGS) affects the dependent variable (FAVORABLE). The  $R^2$  values (for the four regressions across scale labels and points conditions) and t-statistics (for each regression's EARNINGS  $\beta$  coefficients) examining the effect of EARNINGS on FAVORABLE are used to assess power and error (levels of error, including measurement error proxied with  $1-R^2$ ). For this analysis, the dependent variable is rescaled to 9 points consistent with the method used in Experiment 1 (Dawes, 2008). Table 4 Panel A presents the descriptive statistics for the case including the means and the standard deviations.<sup>13</sup>

<Insert Table 4 Here>

---

<sup>13</sup> A fully crossed ANOVA (untabulated) featuring the experimental manipulation (positive results) and the two scale design manipulations (scale length and labels) reveal the main effect for our experimental manipulation (EARNINGS) is significant ( $p < 0.001$ ). Positive results made participants feel more favorable about future earnings.

### ***Measures of Power***

The results of the four regressions (each a different combination of scale design factors crossing labels and points) are presented in Table 4 Panel B<sup>14</sup>. The  $\beta$ s for EARNINGS are consistent across all four regressions, ranging from 1.73 to 2.45. Each regression shows this coefficient to be significant ( $p < 0.001$ ). Still, the t-values for each regression's coefficients vary. The t-value is used to assess power (as higher t-values correspond to higher p-values). Accordingly higher t-values in Panel B demonstrate increased power. The results demonstrate the importance of labels; labeling all points on the scale increases power, as the t-values have larger increases when the scales are fully labeled (4.82 to 6.88 for 5-point scales and 5.30 to 6.99 for 9-point scales). The results also indicate that power, as assessed by t-values of the coefficients, increases slightly as the number of scale points is extended from 5 to 9 points (6.88 to 6.99 in all points labeled, and 4.82 to 5.30 in endpoints labeled).

### ***Measures of Error***

This analysis uses  $1-R^2$  to proxy for error. Since the regression models only vary in the scale design factors tested (scale points and labels), an increase in unexplained variance ( $1-R^2$ ) is attributed to an increase in measurement error. Accordingly, the analysis of  $R^2$  suggests that labeling all points on the scale significantly decreases measurement error, wherein  $R^2$  values are maximized. The  $R^2$  values (Table 4 Panel B) indicate labeling all points maximizes the  $R^2$  (21.27 percent to 35.49 percent for 5-point scales, and 24.64 percent to 36.24 percent for 9-point scales). Additionally,  $R^2$  values experience a small increase as scale points increases from 5 to 9 points.

---

<sup>14</sup> Due to the randomization efforts in conducting the experimental task, cells ended up being unbalanced. In this analysis, all cells are randomly (using a random number generator) balanced with 44 response per cell (the lower limit of all 8 combinations), to prevent incidental bias of t-values or  $R^2$  values. Accordingly, each of the four regression has an  $n=88$ .

## Conclusions

The results from Experiment 2 suggest the importance of labeling all points on scales to increase power and decrease error in behavioral research. Labeling all points on the scale increases the power and decreases error significantly as evidenced by large increases in t-values and  $R^2$ . As in Experiment 1, results concerning scale points are a little more ambiguous.

## DISCUSSION

This paper examines the use of scales in accounting research. First, scales presented in accounting journals are assessed by examining the relative frequency and use of labels and scale points. While the most popular scale lengths are 5, 7, 9, and 11 points, significant variation in the number of scale points is identified. Similarly, although labeling the endpoints of scales is the most popular method, significant variation in how scales are labeled exists. Analysis suggests some commonalities in the scale design strategies employed by accounting researchers, however, considerable variation on the number of scale points and labels used still exist.<sup>15</sup>

Next, two experiments examine how scale design choices affect various statistical properties of data including response patterns, variance, measures of normality, power, and error. Results suggest that scale design matters; especially in how scales are labeled. First, power is maximized and error (including measurement error) is minimized when all points are labeled. Analysis of frequency counts finds that labeling all scale points decreases response bias by producing data less affected by extreme response and central tendency biases. Further, labeling all points on scales significantly increases variance. Interestingly, increasing the number of scale points does not appear to impact variance. This finding is in direct contravention with prevailing

