12-1-2014

Validation of a Personality Scale: A Factor Analytic Approach

Leslie Forbush
University of Nevada, Las Vegas

VALIDATION OF A PERSONALITY SCALE:

A FACTOR ANALYTIC APPROACH

by

Leslie Forbush

Bachelor of Arts in Psychology
University of Nevada, Las Vegas
2008

Master of Science in Educational Psychology
University of Nevada, Las Vegas
2009

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy – Educational Psychology

Department of Educational Psychology & Higher Education
College of Education
The Graduate College

University of Nevada, Las Vegas
December 2014

UNLV | GRADUATE COLLEGE

We recommend the dissertation prepared under our supervision by

Leslie Forbush

entitled

Validation of a Personality Scale: A Factor Analytic Approach

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy - Educational Psychology
Department of Educational Psychology and Higher Education

Tara C. Raines, Ph.D., Committee Chair

Scott A. Loe, Ph.D., Committee Member

Joe Crank, Ph.D., Committee Member

Gene Hall, Ph.D., Graduate College Representative

Kathryn Hausbeck Korgan, Ph.D., Interim Dean of the Graduate College

December 2014

ABSTRACT

Validation of a Personality Scale: A factor analytic approach

by

Leslie Forbush

Dr. Tara C. Raines, Examination Committee Chair
Assistant Professor of Educational Psychology and Higher Education
University of Nevada, Las Vegas

Researchers and clinicians alike make use of various assessment instruments to investigate variables of interest. These instruments include affective assessments, which require individuals to make judgments about themselves. To make these judgments individuals reflect upon their attitudes, interests, values, and personality traits and choose the corresponding response option that they feel best represents their current status with regards to the variable of interest for each question. The optimal number of response options to be included in such affective measurement scales has been greatly debated over the past several decades (Cox, 1980; Foddy, 1993; Jacoby & Matell, 1971; Jones & Loe, 2013; Maydeu-Olivares, et al., 2009).

The purpose of this study is to investigate the factor structure and the impact of altering the number of response options available to an individual on that factor structure for the 50-item scale from the International Personality Item Pool (IPIP-50, http://ipip.ori.org/ipip/). Two de-identified datasets drawn from a collection of approximately 1000 undergraduate and graduate students are used in this analysis. Two separate, but related, investigations are presented using these data. The first is a study wherein the factor structure of the IPIP-50 is investigated. The second study is an investigation using data wherein individuals were randomly assigned to complete the

IPIP-50 when five response options were presented or when two response options were presented. The factor structure of the IPIP-50 is again investigated, looking for differences between the two conditions.

The effects on the factor structure of the IPIP-50 are examined and the hypothesis that the factor structure supports the Five Factor Model of personality in all conditions is tested. The hypotheses were found to be largely supported, and the implications of this for the field of school psychology and within the realm of psychological assessment are discussed.

ACKNOWLEDGEMENTS

My graduate experiences have tended to be characterized by unexpected events and unusual circumstances. When asked about my experience in a certain class or with a particular project my response all too frequently begins with the word "Well…" followed by some sort of qualifying statement as to why my experience may not have been typical. This dissertation process has been no exception and as a result I have many people to acknowledge and thank for their help throughout the process. To begin, none of this could have happened without the support of my committee members. They were flexible and supportive in the face of tragedy, and for that I am very grateful.

I would also like to acknowledge my husband, Eric Forbush, and the role that he has played in pushing me to finish. When I was tired from a long day at work and just wanted to sit on the couch with our puppy instead of writing, he would always encourage me to keep working and to push through. He was there for the hours that I spent working with my data and was patient with me as I neglected him and the dog to write. Without his encouragement I might still be on that couch with this document sitting on my computer only partially completed. He never doubted me even in the moments when I most doubted myself.

There are two other people without whom I would never have gotten through graduate school. The first is Marty Koch in the Educational Psychology and Higher Education department. From day one she has proven her dedication to the students in the department. Whenever I needed a question answered or had trouble finding a form that I needed, she was right there with the answer. Her quick responses and dedication have saved me on multiple occasions. The second person is my twin, Kristen Bjork. Six and a

half years later and I am still not convinced that everyone in the department is able to tell us apart. Our all-night neuroanatomy study parties and afternoon writing sessions at Krispy Kreme made it possible for me to maintain at least a portion of my sanity over the past several years. I did almost have a seizure while we were writing our meta-analyses, but that is a story for a different day. Between our class potluck during my presentation on eating disorders and our cyber-counseling demonstrations, we always kept Dr. Jones on his toes.

Speaking of Dr. Jones, I want to formally acknowledge my genuine gratitude for the support, guidance, and mentorship that he provided for me over the past six years. He saw potential in me when I was just starting out that I had no idea was there. Thanks to him, I have mastered the art of the BDA mental status exam. I only hope that I can someday live up to the expectations that he had for me. I could not have accomplished this without him, and I want him to know that I am almost ready to forgive him for abandoning me. Almost. He will be deeply and sincerely missed.

