

AN EVALUATION OF THE NUMBER OF RESPONSE OPTIONS FOR SCALES IN PSYCHOLOGY

Maria Borjas

APPROVED:

Lynne Steinberg, Ph.D.
College of Liberal Arts and Social Sciences
Thesis Director

Clayton Neighbors, Ph.D.
College of Liberal Arts and Social Sciences
Second Reader

Suzanne Kieffer, Ph.D.
Honors College
Honors Reader

Antonio D. Tillis, Ph.D.
Dean, College of Liberal Arts and Social Sciences

AN EVALUATION OF THE NUMBER OF RESPONSE OPTIONS FOR SCALES IN
PSYCHOLOGY

by
Maria Borjas

A Senior Honors Thesis
Submitted to the Department of Psychology,
College of Liberal Arts of Social Sciences

Chair of Committee: Lynne Steinberg

Committee Member: Clayton Neighbors

Committee Member: Suzanne Kieffer

University of Houston
April, 2020

ACKNOWLEDGEMENTS

First, I wish to express my gratitude to my thesis director, Dr. Lynne Steinberg, who has guided me through every step of the process. This project has been a great learning experience and I am grateful to have had Dr. Steinberg as my supervisor. I would also like to thank my committee members, Dr. Clayton Neighbors and Dr. Suzanne Kieffer for their time and dedication to this project. Lastly, I would like to thank everyone at the Social Influences and Health Behaviors Lab for their constant support.

ABSTRACT

Self-report scales are used widely in the field of psychology. These scales tend to widely differ on scale format for many reasons including consistency, time issues, and convenience. Previous studies have found that scale format has an effect on response variance, and reliability, among other psychometric properties. However, these findings have been mixed. The purpose of this study is to assess the effects of number of response options on response patterns and internal consistency. We used a 5- and 7-point scale of the Rosenberg Self-Esteem measure.

Undergraduate college students were administered this scale with either 5 or 7 response options. We found that frequency and response patterns did not differ between the 2 scales, but differences in response patterns per item were present. There were also mean differences between scales, although these effects were small. The number of response options did not affect reliability. Using descriptive statistics and t-tests, differences were not detected between responses to items presented with the 5- and 7-point response scales. Further research assessing more than one measure and comparing even, and odd numbered scales is needed to better understand the effects of number of response options on response patterns.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	i
ABSTRACT	ii
LIST OF TABLES	iv
LIST OF FIGURES	v
INTRODUCTION	1
Scale Development.....	1
Number of Response Options	3
More responses	5
Midpoint	7
No effect of number of response options found	7
Theoretical Background	8
Present Study.....	9
METHODS	10
Participants	10
Measure	10
Procedure.....	11
Method of Analysis	11
RESULTS	12
DISCUSSION	14
Limitations	15
Future Research.....	17
REFERENCES	19

LIST OF TABLES

Table 1	Summary of Relevant Articles.....	25
Table 2	Mean and Standard Deviations for 5-point and 7-point Scales	27
Table 3	t-Test Comparing 5-point and 7-point Scales	28
Table 4	Inter-Item Correlation Matrix for 5-point Scale	29
Table 5	Inter-Item Correlation Matrix for 7-point Scale	30

LIST OF FIGURES

Figure 1.1	Histograms for Item 1	31
Figure 1.2	Histograms for Item 2	31
Figure 1.3	Histograms for Item 3	31
Figure 1.4	Histograms for Item 4	32
Figure 1.5	Histograms for Item 5	32
Figure 1.6	Histograms for Item 6	32
Figure 1.7	Histograms for Item 7	33
Figure 1.8	Histograms for Item 8	33
Figure 1.9	Histograms for Item 9	33
Figure 1.10	Histograms for Item 10	34

Introduction

Scale development

Self-report rating scales are used widely across fields, including psychology, public health, and other social sciences (Drake, Hargraves, Lloyd, Gallagher, & Cleary, 2014). In psychology, rating scales are among the most widely used method of measurement and assessment (Preston & Colman, 2000). Self-report rating scales are used to measure inter- and intra-personal differences among subjects (Hilbert, Kuchenhoff, Sarubin, Nakagawa, & Buhner, 2016). Their utility, or the extent to which real-world decisions are dependent on them, is based on the quality of these psychological methods of assessment (Clark & Watson, 2019). Progress in psychological sciences and their intersection with other disciplines is dependent on measurement utility. A better understanding of the forms of measurement in psychology may assist in more accurately measuring psychological phenomena.

A notable issue in psychology has been the debate between scaling, measuring the extent of the presence of a particular construct, and categorizing, describing the presence or absence of a construct (DeVellis, 2017, p. 26). A shift in measurement has led to the inclusion of more rating scales in psychological research. DeVellis (2017) aptly defines scales as “measurement instruments that are collections of items combined into a composite score and intended to reveal levels of theoretical variables not readily observable by direct means” (p. 30). As scales are often the only form of measurement used in psychology studies, the psychometric properties of these are of much concern. Psychometrics is defined as a subspecialty in social science pertaining to the measurement of psychological and social phenomena. Scales differ greatly across studies, whether it be to remain consistent with the rest of the study, or to follow the format in past research. Scale format may differ by number of response options offered, response labels, and/or polarity of response categories across studies. These changes are sometimes made in attempt to

address difficulties in data collection that may arise. In other cases, scales may be adapted in response to differences such as educational level, culture, and social structure (Capik & Gozum, 2015). Scale format can influence measures of tendency like mean, variance, and other more complex measurements like correlations between variables, validity, and response bias, putting the comparability and generalizability of the data at stake (Cabooter, Weijers, Geuens, & Vermeir, 2016; Weijtjers, Cabooter, & Schillewaert, 2010). Other factors that should be considered when deciding the number of response options to use are response styles and the target group in the study. Hilbert, et al. (2016) found that variations in response tendencies differ across target groups. Response styles are also greatly affected by how an individual perceives a questionnaire, the scale format, and response categories (Arce-Ferrer, 2006; Baumgartner & Steenkamp, 2001; Paulhus, 1991). All in all, these studies support the idea that variation in response style is not only due to differences in individual characteristics, but to differences in response scale format as well. These findings emphasize the need for further understanding of the effects that response scale formats may have on response patterns.

Broadly, past research in the field found that different assessment formats affect response distribution across content (Hui & Triandis, 1989). That being said, research on the effect of differing formats, particularly number of response options in psychological scales, is scarce and existing research has found contradictory results (Chang, 1994). This remains a prevalent problem in research in psychology, and other disciplines heavily dependent on rating scales as a form of assessment. Duncan (1984) addressed this problem best as he stated, “Measurement is not only the assignment of numerals, etc. It is also the assignment of numerals in such a way as to correspond to different degrees of quality... or property of some object or event” (p.126).