---

<sup>15</sup> The review of accounting literature identified two additional areas of concern. First, some researchers employ multiple formats of both scale lengths and labeling within the same study, typically when combining various previously validated scales into one study. As Dillman et al. (2009) notes, it is important to be consistent when formatting scales. Second, some research articles identified do not provide information regarding scale length (4%) or how scales are labeled (9%). As a science, we could improve the quality of our research by documenting the types and attributes of scales used to improve our ability to replicate results (even if it is simply mentioned in a footnote).

beliefs as variance may be maximized with 7-point scales and additional points do not appear to increase variance. The central finding of this study is the importance of labeling all points on scales. Taken together, results suggest a *fully labeled* 7-point scale may provide the greatest benefits to researchers. Still, this recommendation may be naïve in that it disregards certain characteristics of the experimenters' research design. Context may override this recommendation. However, this analysis demonstrates that regardless of the number of scale points, the majority of accounting studies only label endpoints. Behavioral researchers may be able to improve their data set by labeling all points of their scale.

This study is not without limitations. First, the participants used in this study are all business students. Although business students are common participants, professionals may interpret and use scales differently than students. Second, as earlier stated, context is significant in scale design choices. Accordingly, two different accounting contexts (an ethics case and an investment case) are used. These scale design attributes could also be evaluated in alternative accounting contexts (AIS, managerial, tax, etc.) Thus, our review of existing literature and the use of scales in accounting literature provide for opportunities in future studies of scale design. Interesting future studies should include the impact of the implicit assumptions of the question, forcing a choice (no scale midpoint), and order effects of rating scales (disagree to agree). Other areas for study is the influence of anchoring on 0 instead of 1 as an endpoint or anchoring on positive and negative valences instead of only positive valences (-3 to +3 instead of 1 to 7). Individuals may have an aversion to non-positive numbers (i.e. prospect theory), which could cause a left skew of the data. Further methodological research concerning scale design would be beneficial for behavioral research in accounting.

## REFERENCES

- Bartlett, M. S. 1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*: 268-282.
- Bray, J. H., and M. E. Maxwell. 1985. *Multivariate Analysis of Variance* (No. 54). Sage.
- Chang, L. 1997. Dependability of anchoring labels of Likert-type scales. *Educational and Psychological Measurement* 57(5): 800-807.
- Churchill Jr, G. A., and J. P. Peter. 1984. Research design effects on the reliability of rating scales: A meta-analysis. *Journal of Marketing Research* 21(4): 360-375.
- Cook, D. A., and T. J. Beckman. 2009. Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Advances in Health Science Education* 14: 655–664.
- Cook, C. F., R. Heath, L. Thompson, and B. Thompson. 2001. Score reliability in web or internet-based surveys: Unnumbered graphic rating scales versus Likert-type scales. *Educational and Psychological Measurement* 61: 697-706.
- Cox III, E. P. 1980. The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research* 17(4): 407-422.
- Cummins, R. A., and E. Gullone. 2000. Why we should not use 5-point Likert scales: The case for subjective quality of life measurement. *Proceedings, Second International Conference on Quality of Life in Cities* (pp.74-93). Singapore: National University of Singapore.
- Dawes, J. 2008. Do data characteristics change according to the number of scale points used? *International Journal of Market Research* 50(1): 61-77.