There are many other people that I should have mentioned here, but I have already written a novel and my words are coming in short supply. If I did not include you by name, please do not think that I do not love you. I love and appreciate everyone who has contributed to my accomplishments these past six and a half years. My family members, of course, have been a tremendous source of support and encouragement; to name them all would require too many pages. To any of my friends and family who may read this, thank you for your belief and support. I will not be at all offended if you stop reading now.

DEDICATION

To my parents:

Thank you for always supporting me.

To my husband:

Thank you for carrying me through to the end.

To Mildred:

Thank you for not eating my dissertation.

I love you all.

In memory of Dr. W. Paul Jones.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Purpose of Study

The purpose of the following study is to investigate the factor structure of the

fifty-item scale of the International Personality Item Pool (IPIP-50,

http://ipip.ori.org/ipip/). Greater details pertaining to the IPIP-50, as well as the

importance of factor analytic approaches to the validation of assessment scales will be

introduced in the chapters that follow. Specifically, the next chapter will detail the

theoretical foundation of the IPIP-50 and provide a summary of the empirical research

regarding factor analytic validation approaches as well as review the research regarding

the impact of reducing the number of available response options on affective

measurement scales.

Statement of the Problem

Sources of Validity Evidence

Researchers and clinicians alike make use of various assessment instruments to

investigate variables of interest. These assessment instruments attempt to measure such

factors as cognitive ability, academic achievement, and affect, amongst others. Cognitive

and academic measures typically require the respondent to complete some sort of task,

such as solving arithmetic equations or choosing an item to complete a pattern. Affective

assessments generally require individuals to respond to questions that describe their

interests, attitudes, and values. These types of assessment require the respondent to

provide different types of responses making use of various cognitive processes and raising a diverse set of concerns when responses are being interpreted (Lietz, 2010).

Regardless of the types of responses required, the utility of any assessment hinges upon the ability of the clinician or researcher to make meaningful sense of the assessment's results. In order for an assessment measure to provide meaningful information, it must demonstrate certain psychometric properties; specifically, it must be both reliable and valid. Generally, reliability refers to the consistency of participant responding to the test items (Popham, 2011). This can be consistency from one test administration to another, consistency between different equivalent forms of a test, or consistency of responding within the items of the test. Just like an individual would want their car to reliably get them to work every day, so does a clinician want their assessments to reliably provide information about the individuals completing the assessments.

The second general psychometric property of a test in which researchers and clinicians are interested is validity. In the realm of psychological assessment validity generally refers to the extent to which a particular instrument actually measures what it purports to measure (Anastasi, 1968, cited in Greene, 2000). There are several different aspects of validity. Construct validity has come to be recognized as the fundamental validity concept, and a key method of validating theoretical constructs (Thompson and Daniel, 1996; Anastasi and Urbina, 1997). As each of the other aspects of validity provides important information regarding the constructs measured by a test, the following validation concepts will be discussed: content validity, criterion validity, and construct validity.

Content validation procedures seek to examine the actual content of the test items in search of evidence that the test is asking questions that represent the construct being measured (Anastasi & Urbina, 1997). For example, an assessment designed to measure simple arithmetic ability that is composed of mainly differential functions would lack content validity. Similarly, an assessment designed to measure mathematical computation ability containing solely word problems may not have adequate content validity. It is possible that this assessment, while still requiring students to solve mathematical equations, is actually assessing reading comprehension skills instead. It is therefore important to examine the actual content of the items included in an assessment before making important inferences about an individual based upon their performance on that assessment.

Criterion-related validation procedures seek to examine the extent to which an assessment effectively predicts an individual's performance on some other measure (Anastasi & Urbina, 1997). For example, universities often use applicants' high school grade point average (GPA) as a predictive measure of future academic performance. Investigation of this would look at the relationship between these two measures (high school GPA and college GPA, for example). If a strong positive relationship is found between the two measures, then this would serve as criterion-related validity evidence to support the use of high school GPA as a predictive measure of future academic performance.

Finally, construct validation procedures seek to determine the extent to which an assessment is actually measuring a particular theoretical construct (Anastasi & Urbina, 1997). Theoretical constructs are hypothetical traits that cannot be directly and

concretely measured; they must be investigated through behavioral variables that can be directly observed and measured. For example, a person cannot see an actual physical entity called "anxiety." It is possible, however, to observe a rapidly tapping pencil, or an individual taking short and shallow breaths. These are common behavioral indicators of anxiety symptoms, and can be observed or reported to obtain a measure of anxiety.

Behaviors, however, can be caused by many different outside events. An individual taking short and shallow breaths, for example, may not be doing so because he is feeling anxious, but rather he may have just finished running a marathon. Due to this, there are various ways that researchers can go about examining the construct validity of assessment scales. One of the ways to go about providing evidence of this type of validity is through factor analysis. Factor analysis attempts to reduce the information provided by individuals completing a certain assessment to a relatively small number of common traits (Anastasi & Urbina, 1997). These common traits are the theoretical constructs of interest. In this way, factor analysis can provide an indication of which items on a particular scale appear to be measuring the same construct.

This study investigates the construct validity of the IPIP-50, an online personality scale. This particular scale is based upon the Five Factor Model of personality. Given this, the underlying constructs are already defined and the investigation will be aimed towards finding support for these constructs through a factor analytic approach to the analysis of the data collected. A more in depth review of the Five Factor Model and the utility of factor analysis when investigating the construct validity of a scale is presented in the next chapter.

