AN EXAMINATION OF THE RELATIONSHIPS BETWEEN RESPONSE SCALE LENGTH, LABEL FORMAT, RELIABILITY, AND VALIDITY

by

TYLER HAMBY

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
May 2015

Acknowledgements

I would first like to thank my original mentor Dr. Scielzo for teaching me about psychometrics and introducing me to the field that I eventually decided to specialize in. I would also like to thank Dr. Levine, who has been my mentor over the past year. Being in his lab has allowed me to be very productive in research this year and to learn how to be an active member of a research lab.

I also want to thank the members of my committee. I thank Dr. Ickes for the tremendous help that he offered in our research collaboration over the summer of 2014 and for teaching me how to produce publishable articles. Dr. Dougall and Dr. Han have each strengthened my grasp of statistics with their rigorous classes. I appreciate Dr. Gagne for being a helpful committee member for my Major Area Paper and dissertation and for thinking of me for a potential research collaboration.

Lastly, I appreciate my friends and family for their support. I want to thank Wyn Taylor for opening up several research opportunities throughout the year, and I thank my entire lab for their reliable help. Lastly, I want to acknowledge and thank my Mom, Norm, Garret, Heather, Kayla, Robert, and Nuri for their support over my many years of education.

January 29, 2015

Abstract

AN EXAMINATION OF THE RELATIONSHIPS BETWEEN
RESPONSE SCALE LENGTH, LABEL FORMAT,
RELIABILITY AND VALIDITY

Tyler Hamby, PhD

The University of Texas at Arlington, 2015

Supervising Professor: Daniel S. Levine

This research examined the impact that two qualities of scales—the number of response options and the response scale label format—have on reliability and validity. Based on a meta-analytic pilot study, I expected that these two scale qualities would interact to predict reliability, such that the association between response scale length and reliability would be stronger for fully labeled than for endpoint labeled scales. This study also examined three quantitative variables that have previously been hypothesized to moderate the association between response scale length and reliability: score variability, item homogeneity, and skewness. I randomly assigned 893 participants to one of six scale format conditions to fill out six questionnaires; the reliabilities of these measures were examined. Then, the subjects took two more questionnaires, and the scores of these measures were correlated with the scores from the first six questionnaires for construct validity coefficients. Response scale length and label format did interact to predict reliability but not as expected. However, this outcome may have been due to the characteristics of the chosen sample. I found that college educated respondents had higher reliabilities at seven-point scales, as compared to five-point scales, but this pattern was not seen in the less educated group. The three quantitative variables also

moderated the response scale length and reliability relationship, though not all in the manner anticipated. Finally, the number of response options did influence validity, but the only generalizable conclusion was that fully labeled scales outperformed endpoint labeled scales at seven response options.

Table of Contents

List of Illustrations

List of Tables

Chapter 1

INTRODUCTION

Self-report surveys are among the most important tools in the psychological sciences; yet, to date, we have little understanding of the extent to which various scale modifications might have an impact on the efficacy of these measures. There is, for example, a dearth of research on whether it is beneficial to label all of the categories of a response scale or just the endpoints. Both types of scales are commonly used, and a particular researcher's preference for one over the other is likely not scientifically based. Although much research has explored the relationship between the number of response categories and reliability, the findings tend to be contradictory. Thus, it is likely that researchers still tend to select scale formats based on intuition, familiarity, or preference, instead of the findings from empirical research.

The present research provides an extensive analysis of how these scale modifications affect the psychometric properties of scales. In particular, this study explores an interaction between the number of response categories and the labeling format of those categories in predicting internal consistency reliability and construct validity. I also compare the predictions of various theories about how attributes of scales (score variability, item homogeneity, and skewness) should affect the relationships between the number of response categories and internal consistency reliability. Before reviewing the relevant research, it is necessary to clearly define reliability and validity.

1.1 Reliability and Validity

For a scale to have practical utility, it must demonstrate adequate psychometric properties, including reliability and validity. A succinct and insightful definition of these two constructs was given by Campbell and Fiske (1959, p. 83): "Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods." These two constructs are interrelated

but not interchangeable: It is said that reliability is necessary, but not sufficient, for validity (Nunnally & Bernstein, 1994). Reliability is ultimately a precondition for whether or not scores on a scale can correlate with other meaningful criteria.

1.1.1 Reliability

Aside from the very theoretical definition given above, reliability has some more practical and statistical definitions. In contemporary research, it can refer to temporal stability, alternate forms reliability, or internal consistency (Nunnally & Bernstein, 1994). Temporal stability, or test-retest reliability, refers to the degree that a person will have similar scores when taking the same test at two different points in time. Similarly, alternate forms reliability measures the correspondence between respondents' scores on two tests or questionnaires which purport to measure the same thing. Test-retest and alternate forms reliability are most often measured with a correlation coefficient. Lastly, internal consistency measures how well the test items interrelate. It is the type of reliability that is most important for the present study because it measures the amount of error that is associated with a scale.

Internal consistency is most often measured using coefficient alpha (Cronbach's alpha; Cronbach, 1951). Alpha is essentially the ratio of true variance of a scale, estimated by the average of all the covariances between the items, to the true variance plus error variance, estimated by the average of all the covariances between the items *and* the item variances. Internal consistency is a bit of a misnomer because it reflects both inter-item correlation and the number of items (Nunnally & Bernstein, 1994). It therefore follows that increasing the number of items in a scale should increase alpha, and as the number of items increases, lower levels of inter-item correlations are required to achieve a given level of reliability (Cortina, 1993). In this study, I am only interested in the inter-item correlation factor, so I experimentally control for the influence of the number of items.

1.1.2 Validity

A questionnaire is useless unless it has been shown to have validity. Traditionally, there are three types of validity: content, predictive, and construct (Nunnally & Bernstein, 1994). Content validity depends primarily on how the scale was developed, so this type of validity is not relevant to my purposes. Predictive validity and construct validity are typically assessed with correlation coefficients. Predictive validity involves demonstrating that the instrument relates to some external behavioral criterion. Construct validity is the extent to which an instrument measures what it purports to measure. A self-report scale should correlate with other measures of the same construct and, to a lesser degree, with other related constructs; correlations between unrelated measures should be much smaller than either of these two correlations (Campbell & Fiske, 1959). As I will discuss later, this issue is of particular importance in the proposed study because it is possible that certain scale modifications may artificially inflate reliabilities and correlations by exaggerating response styles. I will examine construct validity in the proposed study, and I will control for any effect of inflated reliabilities by examining the disattenutated correlations.

1.2 Number of Response Categories

When subjects are answering items on a Likert-type scale, they must choose a number from some predefined range of values. In modern research, it is most common for the response alternatives to range from one to five, but other ranges are frequently seen. For scales with more response options, the subject can better qualify the extent of his or her agreement or disagreement with an item, but the directionality of the response should be independent of scale length (Weijters, Cabooter, & Schillewaert, 2010). Ideally, the scale should have enough response categories for the respondent to accurately express him or herself, but it should not have so many response categories that random error is introduced, due to respondents not using all of the response levels (Alwin, 1992; Cox, III, 1980). The actual number of response categories needed to

achieve this ideal probably depends on how refined people's conception of the measured construct is (Krosnick & Presser, 2010). So, perhaps if the items are too vague or abstract, subjects might have difficulty choosing between the response categories for a lengthier scale.

A number of cognitive factors have been proposed to be related to response scale length. Miller (1956) theorized that people are cognitively limited to seven plus or minus two distinctions in terms of memory and judgment. This theory has been used to support using five to nine response categories (Cox III, 1980). If cognitive limitations do influence the optimal response scale length, then the appropriate number of response categories could depend on the cognitive ability of the target population.

Another relevant issue is the distinction between *optimizing* and *satisficing*. These terms have been used to describe how people respond to survey items (e.g., Schwarz & Strack 1985; Tourangeau & Rasinski, 1988). To answer questions optimally, or to optimize, people presumably go through these steps: They interpret the question, deduce its intent, search their memories for relevant information, integrate the information into a single judgment, and then select the most appropriate category for that judgment. Ideally, subjects optimize and thoroughly complete each step. In reality, respondents often satisfice, or settle for a "good enough" answer, by skipping or expending minimal effort on these cognitively demanding steps. The likelihood of satisficing is thought to increase with (1) increased task difficulty, (2) decreased respondents' abilities, and (3) decreased respondents' motivation (Krosnick, 1991). Krosnick and Presser (2010) suggest that subjects might generally have more difficulty responding to lengthier response scales, which could encourage satisficing and result in lower reliability or validity.

A large number of studies have examined the association between the number of response categories and reliability. Studies employing Monte Carlo simulations have shown that using more response categories improves internal consistency reliability

(Bandalos & Enders, 1996; Jenkins & Taber, 1977; Lissitz & Green, 1975; Lozano, García-Cueto, & Muñiz, 2008) and test-retest reliability (Jenkins & Taber, 1977; Lissitz & Green, 1975). One important finding from these studies is that the response scale length apparently has the largest effect on reliability at lower average inter-item correlations or covariances (Lissitz & Green, 1975; Lozano, García-Cueto, & Muñiz, 2008). This makes sense mathematically: It can be shown that any given increase in the average inter-item correlation or covariance will increase alpha to a greater extent at lower, compared to higher, starting levels of average inter-item correlation or covariance. The mathematical details are given in Appendix A, and Figure 1.1 demonstrates the point graphically. These simulation studies are useful, but they do not account for the psychological factors discussed above.

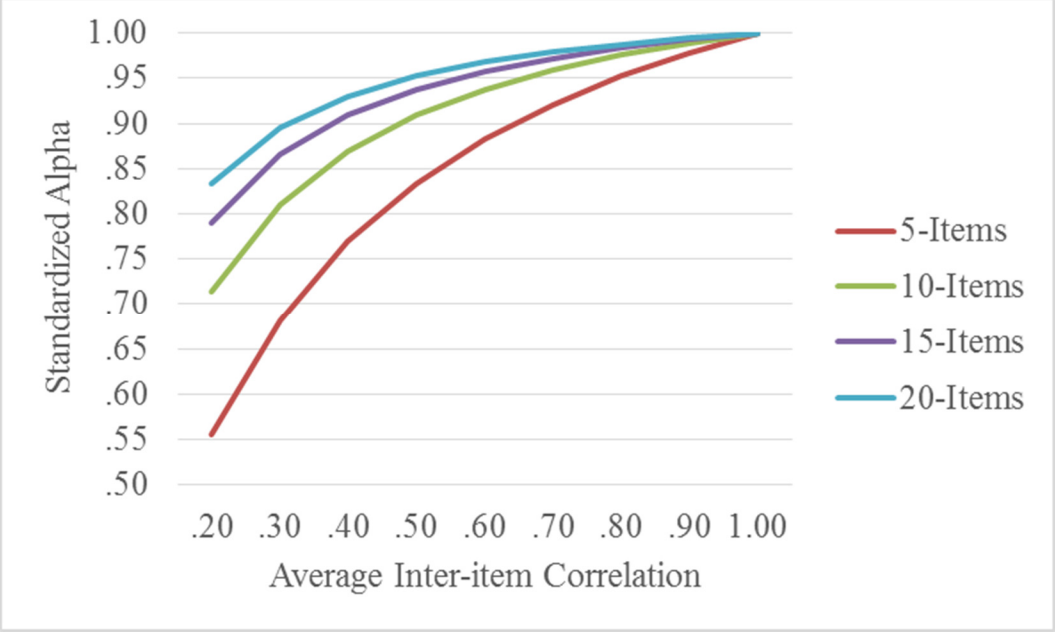

Figure 1.1. The relationships between the average inter-item correlation, the number of items, and standardized alpha.

Although experimental studies have produced more mixed findings, these too seem to favor scales with more response alternatives more often than not. The results of these studies have shown that increasing the number of response alternatives increases

internal consistency (Alwin, 1997; Oaster, 1989; Preston & Colman, 2000; Weng, 2004), test-retest reliability (Preston & Colman, 2000; Weng, 2004), and alternate forms reliability (Oaster, 1989). Meta-analytic studies on coefficient alpha have also shown a positive relationship between the number of response alternatives and internal consistency reliability (Churchill & Peters, 1984; Peterson, 1994).

In spite of the growing body of research suggesting that using more response alternatives improves reliability, some experimental studies have found no relationship (e.g., Bendig, 1953, 1954; Mattell & Jacoby, 1971). Moreover, those studies which *have* found a significant relationship have provided different recommendations for the optimal number of response alternatives. Authors have recommended two or three (Mattell & Jacoby, 1971), five (Jenkins & Taber, 1977; Lissitz & Green, 1975), seven (Finn, 1972; Oaster, 1989) or even more categories (Alwin, 1997; Preston & Colman, 2000). The variability in these recommendations raises the question: What variables influence the relationship between the number of response options and reliability?

Researchers have proposed a number of theories to answer this question. Masters (1974) posited that, for questionnaires with relatively low total score variability at fewer numbers of response categories, reliability benefits from using more categories. This prediction was tested and confirmed with only two scales. Komorita and Graham (1965) suggested that the homogeneity of the item distributions determines whether increasing the response scale length increases reliability: increasing scale length improves reliability for heterogeneous but not homogeneous questionnaires. In this study and in another study (Weng, 2004), item homogeneity was defined as the size of the factor loadings of the items on the first factor. Both studies empirically confirmed the theory; but in both studies the researchers administered only one relatively heterogeneous scale and one relatively homogeneous measure. These studies are obviously limited by the number of constructs measured.

Previous Monte Carlo simulation studies have demonstrated that increasing the number of response alternatives attenuates the deleterious effects of skewness on reliability (Bandolos & Enders, 1996; Bernstein & Teng, 1989). No experimental studies have tested whether skewness moderates the impact of scale length on reliability. Lastly, the present author has obtained meta-analytic results which suggest that the number of alternatives interacts with the labeling format for those alternatives to predict reliability. This study will be described in detail below. Each of these ideas could plausibly account for the mixed findings in the literature, but none of them have been sufficiently tested. Although they are not mutually exclusive, it is uncertain whether any or all of these propositions are actually valid, primarily because of methodological limitations of the supporting studies. Moreover, to date the various theories have not been systematically tested against each other in a single study.

To the extent that scale length affects reliability, increasing the number of response alternatives should improve validity correlations by increasing reliability because imperfect internal consistency reliability attenuates the strength of validity correlations (Nunnally & Bernstein, 1994). Note, however, that Cronbach (1950) argued that any gains in reliability obtained by adding response options are spurious consequences of the idiosyncratic response patterns of subjects. For example, with more response options, respondents with tendencies to give extreme responses will improve reliability more than they would with shorter response scales, but the resulting data will not necessarily be more meaningful. By this reasoning, lengthening the response scale may improve reliability at the expense of validity. Chang (1994) tested these two contrasting predictions by performing confirmatory factor analysis on a multitrait-multimethod matrix for four and six category scales for three different measures. He found that the average validity correlation between scales was higher for six-point scales, but that criterion validity was not improved. Thus, the methodological variance associated with using a longer scale was assumed to inflate the correlations between

scales. He suggested that test-retest reliability and validity correlations between two scales should ideally be tested when the two scales have differing numbers of response categories.

Unfortunately, relatively little research has examined the association between scale length and validity. Matell and Jacoby (1971), using experimental data, concluded that two response options (e.g., true or false) were sufficient to achieve acceptable validity. Regardless, most research has demonstrated that increasing the number of response categories improves validity. Alwin (1997) found that 11-point scales had higher validity coefficients than did seven-point scales in 14 out of 17 different surveys. Similarly, another study (Loken, Pirie, Virnig, Hinkle, & Salmon, 1987) found that 11-point scales had higher criterion validity than either three or four-point scales. Preston and Colman (2000) compared scales with two to 11 response categories, and they found that scales with five or more response alternatives had the highest criterion validities.

1.3 Response Alternative Label Format (All or Endpoint Only)

Some Likert-type rating scales give verbal descriptions for each number on a rating scale. For example, a five-point scale may be labeled 1-*Strongly Disagree*, 2-*Disagree*, 3-*Neither Agree Nor Disagree*, 4-*Agree*, and 5-*Strongly Agree*. On the other hand, other Likert-type scales provide descriptions for the two endpoints only. In the previous example, it would be 1-*Strongly Disagree*, 2, 3, 4, and 5-*Strongly Agree*. This difference raises the question of which scale format results in superior psychometric properties.

There are theoretical reasons why one scale format might result in higher reliabilities than the other. It has been said that response alternatives without labels are more ambiguous in meaning (Alwin & Krosnick, 1991); labeling each category gives the respondent some indication as to when it is appropriate to choose that particular response option, but it is unclear whether this labeling actually achieves the intended effect. Krosnick and Presser (2010) argue that response scales with only verbally

labeled endpoints are more difficult to respond to than completely verbally labeled scales because respondents must generate an appropriate label for each numbered category before selecting the appropriate category. If this is true, then the increased task difficulty could encourage satisficing behavior for people responding to scales with only the endpoints labeled.

Conversely, it has been argued that respondents tend to view numerical response alternatives as being equally spaced (Schaeffer & Presser, 2003). Because there are only numbers between the two poles, the subject should view the scale as being interval in nature, and the various statistical benefits of having a normal, interval scale may in turn come with it. On the other hand, it is debatable whether a Likert-type scale with all points labeled is, in fact, ordinal or interval (Goddard & Villanova, 2006). Some evidence indicates that people do not always perceive the commonly used response category labels as being spaced in the manner that an interval scale would require (Bass, Cascio, & O'Conner, 1974; Dobson & Mothersill, 1979; Spector, 1976), and the labels and positions of those labels can affect how people interpret the response choices (Klockars & Yamagishi, 1988). If fully labeled scales are ordinal and endpoint only scales are interval, then endpoint only scales would have some advantage in terms of reliability and validity. Given the conflicting arguments that have been put forth, empirical data are necessary to determine which scale format results in better psychometric properties.