- Dillman, D. A., J. D. Smyth, and L. M. Christian. 2009. Internet, mail, and mixed-mode surveys: The tailored design method. *New York, NY: John Wiley & Sons.*
- Dixon, P. N., M. Bobo, and R. A. Stevick. 1984. Response differences and preferences for all-category-defined and end-defined Likert formats. *Educational and Psychological Measurement* 44(1): 61-66.
- Felix, R. 2011. The impact of scale width on responses for multi-item, self-report measures. *Journal of Targeting, Measurement and Analysis for Marketing* 19(3/4): 153-164.
- Flory, S. M., T. J. Phillips Jr., R. E. Reidenback, and D. P. Robin. 1992. A Multidimensional Analysis of Selected Ethical Issues in Accounting. *The Accounting Review* 67(2): 284-302.
- Friedman, H. H., and T. Amoo, 1999. Rating the rating scales. *Journal of Marketing Management* 9(3): 114-123.
- Green, P. E., and V. R. Rao. 1970. Rating scales and information recovery. How many scales and response categories to use? *Journal of Marketing* 34(3): 33-39.
- Huck, S. W., and E. J. Jacko. 1974. Effect of varying the response format of the Alpert-Haber Achievement Anxiety Test. *Journal of Counseling Psychology* 21(2): 159-163.
- Hui, C. H., and H. C. Triandis. . 1989. Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20(3), 296-309.
- Jacoby, J., and M. S. Matell. 1971. Three-point Likert scales are good enough. *Journal of Market Research* 8(4): 495-500.
- Kulas, J. T., A. A. Stachowski, and B. A. Haynes. 2008. Middle response functioning in Likert-responses to personality items. *Journal of Business and Psychology*, 22(3), 251-259. doi: 10.1007/s10869-008-9064-2

- Lai, M., Y. Li, and Y. Liu. 2010. Determining the optimal scale width for a rating scale using an integrated discrimination function. *Measurement* 43: 1458-1471.
- Levene, H. 1960. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, I. Olkin et al. eds., Stanford University Press: 278-292.
- Likert, R. 1932. A technique for the measurement of attitude. *Archives of Psychology* 140: 1-55.
- Masters, E. R. 1974. The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement* 11(1): 49-53.
- Miller, G. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63: 81-97.
- Newstead, S. E., and J. Arnold. 1989. The effect of response format on ratings of teaching. *Educational and Psychological Measurement* 49(1): 33-43.
- O'Muircheartaigh, C., G. Gaskell, and D. B. Wright. 1995. Weighing anchors: Verbal and numeric labels for response scales. *Journal of Official Statistics* 11(3): 295-307.
- Olson, C. L. 1976. On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83(4), 579.
- Osborne, J., and E. Waters. 2002. Four assumptions of multiple regression that researchers should always test. *Practical assessment, research & evaluation* 8: (2), 1-9.
- Osgood, C. E. 1952. The nature and measurement of meaning. *Psychological Bulletin* 49(3): 197.
- Pearse, N. 2011. Deciding on the scale granularity of response categories of Likert type scales: The case of the 21-point scale. *The Electronic Journal of Business Research Methods* 9(2): 159-171.

- Preston, C. C., and A. M. Colman. 2000. Optimal number of response categories in the rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica* 104: 1-15.
- Raaijmakers, Q. A., A. V. Hoof, H. T. Hart, T. F. Verbogt, and A. M. Wollebergh. 2000. Adolescents' midpoint response on Likert-type scale items: Neutral or missing values? *International Journal of Public Opinion Research*, 12(2), 208-216.
- Razali, N. M., and Y.B. Wah. 2011. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics* 2(1): 21-33.
- Revilla, M. A., W. E. Saris, and J. A. Krosnick. 2014. Choosing the Number of Categories in Agree–Disagree Scales. *Sociological Methods & Research*, 43(1), 73-97.
- Riley, T. J., G. R. Semin, and A. C. Yen. 2014. Patterns of language use in accounting narratives and their impact on investment-related judgments and decisions. *Behavioral Research in Accounting* 26(1): 59-84.
- Stevens, J. P. 2012. *Applied multivariate statistics for the social sciences*. Routledge.
- Viswanathan, M., Sudman, S., and M. Johnson. 2004. Maximum versus meaningful discrimination in scale response: Implications for validity of measurement of consumer perceptions about products. *Journal of Business Research*, 57(2), 108-124.
- Weathers, D., S. Sharma, and R. W. Niedrich. 2005. The impact of the number of scale points, dispositional factors, and the status quo decision heuristic on scale reliability and response accuracy. *Journal of Business Research* 58(11): 1516-1524.