Number of Response Options on Affective Scales

As previously mentioned, affective measurement scales differ from other types of assessment in that these particular scales require individuals to reflect upon and rate their own or someone else's behavioral or social emotional state. The common response modality for affective assessment measures is a rating scale with Likert-type response options (Popham, 2011). These scales require an individual to read a statement and choose the corresponding point on a continuum that they feel best represents their current status with regards to the statement. An example would be responding to a statement such as "I often feel joyful" by choosing a number 1-7. In such a scale, choosing the number 1 may indicate that the individual never feels joyful while a response of 7 may indicate that this statement describes the person very well.

There are many different ways in which affective measurement scales can be formulated. As described in the above example, individuals can be asked to choose a number on a scale to indicate how little or how much the statement describes their current state. Another formulation is for individuals to choose one of several words that would best categorize their current status with regards to the statement. Using this formulation, in response to the statement "I often feel joyful," an individual may be faced with choosing from category options such as "Never true for me," "Sometimes true for me," "Often true for me," or "Almost always true for me."

In addition to the two formulations already described, such scales can be presented using pictures in place of numbers or words. An example of this would be requiring that the individual respond to a statement such as "I enjoy going to school" by choosing a smiling face to indicate agreement, a frowning face to indicate disagreement,

or a neutral face to indicate that they neither agree nor disagree with the statement. Scales formulated in this way would allow for the use of affective scales with populations commonly exhibiting lower levels of literacy skills, such as young children (Miller, Linn, & Gronlund, 2009).

The optimal number of response options to be included in such affective measurement scales has been greatly debated over the past several decades (Jacoby & Matell, 1971; Cox, 1980; Foddy, 1993; Maydeu-Olivares, et al., 2009; Jones & Loe, 2013). Ideally, scales would be formulated to have enough response options to allow the respondent to accurately indicate their current status, but not so many options as to confuse or unnecessarily tax the individual. These conditions are critical when researchers and clinicians want to use these scales to make any sort of accurate inference about the respondents.

When individuals are not provided enough options, they may feel like none of the provided options accurately reflects their attitude. Similarly, when individuals are presented with so many options that they become overly taxed, they may stop taking the time to respond openly and accurately. Both of these instances can result in the individual responding in an inconsistent or random manner, which has been a continued concern in the realm of affective assessment (Siefert et al., 2012).

It has been proposed that reducing the amount of effort, or perceived effort, required to complete affective measures can reduce the likelihood of invalid responding and participant dropout (Helgeson, Voss, & Terpening, 2002). Forbey and Ben-Porath (2007) suggested that reducing tedium for the individual responding to the instrument would increase the validity of responding. It has also been proposed that reducing the

frustration felt by the test taker would reduce the instance of careless responding (Schmidt, Le, & Ilies, 2003). Reducing the number of response options presented on the scale for each item is one example of a way in which the effort required to complete an affective measure can be addressed.

In this way, an individual filling out an affective measure requiring a rating choice from five possible options would reasonably be expected to spend quite some time completing this measure, especially if the scale is rather lengthy and the individual is giving thoughtful consideration to each item. That same person, however, may require much less time to thoughtfully complete the same affective measure if there were only two options from which to choose a response for each item.

Regardless of the amount of time saved, however, the scale would still need to have adequate psychometric properties with fewer response options in order to be useful to clinicians as well as researchers. This study also aims to investigate the impact on the construct validity of the IPIP-50 when response options are reduced from five to two. A more in depth review of the literature surrounding the optimal number of response options to be included in an affective measurement scale is presented in the next chapter.

Significance of the Problem

As mentioned above, when clinicians and researchers make use of any set of assessment instruments, it is fair to say that they are expecting those instruments to actually measure the construct that the instruments purport to measure. The scores obtained on any assessment are used to make certain inferences about individuals, which can only be done if the assessment instrument has a certain degree of construct validity.

In this way, a school psychologist who administers a depression scale to a student who she suspects of experiencing a certain level of depression is expecting the depression scale to actually be measuring the construct of depression. It does the school psychologist no good if the depression scale is actually measuring anxiety or some other construct.

Not all theoretical constructs are simple. There are often several different components to one particular construct. For example, the construct of anxiety can be broken down into individual components such as worry, physiological anxiety, and social anxiety (Reynolds & Richmond, 2008). It is important for clinicians and researchers to know whether the assessment scales that they are using are actually measuring the constructs in which they are interested, as well as what particular aspects of the constructs are being measured.

Since the factor analytic approach allows for the investigation of the common traits upon which items in a scale tend to group, it is a promising means of providing support for the construct validity of a particular scale. When practitioners have evidence to support the constructs that are being measured by the assessments that they use, they are able to make better choices regarding the measures that they choose to use in their practice. Measures with stronger construct-related validity evidence will allow for the collection of more accurate information regarding the variables of interest, which in turn allows for better decisions to be made based upon the assessment results.