Little research has directly addressed the impact that this choice of labeling may have on reliability or validity, and the research that has been done so far has produced mixed results. Although at least one study found that labeling only endpoints produced higher reliabilities for several surveys (Andrews, 1984), most research has either found that labeling each alternative is preferable or that there is no difference. Three meta-analyses collected studies with great variability in the number of response alternatives, and they found no relationship between labeling format and reliability (Churchill & Peters,

1984; Peterson, 1994; Peterson & Kim, 2013). On the other hand, Alwin and Krosnick (1991) used an archival method, with 13 seven-point scales, and they found that the scales which labeled each response option had higher reliabilities. Also, Peters and McCormick (1966) compared reliabilities for the two label formats on seven-point rating scales, and they found that labeling each category produced higher reliabilities.

Another study (Bendig, 1953) examined the relationship between the number of response alternatives, the labeling format, and reliability. In this study, there was modest evidence in favor of labeling more categories, but there was no significant interaction between scale length and scale format. More recently, Weng (2004) examined both predictor variables. There was no overall difference in reliability between scales that labeled only endpoints and scales that labeled each option. However, Weng concluded that for lengthier scales it was better to label each category in terms of test-retest reliability but not internal consistency reliability, though there was slight evidence of an effect for internal consistency as well. Note that most of the reviewed studies that found an overall effect for label format used scales with seven response categories, and these studies indicated that scales with each category labeled had higher reliabilities. This finding aligns with my prediction that scale length and label format interact to predict reliability. I now present evidence for this interaction effect.

1.4 Pilot Study

For this project, I used a meta-analytic approach to investigate the associations between two qualities of self-report questionnaires—the number of response alternatives and the labeling format of those alternatives—and reliability as measured by coefficient alpha. Importantly, three prior meta-analyses (Churchill & Peters, 1984; Peterson, 1994; Peterson & Kim, 2013) have already explored the impact that these scale modifications have on reliability. These studies generally found that the number of response options positively predicted reliability, but they found no relationship between label format and reliability.

For this study, I had several goals.

- I wanted to replicate the general finding that the reliability of a scale is positively related to the number of response categories.
- I also wanted to determine whether there was a substantial improvement in reliability when adding response categories beyond five.
- I wanted to test whether scales with each category labeled (ALL) differed from scales with only the endpoints labeled (END) in terms of reliability.
- Lastly, I sought to find a satisfactory explanation as to why the prior meta-analyses had found no difference in reliability between ALL and END scales.

To meet these goals, I selected studies for my meta-analysis by drawing from three meta-analyses that have already been done (Chiaburu, Oh, Berry, Li, & Gardner, 2011; Judge & Bono, 2001; Trapman, Hell, Hirn, & Schular, 2007). These three meta-analyses together used studies that measured the Big Five variables, job satisfaction, and self-esteem. For the pilot study data analyses, I generally followed the method described by Rodriguez and Maeda (2006). This method transforms the original alphas (to *T*) to make them more normally distributed, weights them by their inverse variances to compute a weighted mean value, and then transforms that value back to alpha for the mean effect size. I estimated the random variance component using the restricted maximum likelihood estimator (REML) method.

For the moderator analyses, I first applied the Spearman-Brown prophecy formula to each alpha coefficient to control for the number of items in the scale. Specifically, I estimated each alpha at the sample-size-weighted mean number of items:

$$\alpha_{SBjk} = \left(\frac{\overline{\#Items} / \#Items_j * \alpha_{jk}}{1 + \left(\overline{\#Items} / \#Items_j - 1 \right) \alpha_{jk}} \right),$$

where α_{jk} is the initial alpha for study *k* within variable *j*; α_{SBjk} is the transformed alpha; $\#Items_{jk}$ is the number of items used to compute the particular alpha; and $\overline{\#Items_j}$ is the

sample-size-weighted mean number of items for variable *j*. After I estimated the alpha coefficient in this way, I generally followed the procedure described by Rodriguez and Maeda (2006). This method lead me to first transform the alphas (to T_{jk}). Next, I performed weighted least square (WLS) analyses, using the mixed-effects method with REML, separately for each of the seven variables ($j = 1,\ldots,7$). The *T*-transformation maps relatively large values of alpha to relatively small T_{jk} values, so positive relationships between moderators and alpha will produce negative regression coefficients. I also combined the data for the seven variables and ran moderator analyses, using multilevel linear modeling (MLM) with the variable as an added level (Hox, 2010). The equation below shows the model that was used for these analyses.

$$T_{jk} = b_0 + b_1 X_{1jk} + b_2 X_{2jk} + \cdots + b_p X_{pjk} + V_j + S_{jk} + e_{jk}.$$

The dependent variable, T_{jk}, for study *k* and variable *j*, is predicted by the mean T_{jk} value of all study by variable combinations (b_0), *p* predictor variables (e.g., X_{pjk}) and the associated regression coefficients (e.g., b_p), and three error terms, V_j, S_{jk}, and e_{jk}. The error term V_j refers to the random deviation of the variable *j* from the overall effect; S_{jk} refers to the random deviation of study *k* within variable *j* from the mean effect in variable *j*; e_{jk} is the sampling error of the observed T_{jk} from the population effect size for study *k* within variable *j*.

Table 1.1. Coefficients from Regression Tests of Moderation of Number of Response Categories on Reliability

	Full Range			At Least Five Categories		
Variable	*K*	*N*	#RC	*K*	*N*	#RC
Extraversion	37	8,409	-0.03*	34	7,937	-0.03
Agreeableness	46	9,663	-0.03**	43	9,282	-0.03*
Conscientiousness	75	15,882	-0.04***	72	15,501	-0.03**
Neuroticism	53	17,747	-0.01†	43	14,089	-0.00
Openness	46	9,557	-0.03**	44	9,305	-0.03*
Job Satisfaction	62	22,362	-0.04***	53	18,890	-0.04*
Self-Esteem	25	8,110	-0.03	17	5,733	-0.04
Combined	344	48,056	-0.03***	306	46,101	-0.02***

Note. *K* = number of studies; *N* = sample size; #RC = number of response categories. †$p < .10$. *$p < .05$. **$p < .01$. *** $p < .001$.

Table 1.1 reports the WLS regression moderator analyses with the number of response alternatives as the predictor. Remembering that the *T*-transformation reverses the order of the reliabilities, we see that the results for both the full range of scale length and for scales with at least five response alternatives indicated that questionnaires with more response alternatives tended to have higher reliabilities. But this relationship was stronger and more consistent for the full range of response scale lengths than it was for the restricted range. Also, the results were quite consistent across the seven variables, and the results of the multilevel linear models support this generalizability. Thus, I have strong and generalizable evidence that scales with more response alternatives have higher reliabilities.

Table 1.2. Coefficients from Regression Tests of Moderation of Number of Response Categories and Label Format on Reliability

Variable	*K*	*N*	#RC	LF	#RC*LF
Extraversion	36	8,189	-0.03	-0.13*	0.03
Agreeableness	44	9,132	-0.03*	-0.06	0.03
Conscientiousness	72	14,984	-0.04***	0.01	-0.01
Neuroticism	45	12,255	-0.02	0.01	0.03
Openness	46	9,557	-0.03***	-0.07†	0.05*
Job Satisfaction[a]	36	10,363			
Self-Esteem	23	7,365	-0.09**	0.01	0.10**
Combined	302	31,908	-0.03***	-0.01	0.02*

Note. K = number of studies; *N* = sample size; #RC = number of response categories; LF = response category label format with ALL = 0 and END = 1; #RC*LF = interaction between number of response categories and response category label format.

[a]Job satisfaction only had one END scale, so I could not run these analyses.

†$p < .10$. *$p < .05$. **$p < .01$. *** $p < .001$.

I next compared the reliabilities of ALL and END scales. I first performed WLS ANOVAs, but found the results to be rather inconsistent. To better understand these relationships, I then conducted WLS regressions with reliability predicted by the (centered) number of response categories, category label format, and the interaction of these two predictors (see Table 1.2). Although the results tended to somewhat favor END scales, the main effect must be interpreted in light of the significant interaction effect. For ALL scales, using more response alternatives greatly improved reliability. However, interestingly, the relationship between the number of response alternatives and reliability was substantially weaker (and in some cases reversed) for END scales. The significant interaction for the multilevel model analyses across all seven variables provides evidence for generalizability. Figure 1.2 plots the results of the multilevel model with the T_{jk} statistics transformed back into alpha values. The plot shows that the two

response scale formats were roughly equivalent at around five or six response options, but they differed with either fewer or greater numbers of response categories.

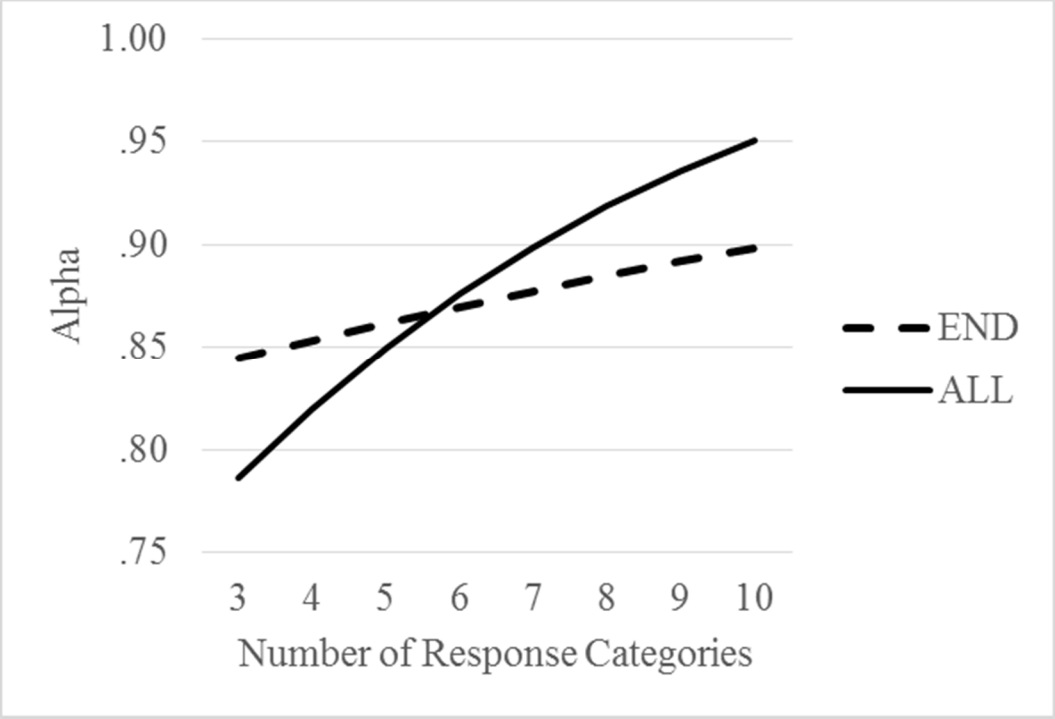

Figure 1.2. Multilevel linear model analysis of the combined dataset with number of response categories, the label format, and the interaction between these variables predicting reliability. END = only endpoints labeled; ALL = all response categories labeled.

This pilot study I conducted served four major purposes. First, I sought to replicate the common finding that the number of response categories positively predicts reliability. Second, I sought to explore whether this relationship would continue beyond five response alternatives. Third, I wanted to determine whether using ALL or END scales affected reliability. Fourth, I hoped to discover why the prior meta-analyses have reported no relationship between scale label format and reliability. The meta-analysis succeeded in meeting each of these goals. I found that the number of response alternatives did positively predict reliability, even for scales with five or more alternatives. The results did not overwhelmingly favor END or ALL scales; instead, the category label

format interacted with the number of categories to predict reliability. For questionnaires with close to an average number of response categories (around five) or fewer, it was better to only label endpoints, but for scales with a large number of response categories, it was better to label each category. I now consider a few factors that might explain this interaction effect.

As stated before, it has been argued that people perceive END scales, but not ALL scales, as being interval. Ordinal, or otherwise nonnormal, data are known to reduce the size of alpha (Greer et al., 2006). On the other hand, Monte Carlo simulations have demonstrated that categorizing continuous data into ordinal levels is most problematic for scales using two to four response options (Johnson & Creech, 1983). So, ALL scales with at least five options may be relatively resistant to the problems of nonnormality, assumed to be caused by the scale being ordinal. Furthermore, it has been shown that people do not perceive the psychological distance between categories as being equally spaced for scales with seven, as opposed to four or five, categories (Wakita, Ueshima, & Noguchi, 2012). Therefore, for longer response scales, even END scales may fail to be perceived as being truly interval. For scales with many response categories, the theorized benefits of labeling each category may then predominate.

This pilot study makes a few important contributions to our knowledge of the relationship between scale reliability and the number of response alternatives. It provides further evidence that, even for scales with at least five options, it is beneficial to add more response alternatives. More importantly, this research is the first to find an interaction between the response scale length and scale label format in predicting internal consistency reliability. This finding suggests a plausible explanation for why the prior meta-analyses have found a null effect for label format on reliability; its influence takes the form of an interaction with response scale length and not the form of a significant main effect.

This pilot study had the advantage of examining numerous questionnaires for seven different psychological constructs. However, I was unable to test the different theoretical ideas about which factors—total score variability, item homogeneity, and skewness—moderate the relationship between scale length and reliability, and I could not explore the effect that these scale modifications might have on validity. Most of all, the pilot study was not an experimental study, so inferring causality is not possible. The present study addresses each of these concerns.

1.5 Present Study

The present study expands upon the findings of the pilot study by experimentally examining the impact that two qualities of scales—the number of response alternatives and the label format for those categories—have on internal consistency reliability and construct validity. The experimental method enables more control over various confounds, and it allows me to directly test different theoretical ideas about which scale qualities moderate the relationship between the scale length and reliability. Lastly, the experimental method allows for proper tests of the construct validities of the various response scale formats. Based on the results of the pilot study and the literature review above, I developed several initial hypotheses and research questions.

H1: The number of response alternatives will positively predict reliability (see pilot study).

H2: There will be an interaction between the number of response alternatives and the response alternative label format in predicting reliability. In particular, END scales will have higher reliabilities for scales with four and five categories, and ALL scales will have higher reliabilities for scales with seven categories (see pilot study).

H3: Relatively low total score variability for shorter scales (Masters, 1974), item heterogeneity (Komorita & Graham, 1965), and greater skewness (Bandolos & Enders, 1996; Bernstein & Teng, 1989) will increase the extent to which scale length predicts reliability.

RQ1: Will total score variability, item homogeneity, or skewness best predict the association between scale length and reliability?

H4: The number of response categories will positively predict validity (e.g., Alwin, 1997; Preston & Colman, 2000).

RQ2: Will there be an interaction between the number of response levels and the response category label format in predicting validity?

RQ3: Will total score variability, item homogeneity, and skewness influence the extent to which scale length predicts validity?

Chapter 2

METHOD

2.1 Participants

The participants in this study were 893 (490 females, 400 males, 3 no response) U.S. workers on Amazon Mechanical Turk who took our surveys in exchange for financial compensation. This data collection method has been found to be acceptable. In a recent study (Paolacci, Chandler, & Ipeirotis, 2010), a large sample from Mechanical Turk was shown to compare favorably to samples collected via other internet sources and typical undergraduate samples in terms of attention, effect sizes for results, and demographics. Indeed, the sample in this study was older (M = 35.35, SD = 12.44) and more ethnically diverse than a typical undergraduate sample, including 75 Black (8%), 60 Asian (7%), 674 White (76%), 48 Hispanic (5%), and 34 multicultural or other (4%) respondents. They also differed in education: 11 (1%) had not completed high school, 93 (10%) were high school graduates, 348 (39%) had completed some college, 327 (37%) had bachelor's degrees, 99 (11%) had master's degrees, and 14 (2%) had doctoral degrees. So, 440 participants (50%) held college degrees.

2.2 Procedure and Materials

Participants were randomly assigned to one of six survey conditions. The Likert-type rating scales for each item in the survey either had each point verbally labeled (ALL), or it only had the endpoints verbally labeled with numerical values between the two poles (END). For these two formats, the scales had four, five, or seven response categories. I decided on these particular response scale lengths because, as discussed in the introduction, it is generally agreed that reliability improves beyond two or three categories, but there is less agreement on whether it improves beyond four or five categories. Also, few authors use more than seven categories. For example, a recent meta-analysis encompassing research from 24 psychology, marketing, management, and education journals, found that only 31 (1%) of 2,524 alpha coefficients arose from scales

with eight or more response categories, and most (91%) used either five or seven levels (Peterson & Kim, 2013). Thus, the present study used a 2 (label format: ALL or END) x 3 (number of response categories: four, five, or seven) between-subjects design.

Each participant was asked to fill out two sets of questionnaires. The first set of questionnaires contained six scales—extraversion and neuroticism from the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991), the Revised Self-Monitoring Scale (RSMS; Lennox & Wolfe, 1984), Rosenberg's Self-Esteem Scale (RSE; Rosenberg, 1965), and the social anxiety and dysphoria scales from the Inventory of Depression and Anxiety Symptoms (IDAS; Watson et al., 2007). These six scales were chosen because they differ in total score variability, item homogeneity, and skewness. The participants were randomly assigned to fill out these six questionnaires using one of the six survey formats described above. The second set of questionnaires measured extraversion and neuroticism with questions from the International Personality Item Pool (IPIP, 2001). Every participant was given five items with four response alternatives, five items with five response alternatives, and five items with seven response alternatives for both of these constructs, but the format for the response alternative labels matched the format that was assigned for the first set of questionnaires. So, for the second set of questionnaires, there were only two conditions: ALL or END. After completing the online informed consent, the participants reported their education level, and at the end of the survey, they were asked to report on age, gender, and ethnicity (see Appendix B for the wording of the survey).

Table 2.1. Response Category Labels for ALL Scales

	Agreement		Quantity		Accuracy	
Level	Five	Seven	Five	Seven	Five	Seven
1	Strongly Disagree	Strongly Disagree	Not At All	Not At All	Very Inaccurate	Very Inaccurate
2	Disagree	Disagree	A Little Bit	A Little Bit	Moderately Inaccurate	Inaccurate
3	Neither Disagree Nor Agree	Slightly Disagree	Moderately	Somewhat	Neither Inaccurate Nor Accurate	Moderately Inaccurate
4	Agree	Neither Disagree Nor Agree	Quite A Bit	Moderately	Moderately Accurate	Neither Inaccurate Nor Accurate
5	Strongly Agree	Slightly Agree	Extremely	Quite A Bit	Very Accurate	Moderately Accurate
6		Agree		Very		Accurate
7		Strongly Agree		Extremely		Very Accurate

Note. The four-point ALL scales use the same categories as the five-point ALL scales without the middle category. The END scales have the same endpoints as the ALL scales, but they only have numerical values between the endpoints.