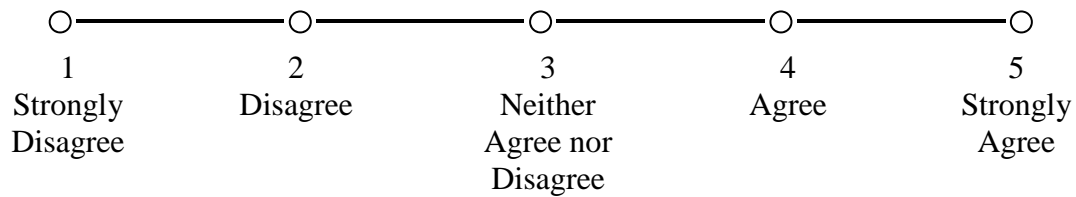
- Weijters, B., E. Cabooter, and N. Schillewaert. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing* 27: 236-247.
- Worcester, R. M., and T. R. Burns. 1975. A statistical examination of the relative precision of verbal scales. *Journal of the Market Research Society*, 17(3), 181-197.
- Wyatt, R. C., and L. S. Meyers. 1987. Psychometric properties of four 5-point Likert type response scales. *Educational and Psychological Measurement* 47(1): 27-35.

**Figure 1**  
**Examples of Likert and Semantic Differential Scales**

**Likert Scale**

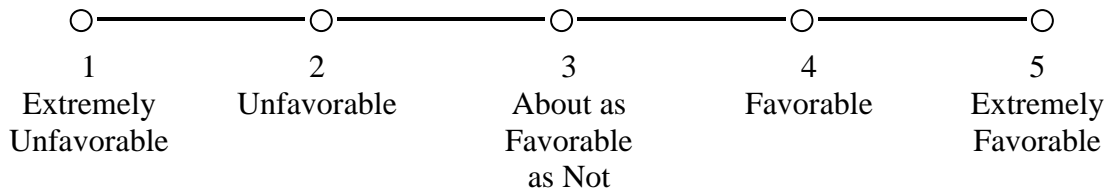
Based on the information described in the scenario, please indicate the extent to which you agree with the following statement:

I believe Tom's actions are acceptable.



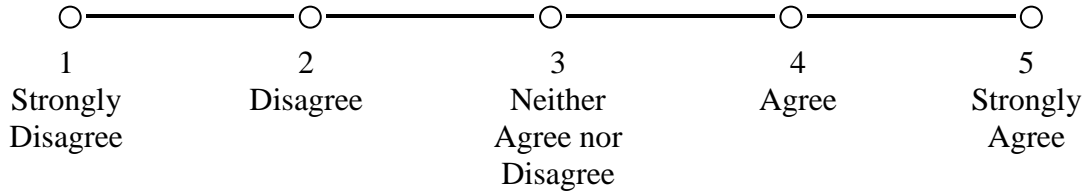
**Semantic Differential Scale**

As an investor, how favorable do you believe the company's financial results will be in the next year?