Similarly, as stated above, when an individual feels overly taxed by an assessment, or that the assessment is too tedious, they may not provide the most accurate responses. If reducing the number of response options that an individual has to read and

contemplate when responding to such a measure can increase the likelihood of accurate responding without decreasing the reliability and validity of the measure, then this could help address this problem of inaccurate responding. With less inaccurate responses, more accurate inferences can be drawn from the assessment results.

Research Questions

This study is composed of two separate but related investigations. The question guiding the first investigation is:

- Is the Five Factor Model of personality evident in the factor structure of the IPIP-50?

The question guiding the second investigation is:

- What is the impact on the factor structure of the IPIP-50 when response options are reduced from 5 to 2?

Hypotheses

These research questions were tested using two null hypotheses. To address the research question guiding the first investigation, the following null hypothesis was used:

1) The Five Factor Model of personality is evident in the factor structure of the IPIP-50.

To address the research question guiding the second investigation, the following null hypothesis was used:

2) Reducing the number of response options on the IPIP-50 from 5 to 2 does not change the factor structure of the scale.

Assumptions

The self-report nature of the instrument used in this study requires certain assumptions to be made. The first assumption is that the participants were able to make accurate self-judgments when completing the scale. The IPIP-50 requires participants to rate certain aspects of their personality based upon five factors. This assumes that a person is able to know himself well enough to actually make those ratings. It also assumes that the participants were able to accurately communicate those self-judgments once they were made.

Also, since an already existing database is used for the investigations in this study, it is assumed that each participant completed the study tasks accurately and to the best of their ability. Since the instruments were administered online and there was therefore no opportunity for a participant to ask for clarification on items, it is assumed that the participants were able to understand their tasks as described in the study materials, as well as the content of each item on the IPIP-50.

Limitations and Delimitations

As with any study, there are some limitations and delimitations to this study. One example is the sample. The sample used for this study is a convenience sample and was delimited to undergraduate and graduate students enrolled in courses in educational psychology at an urban university. As the majority of the subjects are therefore expected to be education majors, the demographics of this particular group may not be adequately representative of the general population, which may limit the generalizability of the results.

Another delimitation of this study is that the data were collected in a purely online format. This could further limit the sample to just those individuals who have access to a computer and the Internet. Given the current state and availability of technology, however, this should not pose too much of a problem. This delimitation brings with it the added limitation, however, that since the participants completed their involvement from a remote location there was no way to verify the identity of the individual responding or ensure that the information they provided is accurate. There was also no way to check the participants for signs that they may have been losing focus or getting off task.

Lastly, due to the nature of the topic and the instrumentation used, this study relies upon self-report data. The participants were asked to make reflections upon themselves and their own personality traits. Since there is no practical way for the responses of the individuals regarding different aspects of their personality to be verified, this study relies upon the participants' ability to make accurate self-judgments and then truthfully portray those self-judgments through their responses to the items on the scale. There is potential for participants to want to portray themselves in a positive (or negative) light. In this way, even if a participant judges herself to be one way in daily life, she may choose to portray herself differently through her responding so as to appear more (or less) socially desirable than she may actually feel that she is. This is another limitation to the study.

This last delimitation, however, actually works in a way to alleviate some of the potential limitations that it initially appears to place upon this study. Due to the fact that the participants are responding to the questions via computer they do not actually come into contact with a person who is attached to the study. This can provide the participants

with a sense of anonymity and take away some of the pressure that they may feel to portray themselves as positively as possible (Albaum, Wiley, Roster, & Smith, 2011; Popham, 2011). Without the physical presence of a proctor, there may be less perceived potential for outside judgment based upon their responses and therefore increase the likelihood of accurate responding.

Implications in School Psychology

Potential implications that can be generated from this study's results may influence the number of response options provided when using or creating an affective measurement scale in both research and clinical school psychology settings. Quite often school psychologists use behavior rating scales to obtain information from teachers, parents, and children themselves, regarding the behavior of their students. School psychologists rely upon the accuracy of this information to make decisions regarding students' eligibility for special education and related services.

The heavy workload of teachers can lead to rushed or less well contemplated ratings of the students on these scales, or potentially not having time to complete them at all. If presenting these assessments to teachers with fewer response options, or in abbreviated forms, can provide protocols that are no less valid than full scales or those including more possible responses, then teachers may feel less taxed by the request to complete them and be more likely to willingly comply with the school psychologist's request for information. Thus, the results of this study can contribute to the body of research pertaining to the greatly debated topic regarding the ideal number of response

options on measurement scales, as well as possibly influencing the future use and development of affective measurement scales.

Additionally, affective scales can be used by school psychologists and teachers to help teachers inform their teaching practices. Considering the push towards using evidence-based interventions in the school system it is important that data exist to support instructional methods and techniques used with students at all levels. Behavior rating scales and affective measurement scales can be utilized by teachers to monitor the progress of their students with regards to certain variables of interest. Making these scales less taxing and cumbersome for teachers to complete, but still maintaining their validity, could increase the likelihood that teachers would actually use them to collect data on their students. These data would provide valuable information as to the effectiveness of the teaching approaches that they are using with their students and would provide them with evidence to support the continuation or change of a particular intervention or teaching method.