The particular labels given for the different response categories for ALL scales with five and seven levels are presented in Table 2.1. ALL scales with four levels are identical to those with five levels except that they lack the middle category, and END scales have the same labels as ALL scales at the endpoints but have only numerical values between the endpoints. The BFI, RSMS, and the RSE use agreement labels, the IDAS scales use quantity labels, and the IPIP scales use accuracy labels. The actual labels were chosen based on empirical work showing that people perceive these category labels as being equally spaced (Bass et al., 1974; Dobson & Mothersill, 1979). See Figure 2.1 for an example item in each of the six formats. The questionnaires that were used in this study are described below.

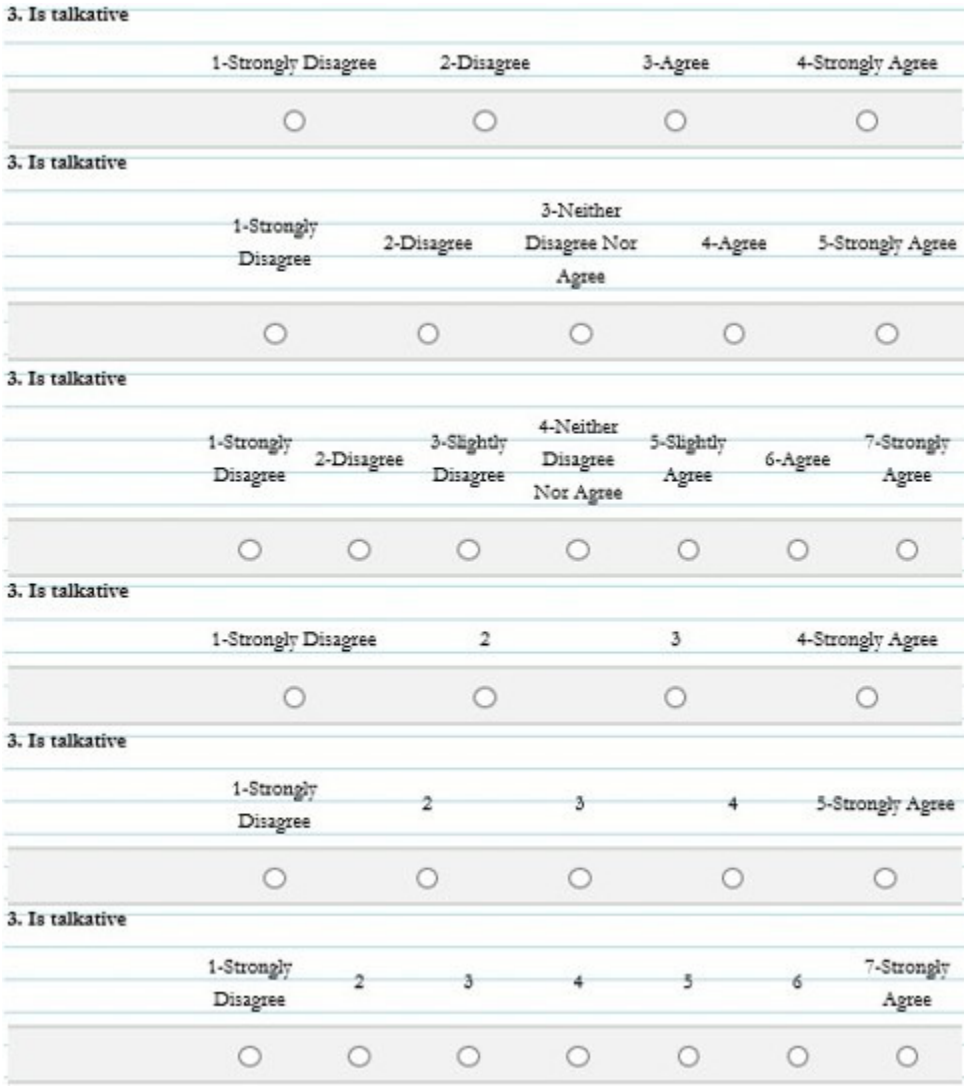

Figure 2.1. An example item with ALL four-point, ALL five-point, ALL seven-point, END four-point, END five-point, and END seven-point conditions, respectively.

2.2.1 The Big Five Inventory (John et al., 1991)

I included the eight-item extraversion and neuroticism scales from the BFI. The BFI scales have been shown to have relatively high reliabilities, usually above α = .80, and the BFI has been shown to perform well on several types of validity tests (John, Naumann, & Soto, 2008). The scales are unidimensional, and they correlate well with other measures of the same two dimensions of the Big Five traits.

2.2.2 The Revised Self-Monitoring Scale (Lennox & Wolfe, 1984)

This scale measures the extent to which people adjust their self-presentation in accordance with the demands of various social situations. This particular measure of self-monitoring has 13 items which form two factors: ability to modify one's self-presentation and sensitivity to the expressive behavior of others. The reliabilities have been adequate but somewhat low for the two factors, $\alpha = .77$ and $\alpha = .70$, and for the full scale, $\alpha = .75$ (Lennox & Wolfe, 1984). The items for the RSMS were selected, in part, to minimize skewness, so the scale should not be notably skewed. The construct of self-monitoring is conceptually related to both extraversion and neuroticism, and it is statistically correlated with measures of these constructs as well (Lennox & Wolfe, 1984).

2.2.3 Rosenberg's Self-Esteem Scale (Rosenberg, 1965)

This instrument is a ten-item measure of global self-esteem. It has demonstrated good reliability and validity (Crandall, 1973). However, recent work has demonstrated that it may form two factors (Huang & Dong, 2012): one for the five positively worded items and one for the five negatively worded items. The RSE is also skewed because most people report having high self-esteem. Self-esteem is correlated with both extraversion and neuroticism (Bosson & Swann, 2009).

2.2.4 Inventory of Depression and Anxiety Symptoms (Watson et al., 2007)

Subjects were asked to fill out the five-item social anxiety and the ten-item dysphoria subscales of the IDAS. Both of these subscales have been shown to be valid and reliable; the internal consistencies were generally well above $\alpha = .80$ in several scale validation samples (Watson et al., 2007). When all items from all subscales were entered into a factor analysis, one factor emerged for each subscale, so it can be assumed that each subscale is unidimensional. Although no skewness statistics have been reported, I expected the scales to be skewed because they measure the normatively unusual symptoms of psychopathology. Based on the results of a large meta-analysis (Kotov,

Gamez, Schmidt, & Watson, 2010), dysphoria should be related to neuroticism, and social anxiety should be related to extraversion and neuroticism.

2.2.5 The International Personality Item Pool Big Five scales (IPIP, 2001)

Lastly, I administered 30 items from the IPIP to assess extraversion and neuroticism. For both constructs, five questions were answered with each of four, five, and seven response levels. These scales come in several formats, but the ten-item versions have been shown to have reliabilities around α = .86 for both constructs (IPIP, 2001). The scales are unidimensional and have been shown to relate well with other measures of the Big Five traits.

2.3 Statistical Analysis

For the hypotheses and research questions concerning scale reliabilities, I first calculated and tabulated the coefficient alphas for each of the six formats of the six scales used in the first set of questionnaires—extraversion and neuroticism from the BFI, RSMS, RSE, and dysphoria and social anxiety from the IDAS. Because alphas are not normally distributed, I used the following transformation for all statistical analyses involving alpha (Hakstian & Whalen, 1976):

$$T = (1 - a)^{1/3}.$$

After transforming these 36 alphas (six questionnaires and six response formats), I obtained the average *T* statistic for each of the 2 (label: ALL, END) X 3 (number of categories: four, five, seven) response scale formats across the six questionnaires. These mean *T* values were then transformed back to alpha for reference purposes:

$$a = 1 - T^3.$$

To test the first hypothesis that the number of response categories positively predicts reliability, I ran two *k*-sample chi-square significance tests for independent alphas (Hakstian & Whalen, 1976) to compare the average reliabilities for the three scale lengths for both END and ALL scales. This chi-square test compares *k* alphas from *k* samples to determine if any alpha(s) is significantly different from any other alpha(s). To

test the second hypothesis that there will be an interaction between the number of response levels and the response category label format, I ran three more chi-square tests to compare the average reliabilities of END and ALL scales at each of the three response scale lengths. Based on my hypotheses, I had predicted that all three tests should yield a significant finding: END scales should have higher average reliabilities for four and five levels, but ALL scales should have the higher average reliability at seven levels.

As an additional test of hypotheses one and two, I computed the item-total correlations for each item within each questionnaire. The item-total correlations, for each questionnaire, represent (though imperfectly) how much each item contributes to the reliability. I separated these correlations by the three response categories and two label format conditions and ran an itemmetric ANOVA. This means that the (N = 54) items from the first set of questionnaires comprised the sample, rather than the respondents. The item-total correlations were the dependent variables and the number of response categories and label format were the within-item independent variables. The effect for the number of response categories and the interaction between response scale length and label format represent item-level tests for hypotheses one and two, respectively. This particular analysis was run because it is a very powerful test of the first two hypotheses. It makes comparisons at the item level, rather than the questionnaire level, and having two sets of within-item variables allows for generalizable tests of the hypotheses that are not as influenced by differences in item-total correlations (reliabilities) between items (questionnaires). The results of this test should be interpreted similarly to the results of the chi-square tests described above. However, this technique examines item-total correlations, whereas the chi-square tests examine alpha.

Next, I computed statistics for total score variability, item homogeneity, and skewness for the first set of questionnaires. The hypotheses for item homogeneity and skewness pertained to the descriptive statistics for the questionnaire in general, instead of the descriptives at various response scale lengths, so I first computed statistics to

represent the average descriptive statistics across the conditions. To make the data comparable across the six scale format conditions, I divided each individual's total score by the number of response categories (four, five, or seven) and combined the data across the six conditions. Then, for both ALL and END scales, I computed score variability, item homogeneity, and skewness statistics. For score variability, I computed the mean score variability for each questionnaire and estimated the statistics at five response categories. For item homogeneity, I performed principal components analyses (PCA) for each questionnaire and I tabulated the proportion of variance accounted for by the first factor.

Masters (1974) hypothesized that questionnaires which have relatively low total score variability at shorter response scale lengths would have relatively higher reliabilities at longer response scale lengths. So, in addition to calculating and tabulating the descriptive statistics for each response scale length, I computed the ratio of these statistics for four and seven-point scales; those with relatively more variability at seven response categories should hypothetically have higher reliabilities at seven response categories. At this point, I calculated the two sets of *k*-sample chi-square significance tests for independent alphas for each of the six scales separately; this allowed me to determine for which questionnaires the scale length and label format predicted reliability. Comparing these results with the pattern of total score variability, item homogeneity, and skewness among the six scales allowed me to test the hypothesis that these factors interact with scale length to predict reliability.

As stated above, every participant answered questions with four, five, and seven levels for the second set of questionnaires—extraversion and neuroticism from the IPIP. I took the participants' mean responses for all three response scale lengths, divided by the scale length, and then averaged these three values. These statistics, for IPIP extraversion and neuroticism, were correlated with the scores on the first set of questionnaires to provide construct validity coefficients, which should not be differentially

influenced by methodological variance for any of the three numbers of response alternatives (Chang, 1994). Analogous to the reliability analyses described above, I applied Fisher's *r*-to-*z* transformation and calculated the average of the absolute values of the correlations between scores on the two sets of questionnaires for each of the six conditions. The test of equality of multiple independent correlations (Hays, 1994) was used to compare the average validities across scale lengths and to compare ALL and END scales at each number of response levels. These analyses were done to test my fourth hypothesis that the number of response categories will positively predict validity, and they also address my research question about whether there will be an interaction between the number of response levels and the response category label format. Moreover, I separately performed the same tests of equality for correlations for each of the six scales and compared these results with the pattern of total score variability, item homogeneity, and skewness among the six scales to assess the third research question regarding these relationships.

Lastly, I conducted one final set of analyses to test the relationship between scale length, label format, and validity. After dividing each respondents' scores on the first six questionnaires by the scale length and combining the data across the six conditions, I performed two sets of multivariate multiple regression analyses—one for each of extraversion and neuroticism from the IPIP. I also performed these analyses with the *T*-transformed alpha of the IPIP scale as a covariate to determine whether the relationships between scale length, label format, and validity continue to hold true after accounting for the impact of reliability. The independent variables in these analyses were the score on the particular IPIP scale being tested, the scale length, the label format, and all interactions between these predictors. The dependent variables were the scores on the first six questionnaires. The interactions between the IPIP score and scale length test hypothesis four, and the triple interactions test research question two. These final

analyses are a powerful test for any generalizable relationship between the various scale modifications and construct validity.

Chapter 3

RESULTS

Prior to analysis, I examined the data for the 893 respondents for missing values and outliers. I ran missing values analyses and found that there were no items that were missing values for more than 5% of the cases, so I concluded that there were no problematic items. I then screened the data for participants who responded to less than 5% of the items; I deleted seven cases based on this criteria, leaving 886 respondents. I examined the six conditions for univariate outliers on all study variables—BFI-E, BFI-N, RSMS, SES, IDAS-D, IDAS-S, IPIP-E, and IPIP-N. For each condition and each questionnaire, I computed the *z*-scores and looked for any values that were outliers at $p < .001$. I only found one case for the RSMS, and I removed this case from all analyses that involved the RSMS.

3.1 Descriptive Statistics

Table 3.1 and Table 3.2 summarize the descriptive statistics for all six questionnaires in the first set of questionnaires on the attributes of sample size, mean score variability, item homogeneity, and skewness. Table 3.1 shows these statistics for the ALL and END conditions, merged across the three response scale length conditions. Table 3.2 shows these statistics for all six conditions, and it also shows the ratio of the descriptive statistics for the seven and four-point response scales. Note that the sample sizes were roughly equivalent for all six conditions. This was true for the ALL ($N = 470$; 53%) versus END ($N = 416$; 47%) manipulation ($\chi^2(1; N = 886) = 3.29$, $p = .07$), and it was true for the response scale length conditions ($\chi^2(2; N = 886) = .52$, $p = .77$): four ($N = 288$; 33%), five ($N = 305$; 34%), and seven ($N = 293$; 33%).

Table 3.1. Summary Descriptive Statistics for Sample Size, Score Variability, Item Homogeneity, and Skewness for the First Set of Questionnaires

	N		Variability		% PCA		Skewness	
Questionnaire	ALL	END	ALL	END	ALL	END	ALL	END
BFI-E	454	403	0.72	0.77	60	59	0.11	0.07
BFI-N	458	401	0.73	0.83	57	58	-0.03	0.06
RSMS	444	396	0.41	0.42	46	41	-0.36	-0.26
SES	444	394	0.70	0.86	60	63	-0.53	-0.61
IDAS-D	453	388	0.88	1.01	61	64	0.52	0.27
IDAS-S	461	406	1.22	1.28	71	73	0.73	0.54

Note. These statistics represent the values obtained after dividing all item statistics by the number of response categories and combining the statistics across response scale length conditions. Variability = the mean score variance corrected to be a five category scale; % PCA = the percentage of variance accounted for by the first component in a principal components analysis; Skewness = the skewness of the scale; BFI-E = Big Five Inventory extraversion scale; BFI-N = Big Five Inventory neuroticism scale; RSMS = Revised Self-Monitoring Scale; IDAS-D = Inventory of Depression and Anxiety Symptoms dysphoria scale; IDAS-S = Inventory of Depression and Anxiety Symptoms social anxiety scale.

Table 3.2. Descriptive Statistics for Sample Size, Score Variability, Item Homogeneity, and Skewness for the First Set of Questionnaires

	N		Variability		% PCA		Skewness	
Questionnaire	ALL	END	ALL	END	ALL	END	ALL	END
Four								
BFI-E	144	134	0.65	0.67	63	53	0.07	-0.24
BFI-N	141	134	0.59	0.69	54	52	-0.09	0.19
RSMS	136	133	0.34	0.36	45	39	-0.15	-0.19
SES	139	130	0.57	0.80	62	60	-0.27	-0.69
IDAS-D	141	126	0.76	0.90	59	59	0.60	0.34
IDAS-S	147	136	1.09	1.17	72	69	0.93	0.55

Table 3.2-*Continued*

Five								
BFI-E	155	143	0.74	0.73	62	60	0.19	0.25
BFI-N	156	142	0.73	0.77	58	59	-0.08	0.19
RSMS	153	141	0.39	0.41	43	46	-0.33	-0.15
SES	150	140	0.68	0.80	60	64	-0.55	-0.56
IDAS-D	154	140	0.74	0.88	57	63	0.29	0.27
IDAS-S	154	144	1.17	1.09	71	71	0.65	0.63
Seven								
BFI-E	155	126	0.73	0.89	56	63	0.13	0.24
BFI-N	161	125	0.82	0.99	57	62	0.16	-0.01
RSMS	155	122	0.50	0.46	49	40	-0.43	-0.35
SES	155	124	0.83	0.98	59	64	-0.60	-0.55
IDAS-D	158	122	1.06	1.22	64	67	0.80	0.32
IDAS-S	160	126	1.30	1.61	71	77	0.84	0.55
7/4[a]								
BFI-E			1.13	1.33	0.88	1.19	1.92	-0.97
BFI-N			1.38	1.43	1.05	1.21	-1.68	-0.08
RSMS			1.46	1.26	1.09	1.05	2.98	1.87
SES			1.45	1.22	0.96	1.07	2.24	0.81
IDAS-D			1.40	1.36	1.08	1.14	1.33	0.95
IDAS-S			1.19	1.38	0.99	1.13	0.90	1.00

Note. Variability = the mean score variance corrected to be a five category scale; % PCA = the percentage of variance accounted for by the first component in a principal components analysis; Skewness = the skewness of the scale.

[a]The ratio of the statistics for seven and four categories.

Table 3.1 shows that the six questionnaires differed greatly on score variability. The two IDAS questionnaires and, to a lesser extent, the BFI-N and SES had the most score variability in general. Also, as indicated in Table 3.2, all six questionnaires increased in relative score variability with increased response scale length. Importantly, counter to Masters' (1974) assumption, the data did not show that the questionnaires with the least total score variability had the greatest concurrent increase in total score variability with increased response scale length.