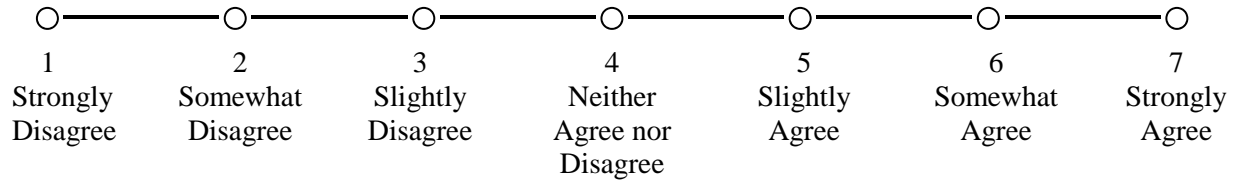


**Figure 2**  
**Scale Labels for Experiment 1**

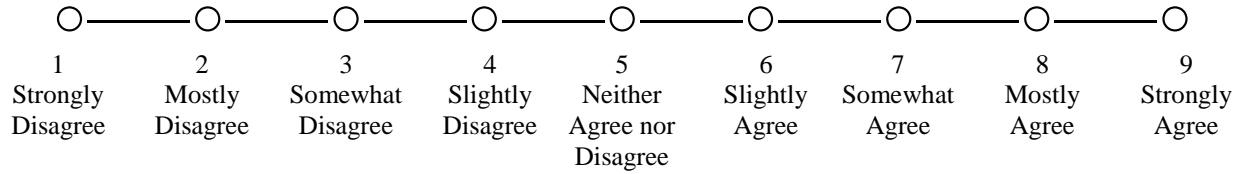
**5-Point Scale**



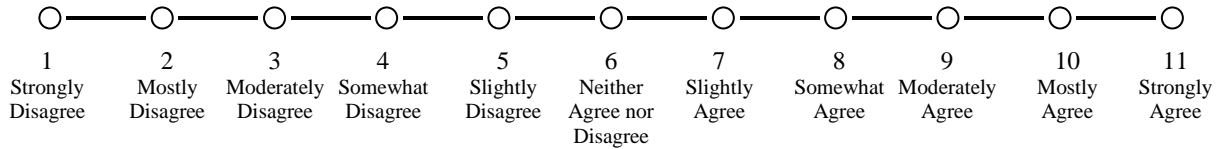
**7-Point Scale**



**9-Point Scale**



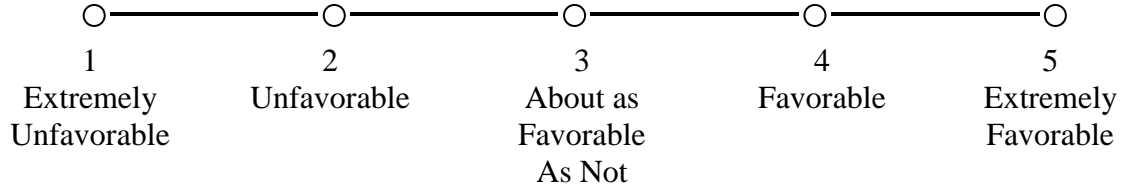
**11-Point Scale**



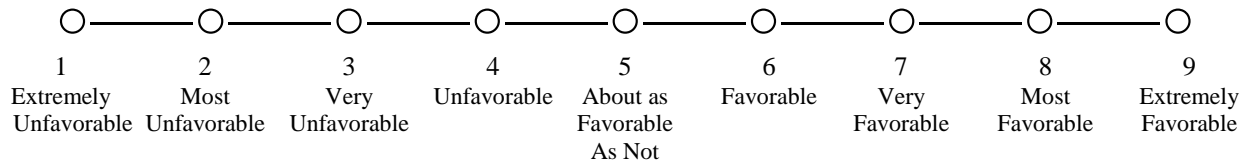
Note: Labels were presented as shown above when the manipulation included all scale points labeled. Only the endpoint labels were included for the endpoint only manipulation.

**Figure 3**  
**Scale Labels for Experiment 2**

**5-Point Scale**



**9-Point Scale**



Note: Labels were presented as shown above when the manipulation included all scale points labeled. Only the endpoint labels were included for the endpoint only manipulation.