CHAPTER 2

BACKGROUND LITERATURE

Theoretical Foundations of the International Personality Item Pool

The International Personality Item Pool (IPIP) is based upon the five-factor model of personality (FFM). The FFM is reportedly the most significant paradigm in personality research (Johnson, 2005). The FFM stems from the factor analytic tradition. Through the factor analyses of vast amounts of personality test and rating scale data, five factors have been found most consistently to account for most of the correlations among the data (Anastasi & Urbina, 1997). Out of this approach the Revised NEO Personality Inventory (NEO PI-R) was born. This measure provides scores on five major dimensions of personality, identified as the following: Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C) (Costa & McCrae, 1992). Anastasi and Urbina (1997) report that the NEO PI-R was originally designed to measure normal personality traits, but it was intended to be useful in clinical, applied, and research settings. This Five-Factor Model has been widely accepted and incited much research activity in this area (Anastasi & Urbina, 1997).

With the help of researchers in the Netherlands and Germany, Goldberg developed a pool of 1252 items which was deemed the International Personality Item Pool (IPIP) (Goldberg, 1999). These scales are an open source resource and have been empirically found to perform equal to or greater than commercial personality tests (Johnson, 2005). The scale used in the present study is one of the scales from the IPIP that is based upon the Five-Factor Model, providing it with a strong theoretical

foundation. The subscales of the IPIP-50, which is the 50 item scale of the International Personality Item Pool, were developed to measure the Big-Five factor markers of Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness to Experience (Goldberg, 1992).

Review of Existing International Personality Item Pool Validation Studies

Several studies were found that directly addresses the issue of validity and the International Personality Item Pool. One investigation, conducted by Johnson (2005), investigates the validity of the IPIP-50 by attempting to analyze the validity of the individual protocols produced by individuals responding to the items on the IPIP-50. In this study the researchers propose statistical methods through which users of this scale can ascertain valid protocols from invalid protocols when the scale is administered in an online forum.

Other researchers (Baker, Victor, Chambers, & Halverson, 2004) attempted to ascertain the construct validity of this scale when the scale was being used to obtain ratings of adolescent personality. Among other statistical methods, confirmatory factor analysis was used in this study to provide an indication of the validity of the scale when used with an adolescent population. This study compared personality ratings for adolescents obtained from teachers, peers, and the adolescents themselves. The researchers found that the five-factor model was largely supported in their data; however, the model was only found to have a moderate fit with TLI values under 0.90 and CFI values above 0.90. The RMSEA values ranged from 0.057 to 0.093 for the different conditions. These values indicated only a moderate fit of the five-factor model.

Another study looked at a short form of the IPIP-50, which consists of only 20 items that were drawn from the IPIP-50 (Cooper, Smillie, & Corr, 2010). This study attempted to investigate the factor structure of this abbreviated scale through the use of confirmatory factor analysis. The researcher found mixed results with RMSEA and RMSR values within acceptable ranges, but with CFI values indicating poor model fit. These researchers also report a generally poor fit found through confirmatory factor analysis of five-factor personality measures (Church & Burke, 1994, cited in Cooper, Smillie, & Corr, 2010). When examining the factor structure of the mini IPIP through exploratory factor analysis, the researchers found better indications of model fit.

A study conducted in Scotland demonstrated support for the five-factor model in the IPIP-50 through the use of exploratory factor analysis (Gow, Whiteman, Pattie, & Deary, 2005). These researchers found only minor deviations from the expected factor loadings for each item in each of their samples. The samples included a mostly undergraduate student sample, a group of volunteers from the general population, and a particular Lothian Birth Cohort. Overall principal components analysis results largely supported the five-factor model.

Overall results of these studies indicate a few important points. First, it appears that there is a paucity of research directly investigating the factor structure of the International Personality Item Pool scales. Secondly, exploratory factor analysis results appear to suggest better model fit than do confirmatory facto analysis results. These studies suggest that there is much room for continued research in this area. This line of research is also important, as this particular scale is a promising resource for researchers and clinicians looking to investigate personality traits.

Background Information on Factor Analytic Approaches to Psychological Research

Fabrigar, Wegener, MacCallum, and Strahan (1999) report that since the development of exploratory factor analytic methods of research in the early 1900s (Spearman, 1904; Spearman, 1927) it has become a widely used statistical procedure within the realm of psychological research. Despite their widespread use, however, factor analytic methods are not immune to criticism. Concerns regarding the ability of factor analytic techniques to add to the development of theories have been raised (Hills 1977; and Overall, 1964) as well as criticisms of the way in which factor analytic techniques are employed (Skinner, 1980).

Fabrigar and colleagues (1999), however, report that factor analytic methods can be useful in psychological research so long as they are implemented in sound ways. They recommend the following five major methodological issues that researchers must consider when conducting a factor analysis:

First, he or she must decide what variables to include in the study and the size and nature of the sample on which the study will be based. Second, a researcher must determine if EFA is the most appropriate form of analysis given the goals of the research project. Third, assuming that EFA is appropriate, a specific procedure to fit the model to the data must be selected. Fourth, the researcher must decide how many factors should be included in the model. Finally, it is usually necessary for a researcher to select a method for rotating the initial factor analytic solution to a final solution that can be more readily interpreted. Each of these decisions can have important consequences for the results obtained. (p. 273)

As each of these decisions can have an impact on the usefulness of the information obtained through the factor analytic techniques employed, it is sound advice to follow when considering such an approach to a line of research.