Number of Response Options

Previous research in this field has attempted to identify the optimal number of response options that should be included in self-report scales (Chang, 1994). Researchers have focused on the effects of number of response options on the reliability, internal consistency, validity, and discriminating power, but there is no clear conclusion on the optimal number of response options (Preston & Colman, 2000). Interestingly, the debate on the optimal number of response options has been present for a significant period of time. One of the earliest reviews on the topic argued that reliability is optimized when seven response options are given, and reliability decreases as more items are added (Symonds, 1924). A major caveat of this paper being that empirical data were not used to arrive at this conclusion, as the author only used models. An early study on the topic found that the relationship between reliability and number of response options varies by score variability and the content being measured (Masters, 1974). More specifically, the authors found that while total score variance is lower when a smaller number of response options is provided, reliability increases when a greater number of response options is provided. Interestingly, the authors also found that this effect on reliability disappeared when opinion is divided on the content being assessed. Preston & Coleman (2000) compared responses for scales with the number of response categories ranging from 2 to 11, plus a 101-point scale, using the same sample for each group. They found that rating scales with a lower number of response categories had lower reliability, validity, and discriminating power, while these were higher for scales with more response categories, up to 7. Rating scales with more than 10 response options had lower test-retest reliability and internal consistency remained consistent between all scales. The authors went a step further by assessing respondent preferences, and they found that this was highest for scales with 10 response categories.

The concept of testing time and its relation to score reliability has also been a topic of discussion, as some argue that as average reaction time increases, so does score reliability (Matell & Jacoby, 1972). This is something to be considered, as there is a relation between number of items and testing time. In a similar vein, Capik & Gozum (2015) found that reducing the number of instrument items helped minimize participant fatigue without impacting results, using the Beck Depression Inventory (Furlanetto, Menlowicz, & Bueno, 2005) and Quality of Life scale (Jakobsson, 2007) as examples. Although this study will not focus on the effects of number of instrument items, it is important to note that a similar association between time, fatigue, and number of response options may also be present. Fatigue and testing time are important factors to consider in psychology research, especially when working with special populations or when a short amount of time is allotted for a study.

More recent studies on the topic have also found that the number of response options yields important implications for the psychometric properties of a questionnaire, where reliability (Cronbach's alpha) increases with the number of response alternatives of the scales, while validity coefficients remain steady (Hilbert, et al., 2016). However, it is important to note that this study compared a dichotomous scale, a 5-point Likert-type scale, and a 100-mm visual analog scale. This could be a problem as there is a lot of variance in number of response options between conditions. This is not reflective of the widely used self-report scales in psychology, and it calls for the need of further analyses on the effects of number of response options.

Simms, Zelazny, Williams, & Bernstein (2019) reported various psychometric implications related to number of response options. The purpose of their study was to identify the optimal number of response options, assess the difference between an odd and an even number of response options, and identify the benefit, if any, in using visual analog items over Likert-type

items. The authors presented participants with response scales ranging from 2 to 11 response options, plus a visual analog, a similar study design as Preston & Coleman (2000). However, Simms, et al. (2019) used a sample of 1,358 undergraduates and randomly assigned them into one of 6 groups to complete odd-even comparisons, where each respondent was presented with both an odd-numbered and an even-numbered scale, while Preston & Coleman (2000) had a group for each response scale. Simms, et al. (2019) concluded that the optimal number of response options is 6, as there was an increase in reliability up to 6 response options, and reliability leveled off after 6 response options. However, these results were not statistically significant, bringing into question the replicability of this study.

More responses

Many studies found that an increase in number of response options is related to an increase in reliability (Garner, 1960; Maydeu-Olivares, Kramp, Garcia-Forero, Gallardo-Pujol, & Coffman, 2009; Oaster, 1989; Symonds, 1924; Weng, 2004). For example, Oaster (1989) found a simultaneous increase in number of response options and test-retest reliability. Participants were presented with 3-point, 5-point, 7-point, or 9-point scales when completing the Texas Social Behavior Inventory. Others have found that providing 6 or 7 response categories optimizes reliability (Green & Rao, 1970; Symonds, 1924).

Similarly, a study implementing rating scales with 4, 6, 8, 10, 12, 16, and 20 response categories found that more information regarding respondent's beliefs was obtained when using the 20 categories as opposed to the other rating scales with fewer categories (Garner, 1960). Maydeu-Olivares, et al. (2009), incorporated classical test theory, item factor analysis, and item response theory methods to assess the effects of number of response categories in rating scales. For Study A they administered questionnaires 3 times with 2, 3, and 5 response categories, and

for Study B, they followed the same design, with the addition of a fourth administration to assess temporal consistency during which the tests were re-administered with either 2, 3, or 5 response categories (Maydeu-Olivares, et al., 2009). They found that increasing the number of response options increased reliability, but it did not have a similar effect on convergent or discriminant validity. This study differs from others in the field as they used a repeated-measures design (Maydeu-Olivares, et al., 2009). Weng (2004) showed similar results, as internal consistency increased from 3 or 9 response options. This study presented 12 groups with subscales of the Teacher Attitude Test, where the conditions varied by number of response options and the presence/absence of response labels. Weitjers, et al. (2010) found that adding more response categories is correlated with a decrease in extreme response style. They also show that scale format affected the mean and internal consistency.

Interestingly, a simulated study conducted by Green & Rao (1970) showed that by using 6 or 7 response options, information retrieval is maximized, but little extra information is gained by increasing the number of categories beyond seven. Chang (1994) found contradicting results regarding the adequate number of response options as providing more response options (6-point scale) increased systematic method variance while using less (4-point scale) showed higher test-retest reliability. Preston & Coleman (2000) found that although the number of response options did not affect the structure derived from the results, convergent validity increased simultaneously with the number of response options. That being said, the authors found that criterion-related validity was not impacted by the number of response options.

All in all, there is existing evidence that an increase in the number of response options used has a positive effect on retest reliability, different types of validity, and internal consistency. An increase in number of response options has also been found to help obtain more information

from the assessment and subdue extreme response style and its detriments. There is also evidence for the opposite effects on increasing number of response options, shedding light on the need for further research on the topic.

Midpoint

Matell & Jacoby (1972) studied whether or not an “uncertain” category should be included in response scales. Results from this study showed that including more intermediate categories around the midpoint reduces the size of the middle category. This allows for respondents to express their attitude even if it is slightly negative or positive (see also, Weems & Onwuegbuzie, 2001). Similarly, Weitjers, et al. (2010) found that the midpoint mitigates extreme response style, wherein extreme response style decreases when more response options are used. Simms, et al. (2019) found that midpoints are clearer and diminish social desirability. This wide array of results suggest that more research is needed to further assess the psychometric benefits provided by implementing midpoints, if any.