Next, consider the statistics for item homogeneity, which I have operationally defined as the percent of variance accounted for by the first component of a PCA. Table 3.1 shows that, as predicted, the RSMS was substantially more heterogeneous than the other questionnaires. The IDAS-S was the most homogeneous questionnaire, and the other four measures were all similar in item homogeneity. Interestingly, Table 3.2 shows that, for the END but not ALL conditions, item homogeneity increased with the number of response alternatives. This finding indicates that for END response scales, increasing scale length improves "factorial validity" (Lozano et al., 2008).

Lastly, Table 3.1 provides generalized skewness statistics. The questionnaires varied in both the direction and the magnitude of skew. As expected, the IDAS questionnaires and the SES were the most skewed, but unexpectedly, the RSMS was fairly skewed as well. Table 3.2 does not give strong evidence for any general relationship between the number of response categories and skewness. If anything, increasing the number of response categories amplifies the observed skew. Overall, the six questionnaires were rather different in these three descriptive statistics, and the data shows that they generally behaved as I had expected.

Table 3.3. Alphas for the First Set of Questionnaires by Label Format

	BFI-E	BFI-N	RSMS	SES	IDAS-D	IDAS-S	Average
ALL							
4	.90†	.88	.89	.93	.92	.90	.90
5	.90	.89	.89	.92	.92	.90	.90
7	.87	.89	.91*	.92	.94	.89	.91
χ^2	2.17	0.61	2.12	0.43	3.08	0.08	0.03
END							
4	.86	.86	.86	.92	.92	.88	.89
5	.89	.90	.90	.94	.94	.90	.91
7	.90	.91	.87	.93	.95	.93*	.92
χ^2	4.12	5.82†	3.22	1.03	3.23	5.39†	2.81

Note. Significant alphas indicate that the scale reliability is significantly higher than the reliability of the scale with the same number of response categories but different category label format. Significant χ^2 statistics indicate that the reliabilities differ among the 4, 5, and 7 response scale lengths for the particular category label format.

†$p < .10$. *$p < .05$.

3.2 H1 and H2: Response Scale Length, Label Format, and Reliability

Table 3.3 provides the reliabilities of the six questionnaires for each of the six conditions. Note that the reliabilities were unusually high for each questionnaire and condition. Indeed, Buhrmester, Kwang, and Gosling (2011) administered the BFI-E, BFI-N, and SES to samples from the Amazon Mechanical Turk website at varying levels of monetary compensation, and they generally had lower reliabilities across conditions for the BFI-E (.85-.88), the BFI-N (.87-.89), and the SES (.90-.91) than reported in Table 3.3. Moreover, the RSMS alphas ranged from .86 to .91 (range = .05), but the scale developers reported an alpha of .75 (Lennox & Wolfe, 1984). I will more thoroughly describe the probable causes of this issue in the discussion section, but this is likely a result of the method used to collect the sample. As mentioned in the introduction and as proven in Appendix A, these inflated reliabilities have the unfortunate consequence of

shrinking differences in reliabilities for any given increase in average inter-item correlation. For example, if the five-point ALL RSMS had a standardized alpha reliability of .75 and if the differences in average inter-item correlation between conditions matched the present study's results, the standardized alphas would have ranged from .66 to .81 (range = .15). The chi-square test for differences in reliability would have been statistically significant for the END, χ^2 (2, k = 3) = 6.22, p = .04, but not ALL, χ^2 (2, k = 3) = 3.28, p = .19, response scale condition. For this reason, I also present the average inter-item correlations in Table 3.4 because these often differ more noticeably between the conditions.

Table 3.4. Average Inter-Item Correlations for the First Set of Questionnaires by Label Format

	BFI-E	BFI-N	RSMS	SES	IDAS-D	IDAS-S	Average
ALL							
4	.53	.47	.40	.57	.53	.64	.53
5	.52	.51	.38	.55	.51	.64	.52
7	.46	.51	.44	.54	.59	.63	.53
END							
4	.42	.44	.32	.55	.54	.61	.49
5	.51	.53	.41	.60	.59	.64	.55
7	.53	.56	.35	.60	.63	.71	.57

Hypothesis one was that there would be a generalizable, positive relationship between the number of response categories and reliability. As shown by the average columns in Table 3.3 and Table 3.4, the average of the coefficient alphas and average inter-item correlations for the six questionnaires did monotonically increase with response scale length for the END response scales but not the ALL response scales. However, because the chi-square tests were not significant for either label format, these analyses did not support the first hypothesis.

I also ran an itemmetric ANOVA—meaning that the items comprise the sample rather than the participants—with the (Fisher *r-z* transformed) item-total correlations for each item in the first six questionnaires ($N = 54$) as the dependent variable. The number of response categories and the category label format were the within-item variables. As shown in Table 3.5, there was a significant main effect for the number of response categories. Using Bonferroni adjustments, the four ($M = .67$, 95% CI = [.64, .70]) and five-point ($M = .69$, 95% CI = [.67, .71]) conditions differed, $p < .01$, and the four and seven-point ($M = .70$, 95% CI = [.68, .73]) conditions differed, $p < .001$, but the five and seven-point conditions did not differ, $p = .23$. Thus, although the analyses of coefficient alpha reliabilities did not support the first hypothesis, the itemmetric analyses showed a positive relationship between the number of response alternatives and item-total correlations.

Table 3.5. Itemmetric ANOVA for the Item-Total Correlations in the First Set of Questionnaires

	Wilk's λ	*p*	η^2
#RC	0.69	< .001	.31
LF	1.00	.63	.00
#RC*LF	0.62	< .001	.38

Note. η^2 = partial eta-squared; #RC = number of response categories; LF = label format.

For the second hypothesis, I predicted an interaction between the number of response alternatives and label format, such that END scales would have higher reliabilities for four and five-point scales and ALL scales would have higher reliabilities for seven-point scales. Examining the average columns for reliabilities in Table 3.3 and for average inter-item correlations in Table 3.4, there was no evidence for this effect. There were no significant differences in reliabilities between the two label formats for any of the

response scale lengths and even the average inter-item correlations were similar. A positive association between response scale length and reliability for ALL but not END scales would also have evidenced this effect. However, this effect was not found either, and in fact, there was some (non-significant) support for the opposite conclusion: There was a minimally stronger effect of response scale length for END than for ALL scales.

I also examined the interaction effect for the itemmetric analyses in Table 3.5. There was a significant interaction between the number of response options and label format in predicting item-total correlations. Using Bonferroni adjustments, for the four-point conditions, the ALL condition (M = .69, 95% CI = [.66, .71]) had significantly higher item-total correlations than the END condition (M = .65, 95% CI = [.62, .68]), $p < .001$. For the five-point conditions, the END condition (M = .70, 95% CI = [.68, .72]) had significantly higher item-total correlations than the ALL condition (M = .68, 95% CI = [.65, .70]), $p = .02$. Lastly, for the seven-point conditions, the END condition (M = .71, 95% CI = [.68, .74]) had significantly higher item-total correlations than the ALL condition (M = .69, 95% CI = [.67, .72]), $p = .04$. For both coefficient alpha for questionnaires and item-total correlations, the hypothesized interaction was not found in the data; instead, there was evidence that the END response scales actually had a stronger positive relationship between scale length and reliability.

3.3 Analyses of Individual Questionnaires and Education Level

The third hypothesis—to be discussed shortly—was that relatively low total score variability for shorter scales, item heterogeneity, and greater skewness would increase the extent to which scale length predicts reliability. To evaluate this hypothesis, I compared the descriptive statistics on individual questionnaires, summarized in Table 3.1 and Table 3.2, to the statistics on the associations between the number of response categories and reliability and the average inter-item correlation for the individual questionnaires. However, before proceeding to make these comparisons, I must know

which questionnaires showed the strongest relationship between response scale length and reliability. I consider the individual questionnaires below.

The Table 3.3 columns for individual questionnaires show a positive, monotonic relationship between the number of response alternatives and reliability for the BFI and IDAS questionnaires in the END conditions, but the relationships only reached marginal significance in two cases: BFI-N and IDAS-S. Again, the effects were more apparent for the average inter-item correlations in Table 3.4, which supports the interpretation that some effects would have reached statistical significance if the reliabilities had not been especially high for each questionnaire.

In a further attempt to identify the questionnaires that were most affected by response scale length, I explored another potential moderator: education level. Table 3.6 and Table 3.7 report the reliabilities and average inter-item correlations, respectively, for the three response scale lengths for two groups: those who hold college degrees and those who do not. The educated group had a significant relationship between response scale length and reliability for two scales—the BFI-N and IDAS-D—though a similar, non-significant pattern emerged for the IDAS-S. The uneducated group had a significant relationship for the RSMS, but some other questionnaires evinced non-significant relationships between response scale length and reliability that were somewhat similar to those for the RSMS. Interestingly, the average column for average inter-item correlations (Table 3.7) clearly shows that the greatest difference in average inter-item correlations was between four and five response levels for the uneducated group, but for the educated group, the largest difference was between five and seven levels. I return to this observation in the discussion section, but it probably results from differences in cognitive ability between the educated and uneducated groups. Whatever the cause, this differential relationship certainly could have precluded finding more statistically significant differences in reliabilities between the response scale lengths in the data reported in Table 3.3.

Table 3.6. Alphas for the First Set of Questionnaires by Education Level

	N	BFI-E	BFI-N	RSMS	SES	IDAS-D	IDAS-S	Average
Uneducated								
4	144-152	.88	.88†	.85	.93	.92	.90	.89
5	137-145	.90	.91	.90	.94	.94†	.90	.92
7	138-144	.89	.90	.91	.93	.94	.90	.91
χ^2		1.77	1.45	9.06*	0.89	2.04	0.23	1.83
Educated								
4	123-132	.88	.84	.90*	.93	.92	.88	.89
5	152-159	.89	.89	.88	.92	.91	.89	.90
7	139-142	.88	.90	.88	.92	.94	.92	.91
χ^2		0.08	6.04*	1.45	0.19	7.01*	4.40	0.88

Note. Significant alphas indicate that the scale reliability is significantly higher than the reliability of the scale with the same number of response categories but different education level. Significant χ^2 statistics indicate that the reliabilities differ among the 4, 5, and 7 response scale lengths for the particular education level. *N* = the range of sample sizes.

†$p < .10$. *$p < .05$.

Table 3.7. Average Inter-item Correlations for the First Set of Questionnaires by Education Level

	BFI-E	BFI-N	RSMS	SES	IDAS-D	IDAS-S	Average
Uneducated							
4	.47	.49	.31	.56	.54	.63	.51
5	.53	.55	.42	.60	.59	.65	.56
7	.49	.54	.43	.58	.61	.66	.56
Educated							
4	.48	.40	.42	.56	.52	.59	.50
5	.49	.50	.36	.55	.51	.62	.51
7	.49	.53	.37	.54	.62	.69	.55

I conducted one final analysis to further elucidate the impact of the confounding factor of education. I ran an itemmetric ANOVA on item-total correlations with three within-item variables: response scale length, label format, and education level. As shown

in Table 3.8, response scale length, and the interaction between response scale length and label format was significant, as with the previous itemmetric ANOVA. Education level was significant; the uneducated group (M = .70, 95% CI = [67, .73]) had higher average item-total correlations than the educated group (M = .68, 95% CI = [.66, .70]). Response scale length and education level also had a significant interaction. For the uneducated group, the four (M = .67, 95% CI = [.64, .70]) and five-point (M = .71, 95% CI = [.69, .74]) conditions differed, $p < .001$, and the four and seven-point (M = .72, 95% CI = [.69, .74]) conditions differed, $p < .001$, but the five and seven-point conditions did not differ, $p = .99$. The pattern was very different for the educated group; the four (M = .67, 95% CI = [.64, .70]) and five-point (M = .67, 95% CI = [.64, .69]) conditions didn't differ, $p = .99$, and the four and seven-point (M = .70, 95% CI = [.67, .73]) conditions didn't differ, $p = .15$, but the five and seven-point conditions differed, $p = .01$. These results agree with those reported for reliability and average inter-item correlations in Table 3.6 and Table 3.7.

Table 3.8. Itemmetric ANOVA for the Item-Total Correlations in the First Set of Questionnaires with Education Level

	Wilk's λ	p	η^2
#RC	0.68	< .001	.32
LF	1.00	.81	.00
Edu	0.78	< .001	.22
#RC*LF	0.70	< .001	.30
#RC*Edu	0.77	< .01	.23
LF*Edu	0.83	< .01	.17
#RC*LF*Edu	0.93	.15	.07

Note. η^2 = partial eta-squared; #RC = number of response categories; LF = label format; Edu = education level.

Lastly, label format and education level significantly interacted. For the ALL conditions, the uneducated (M = .69, 95% CI = [.67, .72]) and educated groups (M = .69,

95% CI = [.66, .71]) did not differ, $p = .55$, but for the END conditions, the uneducated group ($M = .71$, 95% CI = [.68, .74]) had significantly higher average item-total correlations than the educated group ($M = .67$, 95% CI = [.65, .70]), $p < .001$. Many of these results make sense from the perspective of cognitive ability and the theoretical perspective of satisficing versus optimizing responding styles (Krosnick, 1991). As will be described more completely in the discussion section, the uneducated group may have been more prone to satisficing, particularly under some conditions, and this could have lead them to respond to items in a uniform manner. For now, it is important to simply acknowledge that the confounding factor of education level probably made it more difficult to uncover effects of the moderators of initial interest.

3.4 H3 and RQ1: Moderators of Response Scale Length and Reliability

In light of the above analyses, the BFI-N and the IDAS questionnaires seem to most consistently have a monotonic relationship between response scale length and reliability. The RSMS sometimes showed a positive relationship between response scale length and reliability, but it was not consistently monotonic among the various categories of response scale label format and education level. Thus, the BFI-N, the two IDAS questionnaires, and to a lesser degree the RSMS will be considered to be the questionnaires which had a positive relationship between response scale length and reliability for hypothesis three and research question one. I now consider hypothesis three regarding score variability, item homogeneity, and skewness.

Masters (1974) hypothesized that those questionnaires that had the least total score variability at shorter response scale lengths, should benefit the most in reliability by increasing the number of response options. By Table 3.2, the RSMS had the least score variability at four response categories. So, Masters' (1974) hypothesis was not supported by the data. The two IDAS questionnaires and, to a lesser extent, the BFI-N and SES had the most score variability in general, but the RSMS had the least total score variability. Table 3.2 shows that the BFI-N, IDAS-D, and RSMS had the largest increase

in variability with response scale length. To summarize, the data did not show that the questionnaires with lowest variability at four levels had the greatest effect. Instead, there was evidence that the questionnaires that generally had the highest score variability and that increased in variability with response scale length tended to have the strongest positive relationship between the number of response categories and reliability.

I next examined the hypothesis that the questionnaires with the most item heterogeneity should have the strongest effect of response scale length on reliability (Komorita & Graham, 1965). The RSMS had the most item heterogeneity, as defined by the size of the first component of a PCA, but it did not have a particularly strong relationship between response scale length and reliability. Viewing Table 3.1, there was little support for the opposite conclusion. The IDAS questionnaires had the most item homogeneity, but otherwise, the results are inconsistent with this conclusion. By Table 3.2, the BFI-N and the IDAS questionnaires had the strongest increase in item homogeneity with response scale length. As far as item homogeneity or heterogeneity is concerned, the extent to which increasing the number of response options increases item homogeneity seems to best predict whether reliability will increase as well.

Finally, I had predicted based on data from Monte Carlo simulations that the questionnaires with greater levels of skewness would have stronger positive associations between the response scale length and reliability. Table 3.1 shows that the IDAS-S and SES were the most skewed questionnaires, and the IDAS-D and RSMS were more moderately skewed. Thus, the skewness hypothesis explains the positive association between response scale length and reliability for the IDAS questionnaires and the RSMS, but it does not explain the effect for the BFI-N or the null effect for the SES.

Overall, the data indicate that high score variability, co-occurring increases in either score variability or item homogeneity with response scale length, and greater levels of skewness were each associated with a positive relationship between the number of response options and reliability. Therefore, hypothesis three was partially confirmed, but

the descriptive statistics did not all moderate the relationship as predicted. To address research question one, skewness best predicted the relationship in the manner hypothesized. Overall, however, the extent to which increasing response scale length increased item homogeneity was probably the best predictor of this relationship.

3.5 H4: Response Scale Length and Validity

My fourth hypothesis was that the number of response categories would positively predict validity. As I explained in the Method section, I created scores for the IPIP extraversion and neuroticism questionnaires that were independent of response scale length, and I correlated these scores with the first six questionnaires for validity coefficients. Table 3.9 and Table 3.10 include the correlation coefficients for the IPIP extraversion and neuroticism questionnaires, respectively.

Table 3.9. Correlations between First Set of Questionnaires and the International Personality Item Pool Extraversion Measure

	BFI-E	BFI-N	RSMS	SES	IDAS-D	IDAS-S	Average
ALL							
4	.84	-.51	.27	.53	-.53	-.54	.56
5	.85	-.42	.45	.40	-.39	-.51	.53
7	.86	-.40	.48	.48	-.32	-.48	.54
χ^2	0.38	1.60	4.84†	2.03	4.68†	0.51	0.15
END							
4	.83	-.45	.32	.57	-.49	-.61	.57
5	.87	-.49	.36	.53	-.50	-.64	.60
7	.82	-.52	.33	.60	-.42	-.53	.56
χ^2	1.77	0.54	0.19	0.71	0.76	1.70	0.18

Note. Significant correlations indicate that the correlation is significantly higher than the correlation of the scale with the same number of response categories but different category label format. Significant χ^2 statistics indicate that the correlations differ among the 4, 5, and 7 response lengths for the particular category label format.

†$p < .10$.