Review of the Research Regarding the Number of Available Response Options

Empirical Research Supporting the Use of Greater Numbers of Response Options

The debate over the optimal number of response options on affective assessments dates back as far as these sorts of scales have been in use. Murphy and Likert (1938) proposed that the reliability of a scale could be improved through increasing the number of response options. A study presented demonstrated an increase in reliability coefficients from .88 when 44 three-point items were used, to .94 when 36 five-point items were used (Murphy & Likert, 1938, cited in Cox, 1980). Similar to this, Cox (1980) suggests that seven, plus or minus two, is a sort of "magic number," with five options being adequate.

Supporting Cox (1980) in the suggestion of the "magic number" of response options, it has been reported that 5-point to 7-point scales are the most commonly used in questionnaire design (Dillman, 2000; Fink, 2003; Brace, 2004; Lietz, 2010). This practice is supported through research suggesting that a 7-point scale is more reliable (Chronbach, 1951). These researchers lend support to the idea that providing a greater number of response options is more psychometrically desirable.

Affective measurement scales are not limited to the realm of personality assessment and the social sciences, and several reviews of marketing literature were found that address this issue. Churchill and Peter (1984) conducted a review of empirical

marketing studies that used rating scales. In this investigation reliability estimates were gathered and a meta-analysis was conducted. The results of this meta-analysis indicated that the reliability estimate of the affective scales used in the included studies was higher when the scales provided a greater number of response options (accounting for 5% of the variance), thus supporting the notion that more options are better than fewer options.

Empirical Research Supporting the Use of Fewer Response Options

Early support can also be found for the use of a decreased number of response options. Jacoby and Matell (1971) conducted an investigation in to the impacts on reliability and validity of a value scale. To accomplish this, they adapted 18 different versions of the scale, with the number of response options for each item ranging from 2 to 19. Participants were assigned to each condition based upon order of arrival (the first person received the 2-option scale, the second person received the 3-option scale, and so on). After aggregating the data, the researchers found no increase in the estimates of reliability or validity when the number of response options was increased (Jacoby & Matell, 1971). Given no statistically significant increase in psychometric properties when increased numbers of responses are provided, it is recommended that using only three response options would work just as well as using a greater number.

Singer and colleagues (2010) conducted a study investigating the impact that the provided response options have on the relationship between racial attitudes and beliefs about the causes of behavior. This study looked at individuals' responses regarding this topic when questions were presented in a binary-response format (yes/no) versus a 21-point scale. Participants were randomly assigned to either respond to a positively worded 21-point scale, a negatively worded 21-point scale, a positively worded binary scale, or a

negatively worded binary scale. When looking at differences in responding based upon the number of response options, the researchers did not find significant differences in respondents' expressed beliefs (Singer et al., 2010). This indicates that a binary-choice scale is just as effective in allowing respondents to express their opinions as a 21-point scale, thus lending further support to the notion that a large number of response options are not necessary when designing affective measures.

Jones and Loe (2013) conducted a study wherein they found support for the use of fewer response categories with a much more recent sample. This study investigated the impact on the psychometric properties of a vocational personality scale when response options were reduced from ten to three. There were two parts to this study, a simulation study and an empirical investigation. In the simulation study data were collected from individuals completing the scale in the usual 10-response-option format and the psychometric properties of the scale analyzed. Data were then reanalyzed under a simulation condition wherein individual responses were recoded as if the instrument had been administered with a three-point scale (as opposed to ten). This recoding was completed in two different ways, presenting two different possible simulation scenarios. Reliability coefficients were calculated for each of the subscales on the instrument in the 10-response-option condition as well as for the two 3-response-option simulation conditions. No significant differences between the reliability estimates under each condition were found.

Through this process, the researchers were able to look for differences in the estimates of reliability for the vocational personality scale under the different conditions. The results showed no significant differences in the reliability estimates obtained, thus

challenging the notion that affective scales with greater numbers of response options have increased psychometric properties. This study was limited, however, by the fact that simulated data were used to obtain the estimates in the 3-response-option conditions. An attempt to replicate these findings under more strict experimental conditions by Jones and Loe was also reported (Jones & Loe, 2013).

In their follow up study Jones and Loe recruited participants through a university clinic/lab setting to participate (Jones & Loe, 2013). The same instrumentation was used in this study as was used in the initial simulation study, but the design was altered in a way that would allow direct comparisons to be made between participants who were randomly assigned to complete the vocational personality scale in either a 6-point Likert format or a binary choice format. Once again, no significant differences in the reliability estimates for the scale were found between the two conditions. This lends further empirical support to the idea that more response options may not be inherently better when it comes to affective scales.

There is a drawback to this study, however. The core instrument in this investigation requires respondents to make a forced choice between two different vocational personality traits, thus providing ipsative scores. Jones and Loe (2013) report that this ipsative scoring wherein an artificial interdependence exists between the scales of the instrument, which may influence the psychometric properties of the scale. The researchers recommend further study of this issue using an instrument that provides normative scores, rather than ipsative.