No effect of number of response options found

On the other side of the spectrum, Schutz & Rucker (1975) found no relation between number of response options and the response differences. This study analyzed data from rating scales with 2, 3, 6, and 7 response options, excluding 4 and 5 in an effort to compare shorter scales with the longer scales that are more commonly used. Similar to this study, Drake, et al. (2014) found no difference in composite score means and reliabilities in samples of university health center patients who receive 4 response categories, compared to those that received 6 response categories. The authors suggest implementing 4 response categories for more efficiency. Interestingly, in a study on testing time and scale format, Hui & Triandis (1989)

found that although the number of response options affected the usage of the “uncertain” response, it did not have an effect on the consistency, reliability, or validity of the results.

Although there have been many advances for research in this field, it is clear that an uncertainty remains about the number of response options that should be included. Previous studies attempting to tackle this issue tend to compare even and odd numbered scales to each other, disregarding differences between the two, and do not use a model-fitting approach (Chang, 1994). Others have focused on the measure of reliability only, with not much emphasis on validity and other analyses (Lozano, Garcia-Cueto & Muniz, 2008; Preston & Colman, 2000). Because of inconsistencies in findings regarding the effects of number of response options, readers might find Table 1, summarizing the methods and findings discussed in this literature review, useful.

Theoretical Background

Constructs we aim to measure in psychology are derived from psychological theory, and this theoretical background is crucial to better understand how problems in measurement are conceptualized. Various theories of measurement are used as the background for a significant portion of this field of research. The information theory states that the amount of information acquired from a respondent is dependent on scale range (Weitjers, et al., 2010). This suggests that a greater number of response options provides an optimal amount of information compared to scales with less response options (Green & Rao, 1970).

The motivational theory states that respondents aim to not only express their honest attitudes, but also meet the expectations set by the study (Krosnick, 1991). This suggests that extra response categories are necessary for participants to differentiate their responses between the two sides of the spectrum, agreement and disagreement to provide a response that is most

accurate and aligned with their attitudes (Krosnick, 1991). This motivational perspective may also have implications regarding the middle response category (Weitjers, et al., 2010).

Tourangeau & Rasinski (1988) established a four-stage process for answering attitude questions that is widely used in measurement research (Krosnick, 1999; Schwarz, 1999). The four steps are as follows: 1) interpretation of the question, 2) retrieve relevant beliefs and feelings, 3) use information retrieved to render an appropriate judgment, and 4) use judgment to select a response; map judgment onto response scale. Applying this model the question regarding the optimal number of response options helps raise questions like, “How does the number of response options alter the way a question is interpreted?” and “Does a respondent map their judgment differently based on how many response options are administered?”. This model aids in strengthening our understanding of the psychology behind response styles and how they differ as the number of response options is altered.

Present Study

This short review of studies that have focused on the effect of the number of response options show that findings are not consistent across studies. Overall, the question of how many responses should be included remains unresolved. Clark & Watson (2019) note that more research on the topic is needed for a broad range of constructs, samples, and type of format.

This research project aims to compare response patterns for a widely used scale in psychology research by altering the number of response options presented for each. We aim to do so by analyzing how the number of response options affects the frequency of responses in each category. We will also evaluate the frequency of middle responses, 3= neutral for the 5-point Likert-type scale and 4= neutral for the 7-point Likert-type scale, and how this differs between scales based on the number of response options. Being as this is an exploratory study,

there are no established hypotheses. The results of this study may have implications about the future use of response scales in psychology and other fields that heavily depend on the use of self-report scales.

Method

Participants

The sample was composed of undergraduate students from Portland State University and the University of Arkansas. 1,095 undergraduates (n= 404 men, n=686 women, and n=5 did not report gender) participated in the study. Data regarding other demographic characteristics were not assessed. The final sample was composed of 1,090 undergraduate students after eliminating observations with missing data.

Measure

The Rosenberg Self-Esteem (RSE) Scale is a 10-item self-report scale that measures self-esteem (Rosenberg, 1965). It includes items that assess both positive and negative feelings about the self. High scores on the RSE indicate high levels of self-esteem. The RSE includes items like “On the whole, I am satisfied with myself” and “I feel I do not have much to be proud of.” Half of the items were reverse coded, including Items 3, 5, 8, 9, and 10. Respondents are asked to rate how strongly they agree or disagree with each statement. This scale is widely used in research in psychology, and a wide array of other disciplines including kinesiology, and human resources (Koteles, Kollsete, & Kollsete, 2016; Tang, Tang, & Li, 2013). The RSE has been validated with different samples varying by age, place of origin, and language across cultures (Eklund, Backstrom, & Hansson, 2018; Gnamb, Scharl, & Schroeders, 2018; Ventura-Leon, Caycho-Rodriguez, Barboza-Palomino, & Salas, 2018). Self-esteem, as measured by the RSE has been found to be correlated with a multitude of variables, including dangerous mobile phone use, body

objectification, and subjective well-being, among others (Duy & Yildiz, 2019; Lannoy, et al., 2020; Veldhuis, Alleva, Vaate, Keijer, & Konjin, 2020).

Procedure

Undergraduate students were recruited from their lecture psychology classes. Participants from Portland State University answered the questionnaire during the last 10 minutes of class. Participants from the University of Arkansas answered the RSE as part of the screening procedures of a larger study. Participants were randomly assigned to one of two groups. One group was administered the RSE with a 5-point Likert-type scale (1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree), and the other with a 7-point Likert-type scale (1 = strongly disagree, 2 = disagree, 3 = somewhat disagree, 4 = neutral, 5 = somewhat agree, 6 = agree, 7 = strongly agree). The group administered the 5-point scale was composed of 542 undergraduates and the group administered the 7-point RSE scale was composed of 548 undergraduates. The questionnaires were administered by paper and pencil. Participants were randomly assigned to complete the RSE with either the 5- or 7-point response scale.

Method of Analysis

Various analyses were conducted to assess what effect, if any, the number of response options has on response patterns and/or response styles. We first assessed the frequency each response was selected for each item, for both scales. Secondly, conducted t-tests to identify differences between item mean scores between the 5-point and 7-point response scale groups. We also evaluated reliability to assess differences in internal consistency, if any, and correlations between items.

Results

Frequency analyses revealed interesting patterns regarding response variation per item. It should be noted that the RSE items are written such that respondents with moderate amounts of self-esteem endorse the upper end of the response scale; thus, there are very few responses in the lower response categories. Because of this pattern, most of the focus will be on the middle and higher response categories. Figure 1 shows histograms for the percentage of respondents selecting each categorical response; these were made separately for the 5- and 7-point response scales.