**Table 1**  
**Historical Count of Scales in Accounting Research by Number of Scale Points and Labels**

| Scale Points        | Points Labeled |          |           |        |              | Total      |
|---------------------|----------------|----------|-----------|--------|--------------|------------|
|                     | All            | 3 points | Ends      | Other  | Not Reported |            |
| <b>3</b>            | 3              |          |           |        | 1            | 4 (1%)     |
| <b>5</b>            | 11             | 2        | 43        |        | 5            | 61 (13%)   |
| <b>6</b>            | 3              | 1        | 4         |        |              | 8 (2%)     |
| <b>7</b>            | 5              | 18       | 128       |        | 10           | 161 (34%)  |
| <b>8</b>            |                |          | 3         |        |              | 3 (1%)     |
| <b>9</b>            |                | 11       | 23        |        |              | 34 (7%)    |
| <b>10</b>           |                | 1        | 26        |        | 7            | 34 (7%)    |
| <b>11</b>           | 2              | 17       | 83        |        | 3            | 105 (22%)  |
| <b>13-21</b>        |                | 4        | 11        | 1      | 1            | 17 (3%)    |
| <b>100</b>          |                | 2        | 5         |        |              | 7 (1%)     |
| <b>101</b>          |                | 4        | 21        | 3      | 1            | 29 (6%)    |
| <b>Not Reported</b> |                |          |           |        | 17           | 17 (3%)    |
| <b>Total</b>        | 24 (5%)        | 60 (13%) | 347 (72%) | 4 (1%) | 45 (9%)      | 480 (100%) |

Note: Counts are obtained from articles appearing in TAR, JAR, CAR, AOS, Audit, BRIA, and Horizons from January 2000-April 2014.

**Table 2**  
**Frequency of Responses in Experiment 1 by Scale Points and Labels**

| Scale Points | Endpoints labeled |            |                  | All points labeled |           |                  | All participants |            |                  |
|--------------|-------------------|------------|------------------|--------------------|-----------|------------------|------------------|------------|------------------|
|              | End-points        | Mid-point  | All other points | End-points         | Mid-point | All other points | End-points       | Mid-point  | All other points |
| 5            | 30%               | 26%        | 44%              | 21%                | 0%        | 79%              | 25%              | 13%        | 62%              |
| 7            | 21%               | 12%        | 67%              | 30%                | 5%        | 65%              | 25%              | 9%         | 66%              |
| 9            | 27%               | 6%         | 67%              | 19%                | 4%        | 77%              | 23%              | 5%         | 72%              |
| 11           | 21%               | 14%        | 65%              | 26%                | 8%        | 66%              | 23%              | 11%        | 66%              |
| <b>Total</b> | <b>25%</b>        | <b>14%</b> | <b>61%</b>       | <b>24%</b>         | <b>4%</b> | <b>72%</b>       | <b>24%</b>       | <b>10%</b> | <b>66%</b>       |

Note: Table displays percentage of respondents choosing values of endpoints, the midpoint of the scale, or all other points for each scale length and labels condition.

Table 3  
Results from Experiment 1

**Panel A: Descriptive Statistics**

| Scale Labels      | Scale Points | Mean <sup>a</sup> | Variance | Skewness | Skewness Z-Score | Kurtosis | Kurtosis Z-Score |
|-------------------|--------------|-------------------|----------|----------|------------------|----------|------------------|
| <b>Endpoints</b>  | 5            | 2.87              | 5.58     | 0.35     | 1.00             | -0.65    | -0.92            |
|                   | 7            | 3.67              | 8.72     | 0.73     | 2.13*            | -0.14    | -0.19            |
|                   | 9            | 2.97              | 7.51     | 0.98     | 2.89*            | 0.50     | 0.71             |
|                   | 11           | 3.41              | 7.65     | 0.69     | 2.10*            | -0.08    | -0.12            |
| <b>All Points</b> | 5            | 3.57              | 8.85     | 0.87     | 2.52*            | -0.26    | -0.36            |
|                   | 7            | 4.05              | 13.05    | 0.54     | 1.48             | -1.05    | -1.40            |
|                   | 9            | 3.25              | 8.75     | 0.86     | 2.49*            | -0.38    | -0.53            |
|                   | 11           | 3.87              | 12.11    | 0.85     | 2.52*            | -0.43    | -0.62            |
| <b>Combined</b>   | 5            | 3.22              | 7.26     | 0.77     | 3.05*            | -0.02    | -0.04            |
|                   | 7            | 3.85              | 10.68    | 0.64     | 2.49*            | -0.65    | -1.26            |
|                   | 9            | 3.11              | 8.05     | 0.91     | 3.64*            | -0.03    | -0.07            |
|                   | 11           | 3.63              | 9.79     | 0.87     | 3.47*            | -0.13    | -0.26            |

Notes:

<sup>a</sup> In order to compare responses from participants, values are rescaled to 11 points (Dawes 2008).