Table 3.10. Correlations between First Set of Questionnaires and the International Personality Item Pool Neuroticism Measure

	BFI-E	BFI-N	RSMS	SES	IDAS-D	IDAS-S	Average
ALL							
4	-.40	.91*	-.04	-.73	.81	.73	.68
5	-.33	.89	-.20	-.73	.77	.66	.66
7	-.39	.90	-.21	-.75	.76	.62	.66
χ^2	0.52	.60	2.59	0.15	1.41	3.41	0.15
END							
4	-.49	.85	-.17	-.69	.73	.69	.64
5	-.36	.90	-.30	-.74	.80	.59	.67
7	-.37	.86	-.18	-.79	.74	.76*	.67
χ^2	2.10	3.09	1.51	2.94	1.85	6.42*	0.21

Note. Significant correlations indicate that the correlation is significantly higher than the correlation of the scale with the same number of response categories but different category label format. Significant χ^2 statistics indicate that the correlations differ among the 4, 5, and 7 response lengths for the particular category label format.
*$p < .05$.

The average columns in Table 3.9 and Table 3.10 show the averages of the absolute values of the validity correlations, so they provides tests for a generalizable association between response scale length and validity. The chi-square tests were not significant for either IPIP extraversion or neuroticism. As a second test for a general effect of the number of response categories on validity, I performed two multivariate multiple regressions with BFI-E, BFI-N, RSMS, SES, IDAS-D, and IDAS-S as the dependent variables. For the two regressions, I used either IPIP-E or IPIP-N centered, the number of response categories centered, the response scale label format dummy coded, all three double interactions, and the triple interaction of these three predictors as the independent variables. The results for these two tests are reported in Table 3.11.

The interactions between the IPIP scores and the number of response categories test for the relationship between response scale length and validity. This interaction was not significant for either IPIP-E or IPIP-N, indicating a lack of support for hypothesis four.

Table 3.11. Multivariate Regressions of International Personality Item Pool Measures and Response Scale Qualities Predicting the First Set of Questionnaires

	IPIP-E			IPIP-N		
	Wilk's λ	*p*	η^2	Wilk's λ	*p*	η^2
IPIP	0.36	< .001	.64	0.26	< .001	.74
#RC	0.92	< .001	.08	0.91	< .001	.09
LF	0.92	< .001	.08	0.93	< .001	.08
#RC*LF	0.99	.62	.01	1.00	.81	.01
IPIP*LF	0.99	.26	.01	0.99	.15	.01
IPIP*#RC	0.99	.45	.01	0.99	.14	.02
IPIP*#RC*LF	1.00	.79	.01	0.99	.28	.01

Note. η^2 = partial eta-squared; IPIP-E = International Personality Item Pool extraversion scale; IPIP-N = International Personality Item Pool neuroticism scale; #RC = number of response categories; LF = label format.

To further test hypothesis four, I applied the disattenuation formula (Nunnally & Bernstein, 1994) to each correlation:

$$\rho_{12} = \frac{r_{12}}{\sqrt{\alpha_1}\sqrt{\alpha_2}}.$$

Table 3.12 and Table 3.13 include the validity coefficients for these disattenuated correlations. First, I should note that the disattenuated correlation between BFI-N and IPIP-N (Table 3.13) exceeds one in the four-point ALL condition, so I could not properly test for the differences in average correlations for the ALL condition. None of the three remaining average columns in Table 3.12 or Table 3.13 were statistically significant. I also ran the aforementioned multivariate multiple regressions with the *T*-transformed reliabilities for the IPIP scales added as covariates. As seen in Table 3.14, the interactions between the IPIP scores and the number of response categories were not

significant. All of these null effects demonstrate a lack of evidence for any general relationship between the number of response categories and validity as defined by the average associations of six questionnaires with IPIP extraversion and neuroticism scales.

Table 3.12. Disattenuated Correlations between First Set of Questionnaires and the International Personality Item Pool Extraversion Measure

	BFI-E	BFI-N	RSMS	SES	IDAS-D	IDAS-S	Average
ALL							
4	.93	-.57	.30	.58	-.58	-.59	.64
5	.93	-.46	.50	.43	-.42	-.57	.61
7	.96***	-.44	.53	.52	-.35	-.52	.64
χ^2	7.05*	2.55	6.05*	2.73	6.35*	0.76	0.25
END							
4	.94	-.51	.36	.63	-.53	-.68	.66
5	.95	-.53	.40	.57	-.54	-.70†	.68
7	.91	-.57	.37	.65	-.45	-.58	.63
χ^2	6.87*	0.47	0.15	1.13	1.03	2.92	0.45

Note. Significant correlations indicate that the correlation is significantly higher than the correlation of the scale with the same number of response categories but different category label format. Significant χ^2 statistics indicate that the correlations differ among the 4, 5, and 7 response lengths for the particular category label format.

$^{\dagger}p < .10$. $^{*}p < .05$. $^{***}p < .001$.

Table 3.13. Disattenuated Correlations between First Set of Questionnaires and the International Personality Item Pool Neuroticism Measure

	BFI-E	BFI-N	RSMS	SES	IDAS-D	IDAS-S	Average
ALL							
4	-.44	1.01***[a]	-.04	-.79	.87†	.80	***[a]
5	-.36	.98	-.22	-.79	.83	.72	.77
7	-.43	.99***	-.23	-.82	.82	.68	.81
χ^2	0.72	***[a]	3.15	0.57	2.85	5.35†	***[a]
END							
4	-.56	.96	-.19	-.75	.80	.77	.75
5	-.39	.99	-.33	-.79	.86	.65	.79
7	-.41	.95	-.20	-.86	.80	.83**	.76
χ^2	3.60	33.46***	1.74	5.50†	2.88	11.05**	0.50

Note. Significant correlations indicate that the correlation is significantly higher than the correlation of the scale with the same number of response categories but different category label format. Significant χ^2 statistics indicate that the correlations differ among the 4, 5, and 7 response lengths for the particular category label format.

†$p < .10$. **$p < .01$. *** $p < .001$.

[a]Because the disattenuated correlation exceeded one, it is impossible to compare the statistic against the correlation of the scale with the same number of response categories but different label format, to compute the row average correlation, or to compute the χ^2 test for independent correlations.

Table 3.14. Multivariate Multiple Regressions of International Personality Item Pool Questionnaires, Response Scale Qualities, and Reliability Covariate Predicting the First Set of Questionnaires

	IPIP-E			IPIP-N		
	Wilk's λ	*p*	η^2	Wilk's λ	*p*	η^2
IPIP	0.36	< .001	.64	0.26	< .001	.74
#RC	0.95	< .001	.06	0.93	< .001	.07
LF	0.92	< .001	.08	0.98	.01	.03
#RC*LF	0.99	.66	.01	0.99	.44	.01
IPIP*LF	0.99	.26	.01	0.99	.18	.01
IPIP*#RC	0.99	.45	.01	0.99	.14	.02
IPIP*#RC*LF	1.00	.79	.01	0.99	.28	.01
T-IPIP	0.98	.05	.02	0.99	.30	.01

Note. η^2 = partial eta-squared; IPIP-E = International Personality Item Pool extraversion scale; IPIP-N = International Personality Item Pool neuroticism scale; #RC = number of response categories; LF = label format; *T*-IPIP = *T*-transformed alpha for International Personality Item Pool scale.

As a more direct test of hypothesis four, I explored the associations between the BFI and IPIP measures of the same constructs (extraversion and neuroticism) as a function of the number of response categories. The data in Table 3.9 and Table 3.10 do not reveal any significant relationships between response scale length and correlation size for extraversion or neuroticism, respectively. However, the disattenuated correlations in Table 3.12 and Table 3.13 show that all four tests for relationships between scale length and validity were significant. For the END condition, the five response category condition was optimal for both extraversion and neuroticism. The findings for the ALL condition are less clear. For extraversion, seven categories was optimal, but for neuroticism, the four category condition had the highest correlation. The influence of the response scale label format will be described more thoroughly in the next section. For now, it is important to note that—although the data do not give any evidence

that there is a generalizable relationship between response scale length and validity—they do reveal a relationship between the response scale length and concurrent validity for two measures of the same constructs, extraversion and neuroticism. However, the data do not strongly support the hypothesis that this relationship is positive and monotonic.

3.6 RQ2 and RQ3: Moderators of Response Scale Length and Validity

The second research question asks whether there will be an interaction between the number of response categories and the category label format in predicting validity. The average columns in Table 3.9 and Table 3.10 for the correlations with the IPIP extraversion and neuroticism scales, respectively, do not show an interaction effect: There were no statistical differences between any of the correlations and none of the chi-square tests were significant. Moreover, the results reported in Table 3.11 indicate that the triple interactions between the response scale length, label format, and IPIP score were not significant for the multivariate multiple regressions for either extraversion or neuroticism. The disattenuated correlations in Table 3.12 and Table 3.13 show a similar null effect: The only significant results were for the average ALL column in Table 3.13, but the statistics for this column could not be appropriately tested because the four-point, ALL BFI-N disattenuated correlation exceeded one. Lastly, the results in Table 3.14 reveal null effects for the multivariate triple interactions that have accounted for the reliability of the IPIP scale. Thus, there is no evidence of an interaction between the number of response categories and the category label format in predicting general construct validity.

Table 3.15. Correlations between Big Five Inventory and International Personality Item Pool Extraversion and Neuroticism Measures by Education Level

	IPIP-E				IPIP-N			
	Uneducated		Educated		Uneducated		Educated	
	BFI-E	BFI-N	BFI-E	BFI-N	BFI-E	BFI-N	BFI-E	BFI-N
ALL								
4	.84	-.55	.84	-.47	-.40	.94*	-.38	.87
5	.87	-.46	.83	-.38	-.41	.89	-.26	.90
7	.86	-.32	.86	-.49	-.42	.91	-.35	.88
χ^2	0.41	2.75	0.34	0.77	0.01	3.92	0.68	0.52
END								
4	.79	-.39	.88	-.57	-.45	.88	-.57	.78
5	.87	-.50	.87	-.44	-.37	.92	-.33	.86
7	.81	-.60*	.84	-.43	-.45	.87	-.29	.86
χ^2	2.16	2.56	0.79	1.26	0.35	3.10	4.03	1.66

Note. Significant correlations indicate that the correlation is significantly higher than the correlation of the scale with the same number of response categories but different category label format. Significant χ^2 statistics indicate that the correlations differ among the 4, 5, and 7 response lengths for the particular category label format.

*$p < .05$.

Table 3.16. Disattenuated Correlations between Big Five Inventory and International Personality Item Pool Extraversion and Neuroticism Measures by Education Level

	IPIP-E				IPIP-N			
	Uneducated		Educated		Uneducated		Educated	
	BFI-E	BFI-N	BFI-E	BFI-N	BFI-E	BFI-N	BFI-E	BFI-N
ALL								
4	.93	-.61	.93	-.53	-.44	1.04***[a]	-.41	.97*
5	.95	-.51	.91	-.42	-.45	.97	-.28	.99***
7	.95*	-.36	.97**	-.55	-.46	1.01***[a]	-.39	.97
χ^2	1.61	4.07	15.01***	1.17	0.03	***[a]	0.90	18.14***
END								
4	.90	-.43	.98***	-.67	-.51	.97	-.66[†]	.93
5	.94	-.54	.98***	-.49	-.40	1.01***[a]	-.37	.95
7	.89	-.65*	.93	-.47	-.49	.94	-.32	.96
χ^2	2.82	2.92	17.08***	2.71	0.63	***[a]	6.55*	1.59

Note. Significant correlations indicate that the correlation is significantly higher than the correlation of the scale with the same number of response categories but different category label format. Significant χ^2 statistics indicate that the correlations differ among the 4, 5, and 7 response lengths for the particular category label format.

[†]$p < .10$. *$p < .05$. **$p < .01$. *** $p < .001$.
[a]Because the disattenuated correlation exceeded one, it is impossible to compare the statistic against the correlation of the scale with the same number of response categories but different label format or to compute the χ^2 test for independent correlations.

As described above, the disattenuated correlations between the BFI and IPIP measures of the same constructs in Table 3.12 and Table 3.13 do show an effect. For both extraversion and neuroticism, the validity coefficients for the seven-point response scales are stronger for the ALL than the END label format. To better understand the relationships, I further separated these analyses by education level. The regular correlations and disattenuated correlations are reported in Table 3.15 and Table 3.16, respectively. Again, the uncorrected correlations in Table 3.15 yielded few significant results, but the disattenuated correlations in Table 3.16 revealed several significant

associations. Although there were no apparent differences between the educated and uneducated groups, some other important patterns did arise. The ALL conditions tended to increase, or at least not substantially decrease, from five to seven response categories, but the END conditions generally obtained optimal validity correlations at five categories. So, the seven-point ALL conditions tended to have significantly stronger validity correlations than the END conditions. To address research question two, there was no interaction between the response scale length and label format in predicting general construct validity, but there was some support for an interaction effect on concurrent validity correlations between measures of the same construct. After accounting for any impact of reliability, the seven-point ALL condition had stronger validity correlations than the END condition, but there was no evidence that the END condition outperformed the ALL condition at four or five categories.

Finally, research question three asks whether total score variability, item homogeneity, or skewness will influence the extent to which scale length predicts validity. To answer this research question, I examined Table 3.9 and Table 3.10 and then Table 3.12 and Table 3.13 for the uncorrected and disattenuated correlations, respectively. Because the only questionnaires to show any consistent relationship between validity and response scale length were the BFI extraversion and neuroticism scales, I only compared these. The BFI extraversion scale had more of a monotonic relationship between response scale length and validity than did the neuroticism scale, at least for the ALL conditions. However, as indicated by the data in Table 3.1 and Table 3.2, these two measures do not particularly differ on score variability, item homogeneity, or skewness. So, the data do not support the conclusion that these descriptive statistics moderated the relationship between response scale length and validity.

Chapter 4

DISCUSSION

The present research rigorously examined potential moderators of the relationships between the number of response categories, response category label format, reliability, and validity. Based on prior research and my own meta-analytic pilot study, I had several hypotheses and research questions. I expected that response scale length would be positively associated with reliability. I predicted that the number of response categories would interact with response scale label format to predict reliability: Response scale length would predict reliability more strongly for fully labeled, rather than endpoint labeled, scales. I hypothesized that relatively low total score variability for shorter scales, item heterogeneity, and greater skewness would magnify the association between response scale length and reliability, but I asked the research question of which variable would most strongly moderate this association. I hypothesized that the number of response categories would positively predict validity. I also raised the research question of whether response scale length and label format would interact to predict validity, and I asked whether score variability, item homogeneity, and skewness would influence the extent to which scale length predicts validity.

I found partial support for the hypothesis that the number of response categories would predict reliability. However, although the number of response categories did interact with category label format, the nature of the interaction was opposite of prediction: Response scale length actually predicted reliability more strongly for endpoint labeled scales. In examining the individual questionnaires for relationship between response scale length and reliability, I discovered that education level moderated this effect. For college educated respondents, there was an improvement in reliability between five and seven response categories but this improvement was not found for uneducated respondents. The data did not support the hypotheses regarding score variability, item heterogeneity, and skewness in general; the relationships differed among

the three predictors. Instead of finding that low score variability at shorter scale length best predicted the relationship, the data revealed that questionnaires that had high score variability across conditions tended to have a positive relationship between scale length and reliability. Also, the questionnaires in which score variability increased with response scale length were associated with the effect. Similarly, the hypothesis about item heterogeneity was not supported, but I did find that, when item homogeneity increased with scale length, reliability tended to improve as well. Skewness did predict the relationship as expected. As for the question of which of these three variables best moderated the relationship, skewness best predicted the relationship as originally hypothesized, but in general, the extent to which item homogeneity increased with the number of response categories best predicted whether reliability improved too.

The various analyses showed some evidence that the number of response categories predicted concurrent validity correlations of the same construct, but they did not show that the relationship between response scale length and validity was positive and monotonic. For the research question regarding an interaction between response scale length and label format, there was some support for this relationship in the concurrent validity correlations. There was no support that score variability, item homogeneity, and skewness predicted the relationship between response scale length and validly.

4.1 Number of Response Categories and Reliability

Based on an abundance of research, I had expected to find that the number of response categories would in general, positively predict reliability across the six questionnaires in the first set of questionnaires. This first hypothesis was not supported in the most literal interpretation because the average reliabilities of the six formats did not significantly differ across the three response length conditions, despite the fact that there was some weak, though non-significant, evidence for an effect.

For the reasons described in the introduction and demonstrated in Appendix A, I attribute part of the lack of a significant effect to the especially high reliabilities for the six questionnaires. The mean of the average inter-item correlations across questionnaires and conditions was .53, and Figure 1 shows that at this level the curvilinear relationship between average inter-item correlation and alpha is starting to flatten out. For these reasons, the average inter-item statistics were more telling for this study, and these did show a clear relationship for the END scales. The itemmetric analyses of item-total correlations provided the strongest evidence in favor of the hypothesis of a general relationship between the number of response alternatives and reliability. Regardless, it is important to recognize that, though significant, the differences in item-total correlations were very small, and according to the post-hoc analyses, the five and seven-point scales did not differ. So, the evidence in favor of a generalizable relationship was weak at best, and examining results for the six questionnaires showed that the reported relationship was driven by only a few of the questionnaires.

This pattern of mixed results mirrors the overall state of the literature. Some authors have suggested that gains in reliability end after five categories (Jenkins & Taber, 1977; Lissitz & Green, 1975), but other authors have espoused using more than five categories (Alwin, 1997; Finn, 1972; Oaster, 1989; Preston & Colman, 2000). Still others have proposed that the optimal number of response categories lies somewhere between four and seven levels (Bandalos & Enders, 1996; Lozano, García-Cueto, & Muñiz, 2008). Thus, it was foreseen that the average differences in reliability between the three response scale lengths would be minimal. Knowing this, the purpose of this research was not to determine whether the number of response categories predicted reliability but to empirically study the circumstances that allow for this effect.

4.2 Qualitative Moderators of Response Scale Length and Reliability

Based on the results of my pilot study, I had expected to find that the number of response categories would interact with category label format, such that ALL scales

would evidence a stronger relationship between scale length and reliability. The results did not support this conclusion at all. In fact, there was slight evidence supporting the opposite conclusion for the alpha coefficients—that END scales had the stronger relationship. But again, because the average inter-item correlations were high, the effect was more apparent in the average inter-item correlations and the item-total correlations. So, the statistical test for this interaction effect was significant for the itemmetric analysis of item-total correlations, but the tests were not significant for the averages of the alpha coefficients.