There were no identifiable differences between the 5-point and 7-point scale regarding frequency and response variation. Responses per item varied in a consistent pattern for items 1, 2, 3, and 5 for both types of response scales, as shown on Figures 1.1, 1.2, 1.3, and 1.5. Respondents selected the last 2 options with greater frequency, 4 and 5 for the 5-point scale, and 6 and 7 for the 7-point scale. Figures 1.4, 1.6, and 1.7 show response frequencies for Items 4, 6, and 7. These showed interesting patterns, as the second to last response option, 4 for the 5-point scale and 6 for the 7-point scale, were selected more than the endpoint. Item 4 states, "I am able to do things as well as most people," Item 6 states, "I take a positive attitude toward myself," and Item 7 states, "On the whole, I am satisfied with myself." Items 8, 9, and 10 also revealed different response patterns compared to the other items. These items showed responses almost evenly spread across the response scale. This can be seen in Figures 1.8, 1.9, and 1.10.

Item means and standard deviations for the 5-point scale and 7-point scale are presented in Table 2. These item means have been made somewhat comparable by dividing the observed item response mean for the 5-point data by 5 and dividing the observed item response mean for the 7-point data by 7. T-tests were done to evaluate the differences in item means. Table 3 lists

the t-tests, significance, and effect sizes (Cohen's *d*; Cohen, 1988) for the 10 RSE items. Significant differences were found for three items of the scale. There were no statistically significant differences between the 5-point scale and 7-point scale for items 1 through 7. Participants presented with a 5-point scale ($M = 0.70$, $SD = 0.25$), compared to those who were presented with a 7-point scale ($M = 0.65$, $SD = 0.28$) reported significantly higher scores $t(1088) = 3.08$, $p = .002$; $d = 0.27$, for Item 8, which reads, "I wish I could have more respect for myself." For Item 9, participants in the 5-point scale group ($M = 0.72$, $SD = 0.23$) also reported higher scores, $t(1088) = 3.73$, $p = .000$; $d = 0.24$, than the participants in the 7-point scale group ($M = 0.66$, $SD = 0.25$). Item 9 states, "All in all, I am inclined to feel that I am a failure." The same effect was found for Item 10, wherein those in the group with 5-point scales ($M = 0.81$, $SD = 0.23$) reported higher scores, $t(1088) = 2.11$, $p = .035$; $d = 0.24$, than those in the group with a 7-point scale ($M = 0.78$, $SD = 0.24$). This item states "At times I think I am no good at all." Although significant differences were detected, the size of the effect, as indicated by Cohen's *d*, is considered small.

Reliability, as assessed with Cronbach's alpha, for both the 5-point scale and 7-point scale was 0.88. This suggests that the change in number of response options did not have an effect on the internal consistency of each scale. Table 4 shows inter-item correlations for the 5-point RSE scale, and Table 5 shows inter-item correlations for the 7-point scale. Both correlation matrices present similar correlations among the items; for comparison, the range of correlations for the 5-point response scale was .28 to .70, and the range of correlations for the 7-point response scale was .25 to .77.

Discussion

This study aimed to fill the gaps in the literature regarding the number of response options and their effect on response patterns for self-report items. Number of response options did not have an observed effect on frequency and response patterns, but various differences across items were identified. In addition, item mean comparisons between the 5- and 7-point response scales showed significant differences for three RSE items; however, the effect sizes were small. Based on the small effect sizes, it is not certain that a change in the number of response options substantially influences item means. Similarly, due to the equal reliability coefficients for both scales we cannot make a recommendation regarding the optimal number of response options to use in self-report scales.

We did not find differences in frequency between the two response scales. However, there were some interesting patterns that arose between items for both scales. First, various items showed a consistent pattern, as most responses were either option “5” or “7” depending on the scale. This was not the case for all items. Items 4 and 6 showed that most respondents selected either option “4” or “6”, depending on the scale. These items focused on aspects of self-esteem related to comparing oneself to others and attitudes about oneself. These differences may be due to the specific aspects of self-esteem that these items aim to tap into. The last 3 items showed a different response pattern altogether, wherein the responses for each item were much more widespread, showing that respondents varied widely for these items. It is important to note that items 9 and 10, both include the phrase, “at times.” This may be the reason for the large variation for these items as much of the item is left to each respondents’ interpretation of “at times.” For some respondents this may be interpreted as at least one time, while others may interpret this as a more general daily experience. It is possible that this temporal wording may be responsible to the

difference in response patterns for these two items. Item 8, seems to touch more upon the idea of self-growth than self-esteem, using the phrase “I wish I could.” Although this does not tell us about differences across both types of response scales, the fact that these patterns were similar for both scale types might suggest that there is no substantial difference in responses to items presented with 5-point and 7-point scales. Related to this finding, we also found significant differences in the item response means for items 8, 9, and 10 of the RSE (Rosenberg, 1965). Average score per item was slightly higher for participants presented with a 5-point scale, compared with scores for participants presented with a 7-point scale.

Reliability analyses showed that the number of response options did not have a significant effect on the overall reliability of each scale. Both the 5-point and 7-point RSE scales showed equally high reliability. This suggests that adding more response options did not have an effect on the overall internal consistency of the RSE scale. This effect is aligned with results presented by Drake, et al. (2014) where there were no changes in reliability when comparing a 4-point scale and a 6-point scale. Our study differs from theirs as they used 2 even-numbered scales and we used two odd-numbered scales. However, various other studies have found the opposite results (Hilbert, et al., 2016; Masters, 1974; Preston & Coleman, 2010). It is important to note that these studies incorporated larger study designs, assessed both odd- and even-numbered scales, and used various measures.

Altogether these findings show that the number of response options does not have a substantial effect on response patterns, item means, correlations among items, and reliability. This is in line with various other studies in the field (Capik & Gozum, 2015; Drake, et al., 2014; Schutz & Rucker, 1975; Simms, et al., 2019). These findings should be taken with caution, however, due to a variety of study limitations.

Limitations

There are various limitations that should be addressed. First, this study did not use more than one measure. This could be a problem because one measure may not be enough to accurately assess the effects of the number of response options on response patterns. Although we used a widely known scale in psychology, in line with other research on this topic, it would have been more beneficial to use more than one measure. Being as we only used one measure in our study, it would be difficult to generalize our findings. It may be that the effects we found are only present for scales assessing self-esteem, or for 10-item scales.

Another limitation is that we only compared 2 odd-numbered scales. This is not in line with most of the existing research on this topic. Most studies on this topic employ a mix of odd-numbered and even-numbered scales, ranging from 2-101 response options, and some even add other types of measurement like visual analog scales (Garner, 1960; Hilbert, et al., 2016; Maydeu-Olivares, et al., 2009; Oaster, 1989; Preston & Coleman, 2000; Simms, et al., 2019; Weng, 2004). Being as the scales we used both include a midpoint, we cannot make any conclusions about the effects of adding a midpoint on response patterns, as we don't have anything to compare it to.