Empirical Research Reporting Mixed or Conditional Results

Maydeu-Olivares, Kramp, Garcia-Forero, Gallardo-Pujol, and Coffman (2009) postulate that the lack of consistency in the research is due to different researchers using different outcome measures (some studies using reliability and other using validity) and approaching the topic from differing psychometric models (classical test theory, item factor analysis, or item response theory). To address this, the researchers conducted a repeated measures study in which they tested the impact of increasing the number of response options on the psychometric properties of personality scales from three different psychometric models. For this study, Maydeu-Olivares and colleagues examined the data for impacts on internal consistency, internal structure, and convergent and divergent validity from all three of the psychometric models previously mentioned.

The researchers found that when the response options were increased from two to five, the reliability increased, there were small to negligible impacts on predictive validity, and structural validity actually worsened (except for classical test theory, wherein structural validity was not applicable) when examined from the three psychometric models (Maydeu-Olivares, et al., 2009). This line of research would seem to suggest that the optimal number of response categories depends on the psychometric model and the anticipated properties of the data. When employing a classical test theory framework, the researchers recommend using more response options. When employing either of the other two models, however, the number of response options to be used would depend. Under these other two frameworks, if reliability were of concern then a greater number of response options would be recommended. However, if the structural

validity of the scale is of greater importance, than it may prove beneficial to decrease the number of response options presented.

Instead of the number of response options depending on the psychometric model and anticipated properties of the data, Foddy (1993) suggests that the ideal number of response options should be dependent upon the judgments that the respondents are being asked to make. In this way, Foddy recommends that when respondents are asked to make absolute judgments, then a shorter scale (one with fewer response options provided) would be ideal. However, if respondents were being asked to make abstract judgments, then a longer scale (one with more available response options) would be recommended. This recommendation makes sense when the cognitive processes behind such judgments are considered.

When an individual is required to make an absolute judgment (ie. whether or not they would behave in a particular way) there are fewer discriminations possible and therefore fewer response options are necessary. For abstract judgments (ie. how much they like a particular song), however, there is more room for variation and more available response options may be able to better capture the differences in discrimination. From this perspective, the determination of how many response options to provide would be made based upon the types of questions asked more than the psychometric properties of the questionnaire.

Empirical Research Regarding Even or Odd Numbers of Response Options

Further complicating the issue of how many response options to include in affective assessment instruments is the issue of providing an even number of response options (thus eliminating a midpoint option) versus an odd number of response options

(which would include a midpoint). The former alternative forces the respondent to choose an option leaning towards one side or another of the particular scale while the latter alternative allows for respondents to take a more noncommittal, neutral, position through the provision of an exact midpoint. A review of research supporting each practice follows.

Support for the use of an even number of response options. Krosnick (1991) put forth the "satisficing" hypothesis, which suggests that many individuals who are filling out questionnaires will have low motivation and will choose the response options requiring the least amount of thought, commitment, and justification. In many cases this response would be the midpoint. In this way, Krosnick's (1991) hypothesis would suggest that when a midpoint is offered, respondents will be more likely to choose that point rather than thoughtfully decide which end of the spectrum they actually agree with and information will be lost. This hypothesis would lend support to the notion of providing an even number of response options for items on affective scales, so that respondents will be required to make a choice as to which side they agree with rather than taking the so-called easy way out and simply choosing the noncommittal midpoint.

Similarly, Baumbartner and Steenkamp (2001) investigated certain styles of response that an individual can adopt when completing a rating scale. These response styles included a tendency to choose the midpoint regardless of the content of the item. The researchers refer to this style as midpoint responding and report that it may be due to an evasive desire of the respondent to not reveal their true opinion, an indecisive uncertainty as to the respondent's position, or an indifferent lack of interest in the variable being measured (Baumbartner & Steenkamp, 2001). This raises the issue that a

tendency to choose the midpoint when responding to a scale is a form of bias which can contaminate the results.

When an individual's responses to a set of items are used to make inferences about their status with regards to the variables addressed, those inferences rely on accurate responding. If a respondent consistently chooses the midpoint on a scale for the reasons listed above, they are not necessarily providing the accurate information needed for researchers or clinicians to make those valid inferences. In this way, a case could be made that it is advantageous for this sort of measurement scale to provide an even number of response options so as to not allow respondents the opportunity to consistently choose a midpoint and keep their true opinions or attitudes hidden.

A study conducted by Borgers, Hox, and Sikkel (2004) supports this. These researchers investigated the impact on the reliability of self-esteem scales when a midpoint was provided in the scale compared to when there was no midpoint offered. The respondents in this study were children and young adolescents between the ages of eight and sixteen. The results of this study suggest that the optimal number of responses to provide when working with children is four (Borgers et al., 2004). This study further supports the notion that an even number of response options is preferable to an odd number and extends the research base to include samples of children and young adolescents. In this instance, the cognitive development of children is taken into account.

The researchers propose that children do not have the cognitive ability to make reliable discriminations between large numbers of response options and have a tendency to choose the midpoint when provided, even if the midpoint does not accurately reflect their position (Borgers et al., 2004). This would suggest that an even number of response

25

options is ideal in situations where respondents are suspected of not having the ability to make meaningful distinctions between greater numbers of response options and having a tendency to opt towards a midpoint regardless of their true opinions or status with regards to the variables of interest.