Because these results were directly opposite of the results from the pilot study, it is clear that neither direction for the interaction effect is generalizable. Thus, some unknown variable likely moderates this interaction. As mentioned above, I believe that this outcome may be an unforeseen consequence of the method used to collect the sample. The participants in this study were recruited from the Amazon Mechanical Turk website with the requirement that their work, on all prior tasks, has been accepted 90% of the time. Most of these workers (65%) had to have completed over 50 tasks. So, these participants could be considered "experts" at taking surveys. At the same time, the subjects were possibly lower in motivation than laboratory workers because it was a web study that may have been routine for them (Paolacci et al., 2010).

These efficient and possibly less motivated participants were perhaps more susceptible to satisficing (Krosnick, 1991) and, therefore, prone to using shortcuts in responding to the items. They may not have always carefully considered the verbal meaning for each response category label, particularly for the seven-point scales. When only given labels for the endpoints, the subjects were presumably at least able to perceive the scale points as being equidistant (Schaeffer & Presser, 2003) and, perhaps, come to a "good enough" answer more efficiently. This explanation, and the results which support it, weighs against Krosnick and Presser's (2010) assumption that respondents must come up with their own verbal labels for numerically labeled categories

before they can select their answers. Instead, it could be that respondents who are optimizing tend to come up with their own verbal labels for the numerically labeled categories. This interpretation requires further testing. In any case, I suspect that subjects would have responded better to the fully labeled seven-point scales if they were more motivated to carefully read and consider each option (Krosnick, 1999).

Moreover, if those subjects in the seven-point, endpoint labeled condition did satisfice to come to a "good enough" response more efficiently, they were sacrificing the quality of their responses in order to complete the survey more quickly, and they may have inflated reliabilities by giving uniform responses to similar items (Krosnick, 1991). If so, it is questionable whether those responses were as valid as they were reliable. That is, the responses to these items may have been particularly consistent but not particularly meaningful.

In unplanned analyses, I discovered that whether the respondents held a college degree moderated the relationship between the number of response categories and reliability. In particular, for the respondents without college degrees there was generally no difference in reliability or average inter-item correlation between five and seven-point response scales. But for college educated respondents, the greatest improvement in reliability and average inter-item correlation generally occurred between five and seven levels. I believe that differences in intellectual ability best explain these results. More educated and intelligent people should be able to make finer distinctions. They may be able to cognitively distinguish between the categories in seven to nine-point response scales (seven plus two; Miller, 1956), whereas less educated people may be best at distinguishing between categories in five to seven-point response scales (seven minus two).

The results also showed that there was an interaction between education level and label format. The two groups did not differ for ALL scales, but the uneducated group had higher item-total correlations for END scales. This finding is interesting because

uneducated or less intelligent people are assumed to be particularly prone to satisficing (Krosnick, 1991). So, the fact that the uneducated group only had higher item-total correlations for END scales is consistent with the view that respondents had artificially inflated reliabilities for these scales due to satisficing behavior.

4.3 Quantitative Moderators of Response Scale Length and Reliability

I had predicted that three quantitative variables—total score variability, item heterogeneity, and skewness—would moderate the relationship between the number of response alternatives and reliability. First, Masters (1974) hypothesized that questionnaires with relatively low total score variability at lesser numbers of response categories would benefit most in reliability by adding more categories. The results did not support this hypothesis. Instead, they showed that the questionnaires that had the highest score variability, regardless of response scale length, and the questionnaires that increased in variability with response scale length tended to have the strongest positive relationship between response scale length and reliability. I noted above that it was not the case that those questionnaires with relatively small score variability at four levels disproportionately increased in score variability with increases in scale length; in fact, the opposite pattern was observed, which is probably why the results were opposite of what Masters' (1974) hypothesized. I posit that a questionnaire's general tendency towards higher total score variability is indicative that people hold strong opinions about the topic and are capable of making fine distinctions in rating the construct. Consequently, it could actually be that high total score variability is a marker for a questionnaire that could benefit from having more response options.

The results also revealed that those questionnaires which had the greatest increase in score variability with response scale length also tended to have a co-occurring increase in reliability. This observation follows directly from the formula for coefficient alpha:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_{i,i}^2}{\sigma_x^2}\right).$$

As score variability (σ_x^2) increases, the fraction on the right decreases, and alpha (α) increases as a result. This predictable finding is not particularly useful in a practical sense.

With regard to the next prediction tested in this study, Komorita and Graham (1965) theorized that the number of response categories should only predict reliability for relatively heterogeneous scales. However, the present results did not support this conclusion either. The only multidimensional questionnaire, the RSMS, did not particularly benefit from increasing response scale length. Perhaps a multidimensional questionnaire is not a good candidate for increasing scale length. For instance, the weak correlations between items intended to measure two conceptually distinct factors should be less affected by increasing scale length, so the average inter-item correlation (and reliability) should be less affected for these questionnaires. I believe that "item heterogeneity" is an unfortunate label for the construct, which is typically operationally defined as the proportion of variance accounted for by the first factor in a factor analysis. This general definition conflates all of the variables that influence the relative size of the first factor. For example, a highly "heterogeneous questionnaire" may have two factors, or it may have one factor with similar but only moderately sized inter-item correlations for each item. Logically, questionnaires with lower average inter-item correlations (and hence reliabilities) require smaller increases in average inter-item correlations—achieved by increasing the number of response options—to increase alpha by a given amount. Therefore, a more appropriate candidate is a questionnaire with inter-item correlations that are similar but small or moderate in size, producing a unidimensional measure with a relatively weak first component or factor. However, this possibility was not shown in this research because all questionnaires, aside from the RSMS, had high average inter-item correlations. So, the assertion that questionnaires with similar but low average inter-item

will have a stronger relationships between response scale length and reliability should be investigated in future research.

I did find evidence that for those questionnaires in which increasing scale length increased the relative size of the first component or factor, reliability also increased. Indeed, this relationship was determined to be the best quantitative moderator in response to research question one. This relationship may simply reflect the manner by which scale length affects reliability: Increasing scale length presumably increases the average inter-item correlation (and alpha), which increases the size of the first component or factor for a unidimensional construct (Cortina, 1993). Thus, the data shows that the extent to which score variability or item homogeneity increases with the number of response categories predicts whether reliability will simultaneously increase; but both of these relationships appear to be consequences of the mechanisms by which increasing scale length improves reliability. Again, the practical utility of this finding is questionable because it is likely a result, instead of a cause, of a positive relationship between response scale length and reliability.

Lastly, based on findings from Monte Carlo simulation studies, I predicted that the positive relationship between the number of response categories and reliability would be stronger for the more skewed questionnaires. There was moderately strong evidence in favor of this conclusion. This finding is actually very intuitive. The majority of people typically only use a fraction of the response scale for questionnaires measuring skewed concepts. For example, psychopathology questionnaires (e.g., IDAS-D and IDAS-S) are usually positively skewed because most people do not have mental illness. Consequently, most people will use the “disagree” half of the response scale for most positively worded items of these measures, so these people only focus their attention on two or three options in a five-point scale. Therefore, adding more response categories to these scales should allow them to make finer distinctions without overwhelming their cognitive capabilities (Alwin, 1992; Krosnick & Presser, 2010). Of the results concerning

the three quantitative variables, this particular finding probably has the most practical usefulness. A questionnaire's skewness can be somewhat predicted by considering the concept being measured, and that information could then be factored in when deciding what number of response categories to use.

I suggest that these moderators—high total score variability, low reliability and inter-item correlations, and high skewness— should be viewed as affecting the probability that increasing the number of response options will increase reliability. That is, these conditions are not sufficient to guarantee a positive relationship between scale length and reliability. Ultimately, before one decides to add response categories to increase reliability, one must have reason to believe that people are able to make fine distinctions regarding levels of the construct, which map fairly well onto the response scale alternatives (Kuncel, 1973; Schaeffer & Presser, 2003).

In this study, the SES was fairly skewed and it had moderate total score variability, but there was no evidence of a relationship between the number of response alternatives and reliability. The SES items are somewhat vague and very general, but the IDAS and BFI-N items are more specific, succinct, and clear. So, people are probably able to meaningfully distinguish between and choose from the seven categories for the IDAS and BFI-N more easily than for the SES. Although this explanation does help to account for the results in the present study, it should be viewed as tentative because the study design did not allow for a rigorous test of this particular possibility. In the future, one could test this hypothesis more directly. A researcher could deliberately select questions that vary in abstractness and clarity, have participants rate the items on those qualities, and assess the moderating impact that these qualities have on the association between response scale length and item-total correlations, using an itemmetric approach.

4.4 Number of Response Categories and Validity

I predicted that there would be a positive association between the number of response alternatives and construct validity. In this study, construct validity was operationally defined as the correlations between scores on the first six questionnaires and scores derived from IPIP-E and IPIP-N items with four, five, and seven response categories. There was no evidence for any moderating effect of response scale length on the validity correlations in general or between any individual questionnaires. Interestingly, there were significant effects after I applied the disattenuation formula to all validity coefficients. The disattenuation formula estimates what the correlation between two constructs would be if the two measures had perfect internal consistency reliability. In other words, it estimates the optimal correlations, controlling for any differences in reliability between conditions.

These analyses yielded significant and meaningful results only for the validity correlations between measures of the same construct: extraversion and neuroticism. The conceptual relationships between the other constructs in the first set of questionnaires with extraversion and neuroticism may not have been strong enough for the response scale formats to affect the validity coefficients. However, the correlations between the BFI and IPIP questionnaires best exemplify construct validity, so conceptually, these correlations represent the most important tests of the hypothesis.

There was very strong evidence that the number of response categories affected the disattenuated correlations between the BFI and IPIP questionnaires. However, the relationship was certainly not positive and monotonic in general. The pattern of results become clearer after considering the interaction between the response scale length and label format. The ALL conditions generally had similar or higher disattenuated correlations, particularly for the seven-point conditions. Unfortunately, further qualifying the analyses by education level did not elucidate the results; it simply provided more evidence that the fully labeled conditions had superior reliabilities at seven response

categories. Thus, there was an interaction effect, but it was not exactly the one that I predicted.

These results are interesting considering that the endpoint labeled scales tended to have the stronger relationship between the number of response categories and reliability. As I suggested previously, it could be that the high reliabilities for those participants in the seven-point, END condition were due to satisficing response behaviors. Cronbach (1950) argued that the gains in reliability obtained by lengthening the response scale are spurious consequences of the response patterns of subjects. This argument seems to have applied to the END conditions. For END scales, the seven-point conditions generally had the best reliabilities but the four or five-point conditions had the highest disattenuated correlations. Thus, perhaps labeling the response categories helps to prevent this problematic phenomenon: The seven-point, ALL condition outperformed the END condition, and it performed well in general. It is plausible that the category labels encourage the respondent to select the response option more carefully. Unfortunately, satisficing behavior appears to have been an issue for this sample, which may mean that the respondents did not always read the labels for the ALL scales. If the respondents were more motivated to optimize, I suspect that the disattenuated correlations for the seven-point ALL condition would have had the highest validity for both extraversion and neuroticism. This possibility could be tested with a typical undergraduate sample.

I also raised the research question of whether score variability, item homogeneity, and skewness influence the extent to which scale length predicts validity. The data did not show that this was the case. The correlations between the extraversion questionnaires were the only ones to evince a monotonic relationship between response scale length and validity, but extraversion did not greatly differ from neuroticism in terms of these three quantitative variables.

4.5 Limitations

This study presents some interesting findings, but as with any study, there were limitations. The greatest limitation of the present study was that the differences in reliability and validity were generally small. This issue raises the question of whether the results have practical significance. I had foreseen that this would be an issue to some extent simply because response scales with four to seven categories are generally thought to be optimal (e.g., Bandalos & Enders, 1996; Lozano, García-Cueto, & Muñiz, 2008). My purpose was to determine what factors might create any differences between these response scale lengths. Even so, if this study were replicated with a different sample, I believe that the differences would have been much larger. As I have discussed, this problem appears to be a consequence of the unusually high coefficient alphas, and the hypothetical example with the RSMS in the results section demonstrates this point. However, given the present results, it is best to interpret the overall pattern of significant (and non-significant) results, and to focus more on the inter-item correlations, which did differ by conditions.

Related to the previous issue, the sample was evidently atypical in some of their response patterns. The greatest evidence for this point is that the results had inflated reliabilities and correlations, even between conceptually distinct constructs. For example, the fact that the (negative) correlations between measures of extraversion and neuroticism were extremely high suggests that satisficing response behavior was a problem in this study. For the BFI and the IPIP, the items were ordered such that every other extraversion item was followed by a neuroticism item, instead of presenting the items for the different subscales separately (see Appendix B). Some respondents may have assumed the extraversion and neuroticism items measured the same construct (Ostrom, Betz, Skowronski, 1992). If they were satisficing, they may then have given fairly uniform alternating responses to the items (Krosnick, 1991), rather than making the effort to carefully consider each item. Additionally, randomly ordering items is thought to

increase task difficulty (Schriesheim, Solomon, & Kopelman, 1989), which may have increased the incidence of satisficing (Krosnick, 1991).

The problem of satisficing appears to have confounded some of the results. Unfortunately, this problem makes it necessary to speculate what would have occurred with a more typical sample. Most importantly, I attributed the surprising findings regarding the interactions between the number of response categories and category label format to satisficing behavior. For the most part, however, the pattern of results makes sense when compared with previous research. Thus, the results should be considered with this important limitation in mind, and I suggest that the more surprising results should be viewed tentatively and replicated.

4.6 Conclusions

Several conclusions may be drawn from the present research. (1) The association between the number of response categories (four, five, or seven) and reliability depended to a large extent on moderators. (2) The present results suggest that, at least for samples that are prone to satisficing, endpoint labeled response scales have the stronger relationship between response scale length and reliability. This may be a spurious consequence of exaggerated response styles, which seem to have been especially problematic for this particular sample of participants. This conclusion requires further testing. (3) For more educated or intelligent samples, there was an advantage to using seven, rather than five categories, but there was no such difference for less educated or intelligent samples. (4) Two characteristics of questionnaires—generally high score variability and greater skewness—increased the probability that response scale length and reliability were positively associated. However, these conditions were not sufficient to guarantee this association. (5) The relationship between the number of response categories and validity depended on the response scale label format. Fully labeled scales were more valid for the seven-point conditions. Aside from that specific finding, the validities of the different formats did not differ in any generalizable manner.

(6) The research was unable to rigorously test whether the other moderators influenced the associations between response scale length and validity. This is an avenue for further research.

The multitude of findings in the present study allow for some general recommendations. I would recommend fully labeling seven-point response scales, but based on the present research, I cannot confidently advocate one label format over the other for four or five-point scales. I would also recommend using fully labeled seven-point scales when measuring constructs that people tend to hold strong opinions about or that represent skewed traits. Lastly, I would recommend using scales with seven fully labeled categories when administering surveys to educated or intelligent groups.

Appendix A

STARTING AVERAGE INTER-ITEM CORRELATION, CHANGE IN AVERAGE INTER-ITEM CORRELATION, AND STANDARDIZED ALPHA

Below, I demonstrate that a given increase in average inter-item correlation will result a greater improvement standardized coefficient alpha at lower starting values of standardized coefficient alpha. Let α_1 be the starting reliability, with average inter-item correlation, $r > 0$, and let α_2 represent the reliability after a given increase in average inter-item correlation, $x > 0$, to the initial average inter-item correlation, r. Finally, let α_1 come from a correlation matrix with k items and elements r_{ij}. Then,

$$\alpha_1 = \left(\frac{k}{k-1}\right)\left(1 - \frac{\sum r_{ii}}{\sum\sum r_{ij}}\right) = \left(\frac{k}{k-1}\right)\left(1 - \frac{k}{(k+k(k-1)r)}\right) = \left(\frac{k}{k-1}\right)\left(\frac{k(k-1)r}{k+k(k-1)r}\right) = \left(\frac{kr}{1+(k-1)r}\right),$$

and

$$\alpha_2 = \left(\frac{k(r+x)}{1+(k-1)(r+x)}\right).$$

The change in standardized alpha, for a given increase in average inter-item correlation, is:

$$\Delta\alpha = \alpha_2 - \alpha_1 = \left(\frac{k(r+x)}{1+(k-1)(r+x)}\right) - \left(\frac{kr}{1+(k-1)r}\right).$$

Differentiating Δα with respect to r, I get:

$$\frac{d\Delta\alpha}{dr} = \frac{k}{[1+(k-1)(r+x)]^2} - \frac{k}{[1+(k-1)r]^2} =$$

$$\frac{d\Delta\alpha}{dr} = \frac{k}{[1+(k-1)(r+x)]^2[1+(k-1)r]^2}[[1+(k-1)r]^2 - [1+(k-1)(r+x)]^2] =$$

$$\frac{d\Delta\alpha}{dr} = \frac{-k(k-1)}{[1+(k-1)(r+x)]^2[1+(k-1)r]^2}[2x + 2(k-1)rx + (k-1)x^2] < 0.$$

In the last equation, the left part is always negative for $k > 1$, and the right part is always positive.

Thus, for any given increase in average inter-item correlation, x, the resulting improvement in standardized alpha is a decreasing function of the starting average inter-item correlation, r, and therefore the starting reliability, α_1. This point is demonstrated in Figure 1. For any number of items, k, alpha asymptotically approaches one as the average inter-item correlation approaches one, and any given change along the average inter-item correlation axis will have a larger resulting change in alpha at lower levels of starting average inter-item correlation. A similar proof can be given for unstandardized

coefficient alpha. This proof would involve changes in average inter-item covariance, and it would require some assumption about the change in the sum of item variances. For example, one could prove that the improvement in coefficient alpha is greater for a given increase in average inter-item covariance at lower levels of average inter-item covariance, holding the sum of item variances constant.

Appendix B

SUMMARY OF SURVEY WORDING

Item 1: I understand and agree to the information on the informed consent.

- Yes
- No

Item 2: What is your highest level of education?

- Some High School
- High School Graduate
- Some College
- Bachelor's Degree
- Master's Degree
- Doctoral Degree

The following statements concern your perception about yourself in a variety of situations. Your task is to indicate the strengths of your agreement with each statement, utilizing the scales below:

Items 3-17 odd: BFI-E

Items 4-18 even: BFI-N

Please read the following statements and indicate the degree to which each statement is true of you. It is important for you to realize that there are no "right" or "wrong" answers to these questions. People are different, and we are interested in how YOU feel.