Other possible limitations include the lack of demographics information presented and the data collection methods. Although this study does not aim to identify differences across demographics, it would have been helpful to present information regarding common demographic factors like race/ethnicity, socioeconomic status, and age, and assess their possible confounding effects.

Future Research

Although there are various limitations, there are also some strengths that should be noted, including the large sample size, and the demographically diverse sample, based on gender.

Future research on the effects of number of response options on response patterns should aim to address the limitations and expand on the strengths of this study.

First, future studies should incorporate more than one measure. This may ensure that their results can be generalized across different measures. It may be useful to study well-known scales in the field to better assess the effects of number of response options on response patterns. This is in line with various existing studies like Simms, et al. (2019) that uses the Big Five Inventory and Capik & Gozum (2015) that uses the Beck Depression Inventory (Furlanetto, et al., 2005; John & Srivastava, 1999). By analyzing the effects of number of response options on response patterns for other well-known scales, results of these studies can be generalized to a wide range of topics. Scales that are more widely used in the field of psychology, are more likely to have better psychometric properties, leaving little to no room for inconsistencies regarding validity, and reliability due to the scale, making it easier to identify differences solely caused by the number of response options.

Secondly, future studies should assess scales with odd- and even-numbered responses. This will ensure that midpoint effects are analyzed. Recent studies on the use of midpoints in self-report scales have found interesting results. Nadler, Weston, & Voyles (2015) found respondents perceptions of the midpoint widely varied. The word “neither,” was perceived as “no opinion,” “don't care,” “unsure,” “neutral,” “equal/both,” and “neither.” Onwuegbuzie & Weems (2004) found demographic differences between those who selected the midpoint response category and those who did not, where males and younger respondents selected the

midpoint response. Social desirability response styles are also of much concern when using midpoints (Garland, 1991). Interestingly, Viswesvaran & Ones (1999) found that respondents who fake responses by strategically using the midpoint responses did so without affecting overall means. These findings suggest that comparing scales based on the absence or presence of the midpoint may have promising implications for self-report scales across all disciplines.

Using descriptive statistics, t-tests, and reliability analyses, differences were not detected between responses to items presented with the 5- and 7-point response scales. Future research using the methods of item response theory (IRT) may facilitate the identification of differences in the pattern of responses to questions presented with differing numbers of response options.

References

- Arce-Ferrer, A.J. (2006). An investigation into the factors influencing extreme-response style: Improving meaning of translated and culturally adaptive rating scales. *Educational and Psychological Measurement, 66*(3), 374-392. doi:10.1177/0013164405278575
- Baumgartner, H., & Steenkamp, J.-B. E.M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156. doi:10.1509/jmkr.38.2.143.18840
- Cabooter, E., Weitjers, B., Geuens, M., & Vermeir, I. (2016). Scale format effects on response option interpretation and use. *Journal of Business Research, 69*(7), 2574-2485. doi:10.1016/j.jbusres.2015.10.138
- Capik, C., & Gozum, S. (2014). Psychometric features of an assessment instrument with Likert and dichotomous response formats. *Public Health Nursing, 32*(1), 81-86. doi:10.1111/phn.12156
- Chang, L. (1994). A psychometric evaluation of 4-point and 6-point Likert-type scales in relation to reliability and validity. *Applied Psychological Measurement, 18*(3), 205-215. doi:0146-6216/94/030205-11\$1.80
- Clark, L.A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*. doi:10.1037/pas0000626
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.
- DeVellis, R. F. (2017). *Scale Development: Theory and Applications* (4th edition). SAGE Publications.

- Drake, K.M., Hargraves, J.L., Lloyd, S., Gallagher, P.M., & Cleary, P.D. (2014). The effect of response scale, administration mode, and format on responses to the CAPHS Clinician and Group Survey. *Health Services Research, 49*(4), 1387-1399. doi:10.1111/1475-6773.12160
- Duncan, O.D. (1984). Notes on social measurement: Historical and critical. New York: Russell Sage.
- Duy, B., & Yildiz, M.A. (2019). The mediating role of self-esteem in the relationship between optimism and subjective well-being. *Current Psychology, 38*, 1456-1463. doi:10.1007/s12144-017-9698-1
- Eklund, M., Backstrom, M., & Hansson, L. (2018). Psychometric evaluation of the Swedish version of Rosenberg's self-esteem scale. *Nordic Journal of Psychiatry, 72*(5), 318-324. doi:10.1080/08039488.2018.1457177
- Furlanetto, L. M., Mendlowicz, M. V., & Bueno, J. R. (2005). The validity of the Beck Depression Inventory-Short Form as a screening and diagnostic instrument for moderate and severe depression in medical inpatients. *Journal of Affective Disorders, 86*(1), 87–91. doi:10.1016/j.jad.2004.12.011
- Garland, R. (1991). The mid-point in a rating scale: Is it desirable? *Marketing Bulletin, 2*, 66-70.
- Garner, W. R. (1960). Rating scales, discriminability, and information transmission. *Psychological Review, 67*(6), 343–352. doi:10.1037/h0043047
- Gnambs, T., Scharl, A., & Schroeders, U. (2018). The structure of the Rosenberg Self-Esteem Scale: A cross-cultural meta-analysis. *Zeitschrift fur Psychologie, 226*(1), 14-29. doi:10.1027/2151-2604/a000317

- Green, P.E., & Rao, V.R. (1970). Rating scales and information recovery: How many scales and response categories to use? *Journal of Marketing*, 34(3), 33-39. doi:129.7.105.100
- Hilbert, S., Kuchenhoff, H., Sarubin, N., Nakagawa, T.T., & Buhner, M. (2016). The influence of the response format in a personality questionnaire: An analysis of a dichotomous, a Likert-type, and a visual analogue scale. *Testing, Psychometrics, Methodology in Applied Psychology*, 23(1), 3-24. doi:10.4473/TPM23
- Hui, C.H., & Triandis, H. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology*, 20(3), 296-309.
- Jakobsson, U. (2007). Using the 12-item Short Form health survey (SF-12) to measure quality of life among older people. *Aging Clinical and Experimental Research*, 19(6), 457-464. doi:10.1007/BF03324731
- John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (p. 102–138). Guilford Press.
- Koteles, F., Kollsete, M., & Kollsete, H. (2016). Psychological concomitants of crossfit training: Does more exercise really make your everyday psychological functioning better? *Kinesiology*, 48, 39-48. doi:159.9:796.015.15
- Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5, 213-236. doi:0888.4080/91/030213-24\$12.00
- Lannoy, S., Chatard, A., Selimbegovic, L., Tello, N., Linden, M.V., Heeren, A., & Billieux, J. (2019). Too good to be cautious: High implicit self-esteem predicts self-reported

dangerous mobile phone use. *Computers in Human Behavior*, 103, 208-213.

doi:/10.1016/j.chb.2019.09.018

Lozano, L.M., Garcia-Cueto, E., & Muniz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, 4(2), 73-79.

doi:10.1027/1614-2241.4.2.73

Masters, J.R. (1974). The relationship between number of response categories and reliability of Likert-type questionnaires. *Journal of Educational Measurement*, 11(1), 49-53.

doi:129.7.158.4

Matell, M.S., & Jacoby, J. (1972). Is there an optimal number of alternatives for Likert-scale items? *Journal of Applied Psychology*, 56(6), 506-509. doi:

Maydeu-Olivares, A., Kramp, U., Garcia-Forero, C., Gallardo-Pujol, D., & Coffman, D. (2009).