\*Indicates a significant departure from a normal distribution at  $p < 0.05$ .

**Panel B: Analysis of Difference in Variance between Scale Design Characteristics**

| Statistical Test | Labels      | Scale Points |         |          |         |          |          |
|------------------|-------------|--------------|---------|----------|---------|----------|----------|
|                  | End vs. All | 5 vs. 7      | 5 vs. 9 | 5 vs. 11 | 7 vs. 9 | 7 vs. 11 | 9 vs. 11 |
| Levene's Test    | 0.01***     | 0.04**       | 0.61    | 0.14     | 0.13    | 0.64     | 0.32     |
| Bartlett's Test  | 0.01***     | 0.07*        | 0.62    | 0.14     | 0.17    | 0.67     | 0.33     |

Notes:

Difference of variance p-values are calculated using both Levene's Test and Bartlett's Test. Each value is derived from a series of individual ANOVAs, where the dependent variable is the test dependent variable presented in Panel A. The independent variable in the ANOVA is the Scale Comparison item listed above in Panel B (i.e. labels for all points vs. endpoints).

\*\*\*, \*\*, \* Indicate a significant difference in variance at  $p < 0.01$ ,  $p < 0.05$ , and  $p < 0.10$ , respectively (two-tailed tests).

Table 4  
Results of Experiment 2

**Panel A: Univariate Analysis of Dependent Variable FAVORABLE by Scale Points, Labels, and EARNINGS**

| <b>Scale Labels</b> | <b>Scale Points</b> | <b>EARNINGS</b> | <b>n</b> | <b>Mean</b> | <b>Std Dev</b> |
|---------------------|---------------------|-----------------|----------|-------------|----------------|
| Endpoints           | 5 points            | Negative        | 48       | 3.98        | 1.81           |
|                     |                     | Positive        | 50       | 5.81        | 1.51           |
|                     | 9 points            | Negative        | 48       | 3.98        | 1.81           |
|                     |                     | Positive        | 44       | 5.96        | 1.67           |
| All points          | 5 points            | Negative        | 50       | 4.19        | 1.76           |
|                     |                     | Positive        | 48       | 6.42        | 1.67           |
|                     | 9 points            | Negative        | 48       | 4.24        | 1.21           |
|                     |                     | Positive        | 47       | 6.08        | 1.32           |

Notes:

FAVORABLE is the dependent variable where participants assess how favorable a company will be as an investment. Values for FAVORABLE are rescaled to a 9 point scale according to Dawes et al. 2008.

EARNINGS is the case manipulation for whether a company had a positive earnings announcement (growth in revenue, net income, and cash) vs a negative earnings announcement (decline in revenue, net income, and cash).

Scale points and labels are manipulated within the assessment of the dependent variable.

**Panel B: Analysis of Regressions across Scale Points and Labels Conditions**

**Model: FAVORABLE =  $\beta$  EARNINGS + e**

| <b>Scale Labels</b> | <b>Scale Points</b> | <b><math>\beta</math> of EARNINGS</b> | <b>t-Value of <math>\beta</math></b> | <b>Regression R<sup>2</sup> (%)</b> |
|---------------------|---------------------|---------------------------------------|--------------------------------------|-------------------------------------|
| Endpoints           | 5 points            | 1.73                                  | 4.82*                                | 21.27                               |
|                     | 9 points            | 1.97                                  | 5.30*                                | 24.64                               |
| All points          | 5 points            | 2.45                                  | 6.88*                                | 35.49                               |
|                     | 9 points            | 1.86                                  | 6.99*                                | 36.24                               |

Notes:

In the analysis, all cells are randomly balanced with 44 response per cell (the lower limit of all 8 combinations) to prevent incidental bias t-Values or R<sup>2</sup>

\* Indicates significance at p < 0.001