Items 19-31: RSMS

Items 32-36: Satisfaction With Life Scale (not used in this study; Diener, Emmons, Larsen, & Griffin, 1985)

Items 37-46: RSES

Below is a list of feelings, sensations, problems, and experiences that people sometimes have. Read each item to determine how well it describes your recent feelings and experiences. Then select the option that best how much you have felt or experienced things this way during the past two weeks, including today.

Items 47-51 and 54-58: IDAS-D

Items 52, 53, and 59-61: IDAS-S

On the following pages, there are phrases describing people's behaviors. Please use the rating scales below to describe how accurately each item describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age.

Items 62-70 even: IPIP-E items with four response categories

Items 63-71 odd: IPIP-N items with four response categories

Items 72-80 even: IPIP-E items with five response categories

Items 73-81 odd: IPIP-N items with five response categories

Items 82-90 even: IPIP-E items with seven response categories

Items 83-91 odd: IPIP-N items with seven response categories

Item 92: Which best describes you?

- Male
- Female

Item 93: Please indicate your age in years:

Item 94: Please indicate the ethnicity that best describes you:

- Asian
- African American/Black
- Caucasian/White
- Hispanic/Latino
- Multi-racial or other

Item 95: Enter La followed by four random numbers for this question (e.g., La1234). This will be your survey code. Then, submit this survey, enter the survey code on the MTurk page, and submit the HIT.

References

Alwin, D. F. (1992). Information transmission in the survey interview: number of response categories and the reliability of attitude measurement. *Sociological Methodology, 22*, 83-118.

Alwin, D. F. (1997). Feeling thermometers versus 7-point scales: Which are better? *Sociological Methods & Research, 25*, 318–340.

Alwin, D. F, & Krosnick, J. A. (1991). The reliability of survey attitude measurement: The influence of question and respondent attributes. *Sociological Methods & Research, 20*, 139–181.

Andrews, F. M. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly, 48*, 409–442.

Bandalos, D. L., & Enders, C. K. (1996). The effects of nonnormality and number of response categories on reliability. *Applied Measurement in Education, 9,* 151–160.

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

Bendig, A. W. (1953). The reliability of self-rating scales as a function of the amount of verbal anchoring and the number of categories on the scale. *Journal of Applied Psychology, 37*, 38–41.

Bernstein, I.H., & Teng, H. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 76*, 186–204.

Bossom, J. K., & Swann, W. B., Jr. (2009). Self-esteem. In Leary, M. R., & Hoyle, R. H. (Eds.), *Handbook of Individual Differences in Social Behavior* (pp. 527–546). New York: Guilford Press.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*, 3–5.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105.

Chang, L. (1994). A Psychometric evaluation of four-point and six-point Likert-type scale in relation to reliability and validity. *Applied Psychological Measurement, 18*, 205–215.

Chiaburu, D. S., Oh, I., Berry, C. M., Li, N., & Gardner, R. G. (2011). The Five-Factor Model of personality traits and organizational citizenship behaviors: A meta-analysis. *Journal of Applied Psychology, 96,* 1140–1166. doi: 10.1037/a0024004

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

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98–104.

Cox, E. P., III (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*, 407–422.

Crandall, R. (1973). The measurement of self-esteem and related constructs. In J. Robinson & P. Shaver (Eds.), *Measures of social and psychological attitudes* (pp. 45-168). Ann Arbor, MI: Institute for Social Research.

Cronbach, L. J. (1950). Further evidence on response sets and test design. *Educational and Psychological Measurement*, *10*, 3–31.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–224.

Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment, 49,* 71-75.

Dobson, K. S., & Mothersill, K. J. (1979). Equidistant category labels for construction of Likert-type scales. *Perceptual and Motor Skills, 49*, 575–580.

Finn, R. H. (1972). Effects of some variation in rating scale characteristics on the means and reliabilities of ratings. *Educational and Psychological Measurement, 32*, 255–265.

Goddard III, R. D., & Villanova, P. (2006). Designing surveys and questionnaires for research. In F. T. L. Leong & J. T. Austin (Eds.), *The psychology research handbook: A guide for graduate students and research assistants* (2nd ed., pp. 114-124). Thousand Oaks, CA: Sage Publications.

Greer, T., Dunlap, W. P., Hunter, S. T., & Berman, M. E. (2006). Skew and internal consistency. *Journal of Applied Psychology, 91*(6), 1351–1358.

Hakstian, A. R., & Whalen, T. E. (1976). A *k*-sample significance test for independent alpha coefficients. *Psychometrika, 41,* 219–231.

Hays, W. L. (1994). *Statistics* (5th ed.). Orlando, FL: Harcourt Brace.

Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications* (2nd ed.). New York, NY: Routledge.

Huang, C., & Dong, N. (2012). Factor structures of the Rosenberg Self-Esteem Scale. *European Journal of Psychological Assessment, 28*(2), 132–138.

International Personality Item Pool. (2001). A scientific collaboratory for the development of advanced measures of personality traits and other individual differences (Retrieved online May 21, 2014, from: http://ipip.ori.org/).

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

Johnson, D. R., & Creech, J. C. (1983) Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review, 48*, 398–407.

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory—Versions 4a and 5a.* Berkeley: University of California at Berkeley, Institute of Personality and Social Research.

John, O. P., Naumann, L. P., Soto, C. J. (2008). Paradigm shift to the integrative Big Five Trait taxonomy: history, measurement, and conceptual issues. In John, O. P., Robins, R. W., Pervin, L. A. (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 114–158). New York: Guilford Press.

Judge, T. A., & Bono, J. E. (2001). Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology, 86,* 80–92. doi: 10.1037//0021-9010.86.1.80

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25*, 85–96.

Komorita, S. S., & Graham, W. K. (1965). Number of scale points and the reliability of scales. *Educational and Psychological Measurement, 25*, 987–995.

Kotov, R., Gamez, W., Schmidt, F., & Watson, D. (2010). Linking "big" personality traits to anxiety, depressive, and substance use disorders: A meta-analysis. *Psychological Bulletin, 136*, 768–821.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*, 213-236.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology*, *50*, 537-567.

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In J. D. Wright and P. V. Marsden (Eds.), *Handbook of Survey Research* (pp. 263-314). San Diego, CA: Elsevier.

Kuncel, R. B. (1973). The subject-item interaction in itemmetric research. *Educational and Psychological Measurement, 37*, 665-678.

Lennox, R. D., & Wolfe, R. N. (1984). Revision of the Self-Monitoring Scale. *Journal of Personality and Social Psychology, 46*, 1349–1364.

Lissitz, R. W., & Green, S. B. (1975). Effect of the number of scale points on reliability: A Monte Carlo approach. *Journal of Applied Psychology, 60,* 10–13.

Loken, B., Pirie, P., Virnig, K. A., Hinkle, R. L., & Salmon, C. T. (1987). The use of 0–10 scales in telephone surveys. *Journal of the Market Research Society, 29*(3), 353–362.

Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4*(2), 73–79.

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

Matell, M. S., & Jacoby, J. (1971). Is there an optimal number of alternatives for Likert scale items? Study 1: reliability and validity. *Educational and Psychological Measurement, 31*, 657–674.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81-97.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Oaster, T. R. F. (1989). Number of alternatives per choice point and stability of Likert-typescales. *Perceptual and Motor Skills*, *68*, 549-550.

Ostrom, T. M., Betz, A. L., & Skowronski, J. J. (1992). Cognitive representation of bipolar survey items. In N. Schwarz & S. Sudman (Eds.), *Context effects in social and psychological research* (pp. 297–311). New York: Springer-Verlag.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5,* 411-419.

Peters, D. L., & McCormick, E. J. (1966). Comparative reliability of numerically anchored versus job-task anchored rating scales. *Journal of Applied Psychology, 50*, 90–92.

Peterson, R. A. (1994). A meta-analysis of Cronbach's coefficient alpha. *Journal of Marketing Research, 21*, 381–391.

Peterson, R. A., & Kim, Y. (2013). On the relationship between coefficient alpha and composite reliability. *Journal of Applied Psychology, 98*, 194–198.

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.

Rodriguez, M. C., & Maeda, Y. (2006) Meta-analysis of coefficient alpha. *Psychological Methods, 11,* 306–322. doi: 10.1037/1082-989X.11.3.306

Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton Univeristy Press.

Schaeffer, N. C., & Presser, S. (2003). The science of asking questions. *Annual Review of Sociology, 29*, 65–88.

Schriesheim, C. A., Solomon, E., & Kopelman, R. E. (1989). Grouped versus randomized format: An investigation of scale convergent and discriminant validity using LISREL confirmatory factor analysis. *Applied Psychological Measurement, 13*, 19–32.

Schwarz, N., & Strack, F. (1985). Cognitive and affective processes in judgments of subjective well-being: a preliminary model. In H. Brandstatter & E. Kirchler (Eds.), *Economic Psychology* (pp. 439-447). Linz, Austria: Tauner.

Spector, P. E. (1976). Choosing response categories for summated rating scales. *Journal of Applied Psychology, 61,* 374–375.

Tourangeau, R., & Rasinski, K. A. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin*, *3*, 299-314. doi: 10.1037//0033-2909.103.3.299

Trapmann, S., Hell, B., Hirn, J. W., & Schular, H. (2007). Meta-analysis of the Big Five and academic success at university. *Zeitschrift für Psychologie, 215*(2), 132–151.

Wakita, T. Ueshima, N. & Noguchi, H. (2012). Psychological distance between categories in the Likert scale: Comparing different numbers of options. *Educational and Psychological Measurement, 72,* 533–546.

Watson, D., O'Hara, M. W., Simms, L. J., Kotov, R., Chmielewski, M., McDade-Montez, E., et al. (2007). Development and validation of the Inventory of Depression and Anxiety Symptoms (IDAS). *Psychological Assessment, 19,* 253–268.

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.

Weng, L. J. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. *Educational and Psychological Measurement, 64*(6), 956–972.

References for Meta-Analysis

Adams, D, B. (1991). *A model of organizational commitment in staff nurses*. Unpublished doctoral dissertation, University of San Diego, CA.

Allen, T. D., Facteau, J. D., & Facteau, C. L. (2004). Structured interviewing for OCB: Construct validity, faking, and the effects of question type. *Human Performance, 17*, 1-24.

Avis, J. M. (2001). *An examination of the prediction of overall, task, and contextual performance using three selection measures for a servicetype occupation* (Unpublished doctoral dissertation). The University of Southern Mississippi, Hattiesburg, MS.

Avis, J. M., Kudisch, J. D., & Fortunato, V. J. (2002). Examining the incremental validity and adverse impact of cognitive ability and conscientiousness on job performance. *Journal of Business and Psychology, 17*, 87-105.

Baer, M. (2010). The strength-of-weak-ties perspective on creativity: A comprehensive examination and extension. *Journal of Applied Psychology, 95*, 592-601.

Baer, M., & Oldham, G. R. (2006). The curvilinear relation between experienced creative time pressure and creativity: Moderating effects of openness to experience and support for creativity. *Journal of Applied Psychology, 91*, 963-970.

Bagozzi, R. P. (1978). Salesforce performance and satisfaction as a function of individual difference, interpersonal, and situational factors. *Journal of Marketing Research, 15*, 517-531.

Barchard, K.A. (2003). Does emotional intelligence assist in the prediction of academic success? *Educational and Psychological Measurement, 63*, 840-858.

Barrick, M. R., & Mount, M. K. (1993). Autonomy as a moderator of the relationships between the big five personality dimensions and job performance. J*ournal of Applied Psychology, 78*, 111-118.

Barrick, M. R., & Mount, M. K. (1996). Effects of impression management and self-deception on the predictive validity of personality constructs. *Journal of Applied Psychology, 81*, 261-272.

Batlis, N. C. (1980). Job involvement and locus of control as moderators of role-perception/individual-outcome relationships. *Psychological Reports, 46*, 111-119.

Bauer, K.W., & Liang, Q. (2003). The effect of personality and precollege characteristics on first-year activities and academic performance. *Journal of College Student Development, 44*, 277-290.

Berhman, D. N., & Perreault, W. D. (1984). A role stress model of the performance and satisfaction of industrial salespersons. *Journal of Marketing, 48*, 9-21.

Bhagat, R. S., & Allie, S. M. (1989). Organizational stress, personal life stress, and symptoms of life strains: An examination of the moderating role of sense of competence. *Journal of Vocational Behavior, 35*, 231-253.

Blakely, G. L., Andrews, M. C., & Fuller, J. (2003). Are chameleons good citizens? A longitudinal study of the relationship between selfmonitoring and organizational citizenship behavior. *Journal of Business and Psychology, 18,* 131-144.

Blau, G. (1987). Locus of control as a potential moderator of the turnover process. *Journal of Occupational Psychology, 60*, 21-29.

Blickle, G., Momm, T., Schneider, P., Gansen, D., & Kramer, J. (2009). Does acquisitive self-presentation in personality self-ratings enhance validity? Evidence from two experimental field studies. *International Journal of Selection and Assessment, 17*, 142-153.

Boudreau, J. W., Boswell, W. R. & Judge, T. A. (2001). Effects of personality on executive career success in the United States and Europe. *Journal of Vocational Behavior, 58*, 53-81

Brookings, J. B., Bolton, B., Brown, C. E., & McEvoy, A. (1985). Self-reported job burnout among female human service professionals. *Journal of Occupational Behaviour, 6*, 143-150.

Brown, J., Cooper, G., & Kirkcaldy, B. (1996). Occupational stress among senior police officers. *British Journal of Psychology, 78*, 31-41.

Cellar, D. F., DeGrendel, D. J. D., Klawsky, J. D., & Miller, M. L. (1996). The validity of personality, service orientation, and reading comprehension measures as predictors of flight attendant training performance. *Journal of Business and Psychology, 11*, 43-54.

Cellar, D. F., Miller, M. J., Doverspike, D. D., & Klawsky, J. D. (1996). Comparison of factor structures and criterion-related validity coefficients for two measures of personality based on the five factor model. *Journal of Applied Psychology, 81*, 694-703.

Chan, D., & Schmitt, N. (2002). Situational judgment and job performance. *Human Performance, 15*, 233-254.

Chandler, C. W. (2004). *Understanding organizational citizenship behaviors: Do motives make a difference?* (Unpublished doctoral dissertation). George Mason University, Fairfax, VA.

Coˆte´, S., & Miners, C. T. H. (2006). Emotional intelligence, cognitive intelligence, and job performance. *Administrative Science Quarterly, 51*, 1-28.

Colquitt, J.A., & Simmering, M.J. (1998). Conscientiousness, goal orientation, and motivation to learn during the learning process: A longitudinal study. *Journal of Applied Psychology, 83*, 654-665.

Cooper, C. L., & Williams, J. (1991). A validation of the OSI on a blue-collar sample. *Stress Medicine, 7*, 109-112.

Cortina, J. M., Doherty, M. L., Schmitt, N., Kaufman, G., & Smith, R. (1992). The "Big Five" personality factors in the IPI and MMPI: Predictors of police performance. *Personnel Psychology, 45*, 119-140.

Crant, J. M. (1995). The Proactive Personality Scale and objective job performance among real estate agents. *Journal of Applied Psychology, 80*, 532-537.

Deluga, R. J. (1998). Leader-member exchange quality and effectiveness ratings: The role of subordinate-supervisor conscientiousness similarity. *Group & Organization Management, 23*, 189 -216.

Detert, J. R., & Burris, E. R. (2007). Leadership behavior and employee voice: Is the door really open? *Academy of Management Journal, 50*, 869-884.

Dewett, T. (2002). *Differentiating outcomes in employee creativity: Understanding the role of risk in creative performance* (Unpublished doctoral dissertation). Texas A&M University, College Station, TX.

Diefendorff, J. M., Brown, D. J., Kamin, A. M., & Lord, R. G. (2002). Examining the roles of job involvement and work centrality in predicting organizational citizenship behaviors and job performance. *Journal of Organizational Behavior, 23*, 93-108.

Diseth, A. (2003). Personality and approaches to learning as predictors of academic achievement. *European Journal of Personality, 17*, 143-155.

Draves, P. R. (2003). *An examination of potential moderating effects of personality on the relationship between job attitudes and organizational citizenship behaviors* (Unpublished doctoral dissertation). University of South Florida, Tampa, FL.

Duff, A., Boyle, E., Dunleavy, K., & Ferguson, J. (2004). The relationship between personality, approach to learning, and academic performance. *Personality and Individual Differences, 36*, 1907-1920.

Fellenz, M. R. (1996). *Individual flexibility in organizations: A conceptual and empirical investigation* (Unpublished doctoral dissertation). The University of North Carolina at Chapel Hill, Chapel Hill, NC.

Ferguson, E., James, D., O'Hehir, F., &Sanders, A. (2003). Pilot study of the roles of personality, references, and personal statements in relation to performance over the 5 years of a medical degree. *British Medical Journal, 326*, 429-431.

Ferris, G. R., Witt, L. A., & Hochwarter, W. A. (2001). Interaction of social skill and general mental ability on job performance and salary. *Journal of Applied Psychology, 86*, 1075-1082.

Galperin, B., & Burke, R. J. (2006). Uncovering the relationship between workaholism and workplace destructive and constructive deviance: An exploratory study. International *Journal of Human Resource Management, 17*, 331-347.

Gebbia, M. (1999). *Transforming the work environment: Do norms influence organizational citizenship behavior?* (Unpublished doctoral dissertation). The City University of New York, New York, NY.

Gellatly, I. R., & Irving, P. G. (2001). Personality, autonomy, and contextual performance of managers. *Human Performance, 14*, 231-245.

George, J. M., & Zhou, J. (2001). When openness to experience and conscientiousness are related to creative behavior: An interactional approach. *Journal of Applied Psychology, 86*, 513-524.

Goff, M., & Ackerman, P.L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology, 84*, 537-552.

Goh, S. C., & Mealiea, L. W. (1984). Fear of success and its relationship to the job performance, tenure, and desired job outcomes of women. *Canadian Journal of Behavioural Science, 16*, 65-75.