The effect of varying the number of response alternatives in rating scales: Experimental evidence from intra-individual effects. *Behavior Research Methods*, 41(2), 295-308.

doi:10.3758/BRM.41.2.295

Nadler, J. T., Weston, R., & Voyles, E.C. (2015) Stuck in the middle: The use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology*, 142(2), 71-89. doi: 10.1080/00221309.2014.994590

Oaster, T.R.F. (1989). Number of alternatives per choice point and stability of Likert-type scales.

Perceptual and Motor Skills, 68, 549-550. Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of social psychological attitudes*, Vol. 1. *Measures of personality and social psychological attitudes* (p. 17–59). Academic Press. doi:10.1016/B978-0-12-590241-0.50006-X

- Onwuegbuzie, A. J., & Weems, G. H. (2004). Response Categories on Rating Scales Characteristics of Item Respondents who Frequently Utilize Midpoint. *Research in the Schools*, 11(1), 50–59.
- Paulhus, D. L. (1991). “Measurement and control of response bias,” in *Measures of Personality and Social Psychological Attitudes*, eds J. P. Robinson, P. R. Shaver, and L. S. Wrightsman (San Diego, CA: Academic Press), 17–59.
- Preston, C.C., & Colman, A.M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, 104, 1-15. doi:0001-6918/00/
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Schutz, H.G., & Rucker, M.H. (1975). A comparison of variable configurations across scale lengths: An empirical study. *Educational Psychological Measurement*, 35, 319-324. doi:
- Schwarz, (1999). Self-reports: How the questions shape the answers. *American Psychologist*, 54(2), 93-105. doi:0003-066X/99/S? 00
- Simms, L.J., Zelazny, K., Williams, T.F., & Bernstein, L. (2019). Does the number of response option matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment*, 31(4), 557-566. doi:1040-3590/19/\$12.00
- Symonds, P.M. (1924). On the loss of reliability in ratings due to coarseness of the scale.
- Tang, C.S., Tang, T.L., Li, X. (2013). Chinese core self-evaluations and job performance: Entrepreneurs in private small and medium enterprises. *Journal of Chinese Human Resource Management*, 4(2), 151-170. doi:10.1108/JCHRM-01-2013-0001

- Tourangeau, R., & Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, 103(3), 299-314. doi:0033-2909/88/S00.75
- Veldhuis, J., Alleva, J.M., Vaate, A.J.D.B., Keijer, M., & Konjin, E.A. (2018). Me, my selfie, and I: The relations between selfie behaviors, body, image, self-objectification, and self-esteem in young women. *Psychology of Popular Media*, 9(1), 3-13. doi:2689-6567/20/\$12.00
- Ventura-Leon, J., Caycho-Rodriguez, T., Barboza-Palomino, M., & Salas, G. (2018). Evidencias Psicometricas de la escala de autoestima de Rosenberg en adolescents Limenos. *Revista Interamericana de Psicologia*, 52(1), 44-60.
- Viswesvaran, C. , & Ones, D. S. (1999). Meta-analyses of fakability estimates: Implications for personality measurement. *Educational and Psychological Measurement*, 59 , 197-210. doi:10.1177/00131649921969802
- Weems, G.H., & Onwuegbzie, A.J. (2001). The impact of midpoint responses and reverse coding on survey data. *Measurement and Evaluation in Counseling and Development*, 34, 166-176.
- Weitjers, B., Cabooter, E., & Schillewaert, N. (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. doi:10.1016/j.ijresmar.2010.02.004
- Weng, L., (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test-retest reliability. *Educational and Psychological Measurement*, 64(6), 956-972. doi:10.1177/0013164404268674

Table 1. Summary of Relevant Articles

Author	Number of Response Options	Labels	Findings
Capik & Gozum (2015)	Dichotomous and 5	2: Yes or No 5: Strongly Disagree to Strongly Agree	No difference.
Chang (1994)	4-and 6	4: Disagree to Agree 6: Strongly Disagree to Strongly Agree	No change in validity, but the 4-point scale had higher reliability.
Drake, Hargraves, Lloyd, Gallagher, & Cleary (2014)	4 and 6	Never to Always	No difference.
Garner (1960)	4, 6, 8, 10, 16, and 20	Rating the legibility of handwriting samples	More information is obtained when more options are given.
Hilbert, Kuchenhoff, Sarubin, Nakawaga, & Buhner (2016)	Dichotomous, 5, and 100-mm visual analogue scale	2: Yes or No 5: Completely Incorrect to Completely Correct	Reliability increased with the increase in number of response option. Validity remained the same.
Masters (1974)	2, 3, 4, 5, 6, and 7	2 & 3: Agree to Disagree 4 & 5: Strongly Agree to Strongly Disagree 6 & 7: Very Strongly Agree to Very Strongly Disagree	Reliability increased with the increase in number of response options, dependent on low score variability.
Matell & Jacoby (1972)	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, and 18	Not specified	Score variability did not change. Testing time and use of the midpoint category increased with the increase in number of response options.
Maydeu-Olivarez, Kramp, Garcia-Forero, Gallardo-Pujol & Coffman (2009)	2, 3, and 5	2: Yes or No 3: False to True 5: Very False to Very True	Reliability increased, goodness of fit decreased, and validity stayed the same with the increase in number of response options.
Oaster (1989)	3, 5, 7, and 9	Marking a number on a scale	Reliability increased with the increase in

			number of response options, up to 7.
Preston & Coleman (2000)	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 101	Very Poor to Very Good	Reliability, validity, and discriminating power were low for scales with 2-4 response options. Reliability increased with the increase in number of response options up to 7.
Schutz & Rucker (1975)	2, 3, 6, and 7	Appropriate to Inappropriate	No difference.
Simms, Zelazny, Williams, & Bernstein (2019)	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and visual analog scale	2 & 3: Disagree to Agree 4-7: Strongly Disagree to Strongly Agree 8-11: Very Strongly Disagree to Very Strongly Agree	No difference.
Weitjers, Cabooter, & Schillewaert (2010)	4, 5, 6, and 7	Strongly Disagree to Strongly Agree	The midpoint has an effect on response patterns.
Weng (2004)	3, 4, 5, 6, 7, 8, and 9	Does not describe me at all to Describes me completely	Test-retest reliability increased with the increase in number of response options.

Table 2. Mean and Standard Deviation for 5-point and 7-point Scales

Item		Mean	SD	Mean	SD
		5-point scale		7-point scale	
1	I feel I am a person of worth, at least on an equal basis with others.	.92	.12	.91	.14
2	I feel that I have a number of good qualities.	.90	.13	.90	.12
3	All in all, I am inclined to think I am a failure.	.87	.16	.85	.18
4	I am able to do things as well as most people.	.84	.15	.84	.15
5	I feel that I do not have much to be proud of.	.87	.18	.85	.21
6	I take a positive attitude toward myself.	.80	.17	.81	.17
7	On the whole, I am satisfied with myself.	.79	.17	.79	.18
8	I wish I could have more respect for myself.	.70	.25	.65	.28
9	I certainly feel useless at times.	.72	.23	.66	.25
10	At times I think I am no good at all.	.81	.23	.78	.24

Note: n=542 for the 5-point scale and n= 548 for the 7-point scale

Table 3. *t*-Test Comparing 5-point and 7-point Scales

Item		t	df	<i>p</i>	Cohen's <i>d</i>
1	I feel I am a person of worth, at least on an equal basis with others.	0.68	1088	.495	
2	I feel that I have a number of good qualities.	-0.50	1088	.614	
3	All in all, I am inclined to think I am a failure.	1.59	1088	.113	
4	I am able to do things as well as most people.	0.57	1088	.568	
5	I feel that I do not have much to be proud of.	1.53	1088	.127	
6	I take a positive attitude toward myself.	-1.03	1088	.305	
7	On the whole, I am satisfied with myself.	-0.35	1088	.727	
8	I wish I could have more respect for myself.	3.08	1088	.002**	.266
9	I certainly feel useless at times.	3.73	1088	.000**	.240
10	At times I think I am no good at all.	2.11	1088	.035*	.235

* significant at $p < 0.05$; ** significant at $p < 0.005$.