Gorini, H. M. (1991). *An individual's correspondence preference as related to work related complaints, neuroticism and job satisfaction* Unpublished doctoral dissertation, University of Illinois, Urbana.

Grandmaison, L. J. (2006). *Assessing the incremental validity of personality on direct leadership in the Canadian Forces* (Unpublished doctoral dissertation). Carleton University, Ottawa, Ontario, Canada.

Grant, A. M., & Berry, J. (2011). The necessity of others is the mother of invention: Intrinsic and prosocial motivations, perspective-taking, and creativity. *Academy of Management Journal, 54*, 73-96.

Grant, A. M., & Wrzesniewski, A. (2010). I won't let you down . . . or will I? Core self-evaluations, other-orientation, anticipated guilt and gratitude, and job performance. *Journal of Applied Psychology, 95*, 108-121.

Gray, E.K., & Watson, D. (2002). General and specific traits of personality and their relation to sleep and academic performance. *Journal of Personality, 70*, 177-206.

Greenberger, D. B., Strasser, S., Cummings, L. L., & Dunham, R. B. (1989). The impact of personal control on performance and satisfaction. *Organizational Behavior and Human Decision Processes, 43*, 29-51.

Greguras, G. J., & Diefendorff, J. M. (2010). Why does proactive personality predict employee life satisfaction and work behaviors? A field investigation of the mediating role of the self-concordance model. *Personnel Psychology, 63*, 539-560.

Griffin, B., & Hesketh, B. (2003). Adaptable behaviours for successful work and career adjustment. *Australian Journal of Psychology, 55*, 65-73.

Griffin, B., & Hesketh, B. (2005). Are conscientious workers adaptable? *Australian Journal of Management, 30*, 245-259.

Gutkowski, J. M. (1997). *Investigating a three component model of job performance: A construct-oriented approach* (Unpublished doctoral dissertation). University of Houston, Houston, TX.

Hajnal, V. (1991). *The pay satisfaction and efficacy of educators: A multivariate analysis*. (Unpublished doctoral dissertation). University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

Halbesleben, J. R. B., Harvey, J., & Bolino, M. (2009). Too engaged? A conservation of resources view of the relationship between work engagement and work interference with family. *Journal of Applied Psychology, 94*, 1452-1465.

Han, T. (2003). Multilevel approach to individual and team adaptive performance (Unpublished doctoral dissertation). The University at Albany, State University of New York, Albany, NY.

Hassell, B. L. (1991). *The effects of ageism and age discrimination on older workers: A field study.* (Unpublished doctoral dissertation). Florida State University, Tallahassee.

Hattrup, K., O'Connell, M. S., & Wingate, P. H. (1998). Prediction of multidimensional criteria: Distinguishing task and contextual performance. *Human Performance, 11*, 305-319.

Hense, R. L. (2001). *The Big Five and contextual performance: Expanding person-environment fit theory* (Unpublished doctoral dissertation). University of South Florida, Tampa, FL.

Janssen, O., & Van Yperen, N. W. (2004). Employees' goal orientations, the quality of leader-member exchange, and the outcomes of job performance and job satisfaction. *Academy of Management Journal, 47*, 368-384.

Jiang, C., Wang, D., & Zhou, F. (2009). Personality traits and job performance in local government organizations in China. *Social Behavior and Personality, 37*, 451-457.

Johnson, A. L., Luthans, F., & Hennessey, H. W. (1984). The role of locus of control in leader influence behavior. *Personnel Psychology, 37*, 61-75.

Judge, T. A., Bono, J. E., & Locke, E. A. (2000). Personality and job satisfaction: The mediating role of job characteristics. *Journal of Applied Psychology, 85*, 751-765.

Judge, T. A., LePine, J. A., & Rich, B. L. (2006). Loving yourself abundantly: Relationship of the narcissistic personality to self- and other perceptions of workplace deviance, leadership, and task and contextual performance. *Journal of Applied Psychology, 91*, 762-776.

Kaldenberg, D. O. (1991). Test of the Korman hypothesis: Performance, self-esteem, and job satisfaction among dentists. *Psychological Reports, 69*, 201-202.

Keller, R. T. (1983). Predicting absenteeism from prior absenteeism, attitudinal factors, and non-attitudinal factors. *Journal of Applied Psychology, 68*, 536-540.

Keller, R. T. (1987). Cross cultural influences on work and nonwork contributors to quality of life. *Group and Organization Studies, 12*, 304-318.

Keller-Glaze, H. (2001). *Organizational citizenship behavior as impression management* (Unpublished doctoral dissertation). Central Michigan University, Mount Pleasant, MI.

Kemmerer, B. E. (1990). *The moderating effect of personality differences on job stress: A longitudinal investigation* (Unpublished doctoral dissertation). University of Nebraska, Lincoln.

King, E. B., George, J. M., & Hebl, M. R. (2005). Linking personality to helping behaviors at work: An interactional perspective. *Journal of Personality, 73*, 585-608.

Koeske, G. F., & Kelly, T. (1995). The impact of overinvolvement on burnout and job satisfaction. *American Journal of Orthopsychiatry, 65*, 282-292

Konovsky, M. A., & Organ, D. W. (1996). Dispositional and contextual determinants of organizational citizenship behavior. *Journal of Organizational Behavior, 17*, 253-266.

Kraus, E. (2002). *Personality and job performance: The mediating roles of leader-member exchange quality and action control* (Unpublished doctoral dissertation). Florida International University, Miami, FL.

Krautheim, M. D. (1997). *The development and validation of a customer service orientation scale for university resident assistants* (Unpublished doctoral dissertation). The University of Tennessee, Knoxville, TN.

Ladd, D., & Henry, R. (2000). Helping coworkers and helping the organization: The role of support perceptions, exchange ideology, and conscientiousness. *Journal of Applied Social Psychology, 30*, 2028-2049.

Landsbergis, P. A., Schnall, P. L., Deitz, D., Friedman, R., & Pickering, T. (1992). The patterning of psychological attributes and distress by "job strain" and social support in a sample of working men. *Journal of Behavioral Medicine, 15*, 379-405.

LaRocco, J. M., & Jones, A. P. (1978). Co-worker and leader support as moderators of stress-strain relationships in work situations. *Journal of Applied Psychology, 63*, 629-634.

Le, H., Oh, I.-S., Robbins, S. B., Ilies, R., Holland, E., & Westrick, P. (2011). Too much of a good thing: Curvilinear relationships between personality traits and job performance. *Journal of Applied Psychology, 96*, 113-133.

Lee, S.-H. (2000). *Cross-cultural validity of personality traits for predicting job performance of Korean engineers* (Unpublished doctoral dissertation). The Ohio State University, Columbus, OH.

Lee, Y. H., Yang, L. S., Wan, K. M., & Chen, G. H. (2010). Interactive effects of personality and friendship networks on contextual performance. *Social Behavior and Personality, 38*, 197-208.

Lehman, E. K., & Simpson, D. D. (1992). Employee substance use and on-the-job behaviors. *Journal of Applied Psychology, 77*, 309-321.

Liao, H. (2002). *A cross-level analysis of organizational citizenship behaviors in work groups* (Unpublished doctoral dissertation). University of Minnesota, Twin Cities, Minneapolis, MN.

Lopez, E. M. (1982). A test of the self-consistency theory of the job performance-job satisfaction relationship. *Academy of Management Journal, 25*, 335-348.

Lounsbury, J.W., Sundstrom, E., Loveland, J.M., & Gibson, L.W. (2003). Intelligence, Big Five personality traits, and work drive as predictors of course grade. *Personality and Individual Differences, 35*, 1231-1239.

Lusch, R. F., & Serpkenci, R. R. (1990). Personal differences, job tension, job outcomes, and store performance: A study of retail store managers. *Journal of Marketing, 54*, 85-101.

Madjar, N. (2008). Emotional and informational support from different sources and employee creativity. *Journal of Occupational and Organizational Psychology, 81*, 83-100.

Majumder, R. K., MacDonald, A. P., & Greever, K. B. (1977). A study of rehabilitation counselors: Locus of control and attitudes toward the poor. *Journal of Counseling Psychology, 24*, 137-141.

Mann, S. L. (2007). *Values as incremental predictors of organizational citizenship behavior* (Unpublished doctoral dissertation). University of Toronto, Toronto, Ontario, Canada.

McIlroy, D., & Bunting, B. (2002). Personality, behavior, and academic achievement: Principles for educators to inculcate and students to model. *Contemporary Educational Psychology, 27*, 326-337.

McManus, M. A., & Kelly, M. L. (1999). Personality measures and biodata: Evidence regarding their incremental predictive value in the life insurance industry. *Personnel Psychology, 52*, 137-148.

Meir, E. I., Melamed, S., & Abu-Freha, A. (1990). Vocational, avocational and skill utilization congruences and their relationships with well-being in two cultures. *Journal of Vocational Behavior, 36*, 153-165.

Moon, H., Kamdar, D., Mayer, D. M., & Takeuchi, R. (2008). Me or we? The role of personality and justice as other-centered antecedents to innovative citizenship behaviors within organizations. *Journal of Applied Psychology, 93*, 84-94.

Moon, S.M., & Illingworth, A.J. (2005). Exploring the dynamic nature of procrastination: A latent growth curve analysis of academic procrastination. *Personality and Individual Differences, 38*, 297-309.

Mossholder, K. W., Bedeian, A. G., & Armenakis, A. A. (1981). Role perceptions, satisfaction, and performance: Moderating effects of self-esteem and organizational level. *Organizational Behavior and Human Performance, 28*, 224-234.

Mount, M. K., Barrick, M. R., & Stewart, G. L. (1998). Five-factor model of personality and performance in jobs involving interpersonal interactions. *Human Performance, 11*, 145-165.

Mount, M. K., Barrick, M. R., & Strauss, J. P. (1994). Validity of observer ratings of the Big Five personality factors. *Journal of Applied Psychology, 79*, 272-280.

Mount, M. K., Oh, I.-S., & Burns, M. (2008). Incremental validity of perceptual speed and accuracy over general mental ability. *Personnel Psychology, 61*, 113-139.

Musgrave-Marquart, D., Bromley, S.P., & Dalley, M.B. (1997). Personality, academic attribution, and substance use as predictors of academic achievement in college students. *Journal of Social Behavior and Personality, 12*, 501-511.

Nelson, A., Cooper, G. L., & Jackson, P. R. (1995). Uncertainty amidst change: The impact of privatization on employee job satisfaction and well-being. *Journal of Occupational and Organizational Psychology, 68*, 57-71.

Neuman, G. A., & Kickul, J. R. (1998). Organizational citizenship behaviors: Achievement orientation and personality. *Journal of Business and Psychology, 13*, 263-279.

Norris, D. R., & Niebuhr, R. E. (1984). Attributional influences on the job performance-job satisfaction relationship. *Academy of Management Journal, 27*, 424-431.

Norris, G. W. (2002). *Using measures of personality and self-efficacy to predict work performance* (Unpublished doctoral dissertation). The Ohio State University, Columbus, OH.

O'Brien, K. E., & Allen, T. D. (2008). The relative importance of correlates of organizational citizenship behavior and counterproductive work behavior using multiple sources of data. *Human Performance, 21*, 62-88.

O'Connell, M. S., Doverspike, D., Norris-Watts, C., & Hattrup, K. (2001). Predictors of organizational citizenship behavior among Mexican retail salespeople. International *Journal of Organizational Analysis, 9*, 272-280.

Oh, I.-S., & Berry, C. M. (2009). The five-factor model of personality and managerial performance: Validity gains through the use of 360 degree performance ratings. *Journal of Applied Psychology, 94*, 1498-1513.

Okun, M.A., & Finch, J.F. (1998). The Big Five personality dimensions and the process of institutional departure. *Contemporary Educational Psychology, 23*, 233-256.

Oswald, F.L., Schmitt, N., Kim, B.H., Ramsay, L.J.,& Gillespie, M.A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89*, 187-207.

Pace, V. L., & Brannick, M. T. (2010). Improving prediction of work performance through frame of reference consistency: Empirical evidence using openness to experience. *International Journal of Selection and Assessment, 18*, 230-235.

Parasuraman, S., & Alutto, J. (1984). Sources and outcomes of stress in organizational settings: Toward the development of a structural model. *Academy of Management Journal, 27*, 330-350.

Peeters, H., & Lievens, F. (2005). Situational judgment tests and their predictiveness of college students' success: The influence of faking. *Educational and Psychological Measurement, 65*, 70-89.

Pierce, J. L., Gardner, D. G., Cummings, L. L., & Dunham, R. B. (1989). Organization-based self-esteem: Construct definition, measurement, and validation. *Academy of Management Journal, 32*, 622-648.

Pierce, J. L., Gardner, D. G., Dunham, R. B., & Cummings, L. L. (1993). Moderation by organization: Self-esteem of role condition-response relationship. *Academy of Management Journal, 36*, 271-288.

Porac, J. F., Ferris, G. R., & Fedor, P. D. (1983). Causal attributions, affect and expectations for a day's work performance. *Academy of Management Journal, 26*, 285-296.

Pulakos, E. D., Schmitt, N., Dorsey, D. W., Arad, S., Borman, W. C., & Hedge, J. W. (2002). Predicting adaptive performance: Further tests of a model of adaptability. *Human Performance, 15*, 299-323.

Radwinsky, R. L. (1999). *The effect of psychological contracts on the performance of temporary employees* (Unpublished doctoral dissertation). The University of Tulsa, Tulsa, OK.

Renn, R. W., & Prien, K. O. (1995). Employee responses to performance feedback from the task. *Group and Organization Management, 20*, 337-354.

Richards, D. A., & Schat, A. C. H. (2011). Attachment at (not to) work: Applying attachment theory to explain individual behavior in organizations. *Journal of Applied Psychology, 96*, 169-182.

Ridgell, S.D., & Lounsbury, J.W. (2004). Predicting academic success: General intelligence, Big Five personality traits, and work drive. *College Student Journal, 38*, 607-619.

Riggs, M. L., & Knight, P. A. (1994). The impact of perceived group success-failure on motivational beliefs and attitudes: A causal model. *Journal of Applied Psychology, 79*, 755-766.

Rogg, K. L. (1997). *Organizational commitment in the post-loyal era: Perceived organizational support, multiple commitments, and other antecedents' effects on turnover intentions and job performance* (Unpublished doctoral dissertation). Kansas State University, Manhattan, KS.

Saks, A. M., & Ashforth, B. E. (1997). A longitudinal investigation of the relationships between job information sources, applicant perceptions of fit, and work outcomes. *Personnel Psychology, 50*, 395-426.

Schmidt, J. (2008). *Personality, group context, and performance behaviors in university football teams* (Unpublished doctoral dissertation). University of Calgary, Calgary, Alberta, Canada.

Schmitt, N., & Bedeian, A. G. (1982). A comparison of LISREL and two-stage least squares analysis of a hypothesized life-job satisfaction reciprocal relationship. *Journal of Applied Psychology, 67*, 806-817.

Schwoerer, C. E., & May, D. R. (1996). Age and work outcomes: The moderating effects of self-efficacy and tool design effectiveness. *Journal of Organizational Behavior, 17*, 469-487.

Sears, G. (2005). *The dispositional antecedents of leader-member exchange and organizational citizenship behaviour: A process perspective* (Unpublished doctoral dissertation). McMaster University, Hamilton, Ontario, Canada.

Shahani, C., Dipboye, R. L., & Phillips, A. P. (1990). Global self-esteem as a correlate of work-related attitudes: A question of dimensionality. *Journal of Personality Assessment, 54*, 276-288.

Smith, C. A., Organ, D. W., & Near, J. P. (1983). Organizational citizenship behavior: Its nature and antecedents. *Journal of Applied Psychology, 68*, 653-663.

Steffensmeier, J. (2008). *Situational constraints and personality as antecedents of organizational citizenship behaviors* (Unpublished doctoral dissertation). Clemson University, Clemson, SC.

Stewart, G. L., & Carson, K. P. (1995). Personality dimensions and domains of service performance: A field investigation. *Journal of Business and Psychology, 9*, 365-378.

Taylor, S., Kluemper, D., & Mossholder, K. (2010). Linking personality to interpersonal citizenship behaviour: The moderating effect of empathy. *Journal of Occupational and Organizational Psychology, 83*, 815-834.

Tharenou, P., & Marker, P. (1984). Moderating influence of self-esteem on relationships between job complexity, performance, and satisfaction. *Journal of Applied Psychology, 69*, 623-632.

Tokar, D. M., & Subich, L. M. (1997). Relative contributions of congruence and personality dimensions to job satisfaction. *Journal of Vocational Behavior, 50*, 482-491.

Van Dyne, L., Graham, J. W., & Dienesch, R. M. (1994). Organizational citizenship behavior: Construct redefinition, measurement, and validation. *Academy of Management Journal, 37*, 765-802.

Venkataramani, V., & Dalal, R. S. (2007). Who helps and harms whom? Relational antecedents of interpersonal helping and harming in organizations. *Journal of Applied Psychology, 92*, 952-966.

Vigoda, E. (2001). Reactions to organizational politics: A cross-cultural examination in Israel and Britain. *Human Relations, 54*, 1483-1518.

White, A. T., & Spector, P. E. (1987). An investigation of age-related factors in the age-job satisfaction relationship. *Psychology and Aging, 2*, 261-265.

Williams, S. (2004). Personality, attitude, and leader influences on divergent thinking and creativity in organizations. *European Journal of Innovation Management, 7*, 187-204.

Wolfe, R.N., & Johnson, S.D. (1995). Personality as a predictor of college performance. *Educational & Psychological Measurement, 55*, 177-185.

Zhou, J., & George, J. M. (2001). When job dissatisfaction leads to creativity: Encouraging the expression of voice. *Academy of Management Journal, 44*, 682-696.

Biographical Information

Tyler Hamby earned his Bachelor's degree in Mathematics in May 2008 from Lubbock Christian University. Then, he studied Pure Mathematics at Texas Tech University and graduated with a Master's degree in May 2010. Since starting his graduate education in psychology at the University of Texas at Arlington, his research has focused primarily on survey methodology and psychometrics. After completing his Ph.D. in Psychological Sciences, he plans to work as a Survey Statistician or Psychometrician, preferably in a research setting.