Figure 1. Histograms for Each Item

Figure 1.1 Histogram for Item 1

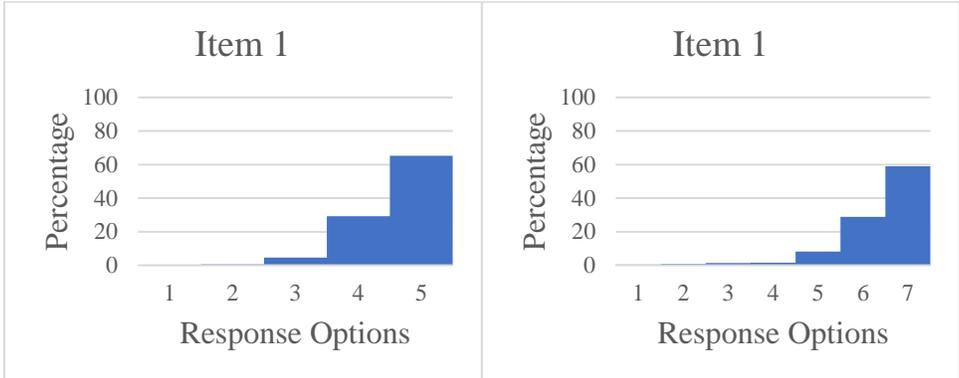


Figure 1.2 Histogram for Item 2

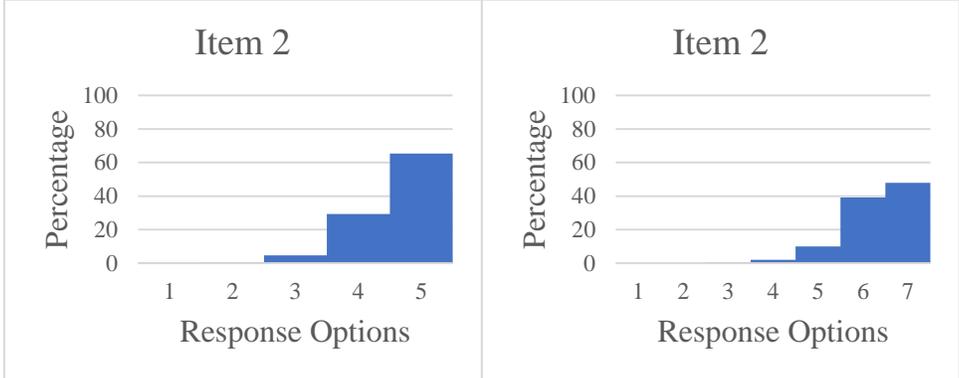


Figure 1.3 Histogram for Item 3

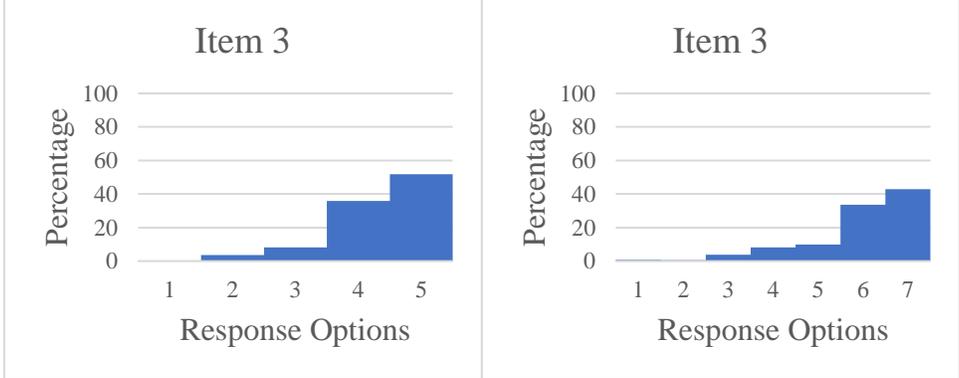


Figure 1.4 Histogram for Item 4

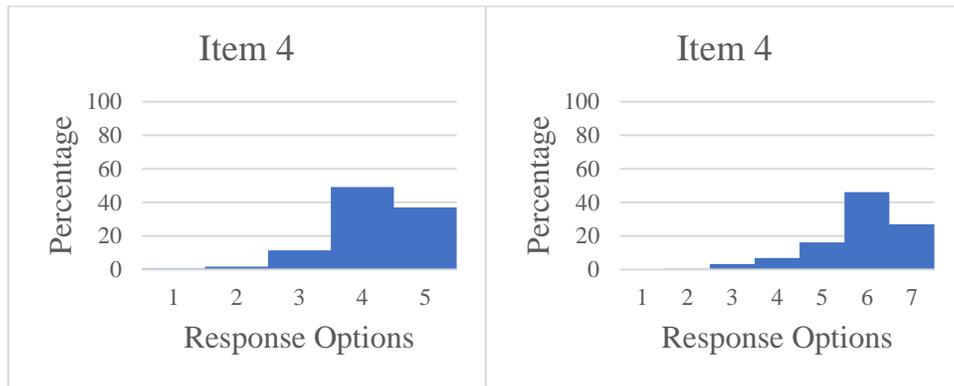


Figure 1.5 Histogram for Item 5

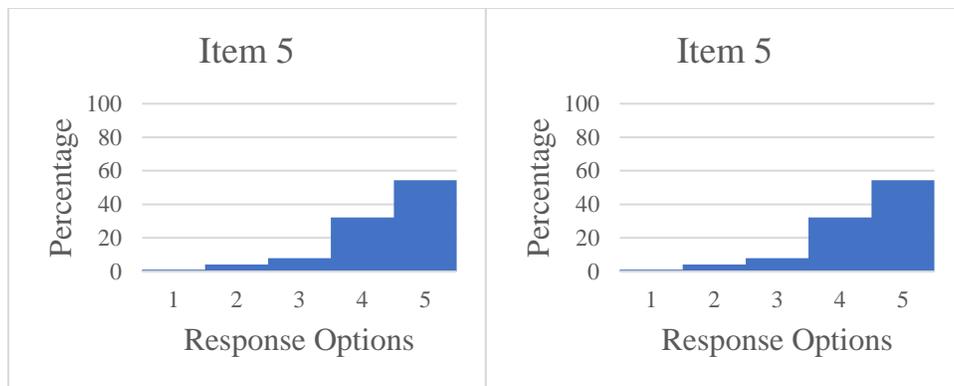


Figure 1.6 Histogram for Item 6

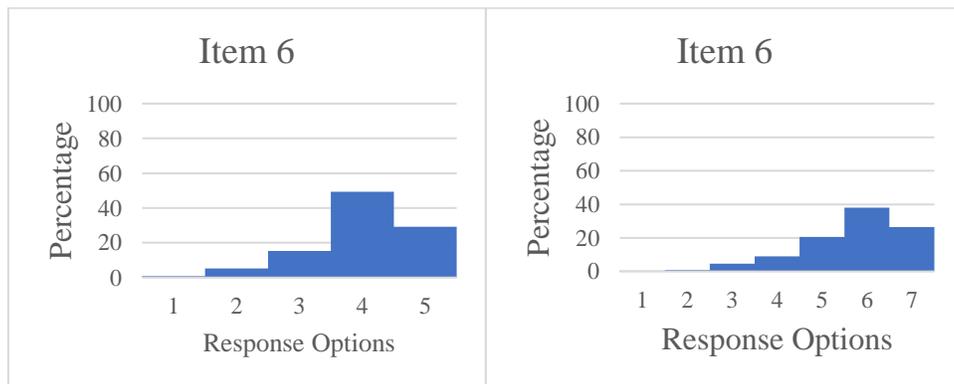


Figure 1.7 Histogram for Item 7

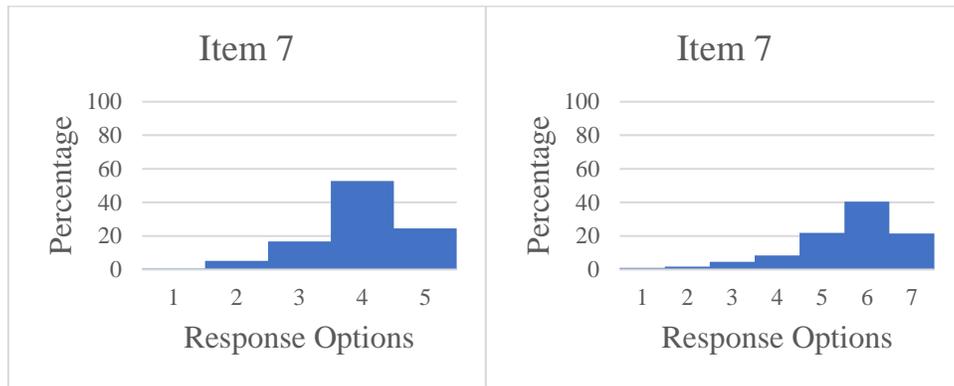


Figure 1.8 Histogram for Item 8

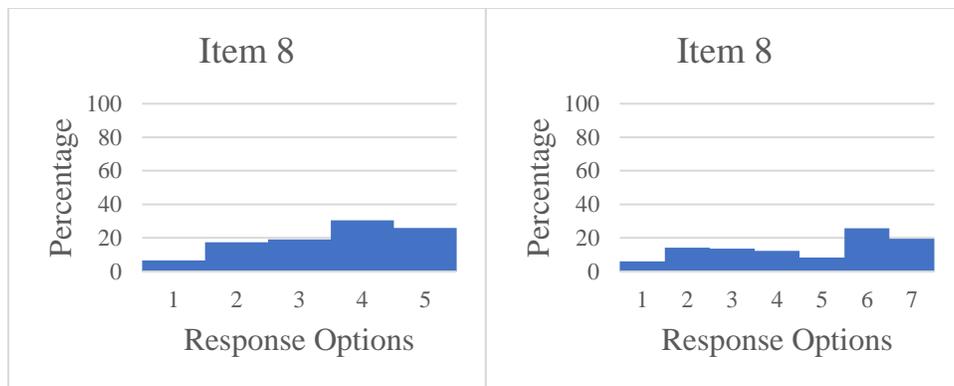


Figure 1.9 Histogram for Item 9

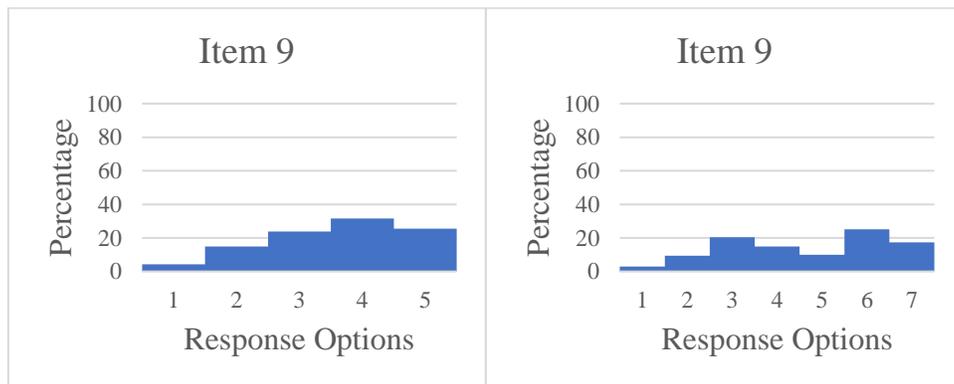


Figure 1.10 Histogram for Item 10