Support for the use of an odd number of response options. A team of researchers tested Krosnick's hypothesis presented above (O'Muircheartaigh, Krosnick, & Helic, 2000, cited in Lietz, 2010). These researchers did not find support for Krosnick's hypothesis. Indeed, they found that scales without a midpoint demonstrated higher levels of random error variance and lower validity, suggesting that including a midpoint in the response options increases the psychometric properties of a scale. Meta-analytic research conducted by Saris and Gallhofer (2007) also supports the increase in validity and reliability of a scale when a midpoint is included in the available response options.

Traditionally the inclusion of a midpoint has been recommended in survey development (Madden & Klopfer, 1978; Rappaport, 1982; Courtenay & Weidemann, 1985). The basis of this recommendation was that the inclusion of a midpoint will increase the reliability of a measure and it also provides an option for individuals who truly do not have an opinion on the given topic (Weems & Onwuegbuzie, 2001). This makes sense as a rationale for the following reasons.

When using test scores or survey responses to make meaningful inferences about people's attitudes, it is essential that those responses reflect the individuals' true attitudes. If a respondent truly has no opinion on a matter, or truly feels neutrally about an issue, then they should be provided a means by which to make that opinion known. Removing

the option of a midpoint and thereby forcing the respondent to make a choice when they truly are not drawn to either side of an issue does not improve the understanding of that individual's true feelings. To address this issue, Albaum and his colleagues (2011) advocate the inclusion of a "Prefer Not to Answer" option when a forced answering method, such as providing an even number of response options, is used.

Alternate Validation Approaches

Internal Validity Indices

In the realm of psychological assessment, the term validity tends to refer to the extent to which a particular instrument actually measures what it purports to measure (Anastasi, 1968, cited in Greene, 2000). Within the domain of personality assessment, this frequently equates to the accuracy with which an individual portrays himself or herself when filling out the questionnaire. Since it is often difficult to verify each individual respondent's claims, some of the better-established measures have validity indices built into them. These validity indices give an estimate of the degree to which individuals may be distorting their portrayal of themselves. One example of this is the Minnesota Multiphasic Personality Inventory (MMPI: Hathaway & McKinley, 1940) and the most recent revision, the MMPI-2 (Butcher, Dahlstrom, Graham, Tellegen, & Kraemmer, 1989).

Greene (2000) recounts the process by which the developers of the MMPI came to include such indices. When developing the clinical scales that would become the MMPI, the reality was acknowledged that sometimes people do not provide an accurate self-report. These inaccurate reports could be due to a variety of reasons. Some likely

explanations involve inherent ambiguity in the items, individual variations in self-accuracy and ability and/or willingness to provide accurate reports (Greene, 2000). In this way, individuals may differ in their responding due to different interpretations of an item (Saris & Gallhofer, 2007).

For example, one individual may report that they do feel sad often because they experience that emotion once a month, but another individual may report that they do not feel sad often because they only experience the feeling once a week. These two individuals are interpreting the term "often" in two very different ways. Similarly, self-report inventories make the (often unwarranted) assumption that individuals are able and willing to provide accurate ratings of themselves and various aspects of their personality. Individuals who want to portray themselves a certain way, or who are simply unable to make such self-judgments, will not provide a valid representation of their current affective status.

Given this limitation inherent to self-report personality inventories, the developers of the MMPI developed three validity indices to be included in the instrument (Greene, 2000). These three scales are the L scale, the F scale, and the K scale. The L scale, or the Lie scale, includes items reflecting socially desirable behaviors that are rarely all true for any one individual. Frequent endorsement of these items would therefore indicate a distortion in the responses of that individual. Similarly, the F scale is comprised of items from a variety of areas that were endorsed relatively infrequently by the majority of the normative group. In a similar manner as the L scale, frequent endorsement of items on this scale would indicate response distortion. Whereas the L scale generally identifies individuals seeking to present themselves in a socially desirable light, the F scale

28

identifies individuals who are attempting the intentionally fake pathology. Finally, the K scale was developed to differentiate between individuals who are known to be suffering from psychopathology who yet obtain normal profiles from normal individuals who somehow obtain elevated profiles on the MMPI.

Review of Commonly Used Affective Assessment Instruments Considering the Number of Response Options

Minnesota Multiphasic Personality Inventory, Second Edition (MMPI-2). One example of a commonly used affective measure in the realm of psychological assessment is the Minnesota Multiphasic Personality Inventory, Second Edition (MMPI-2). The MMPI-2 is a prime example of a well-established, highly researched, affective measurement scale that has, throughout its development, offered only two response options for each item on the scale. According to Greene (2000), the original MMPI was created in a two response option format, which was the commonly used and accepted format for affective assessments at the time. Since then, the MMPI was undergone decades of research attempting to fine-tune the instrument and increase its psychometric properties and practical utility (Greene, 2000). Throughout the course of research and development of the scale, the two-response option format has never been altered.

Behavior Assessment System for Children, Second Edition (BASC-2). A contrasting example would be the Behavior Assessment System for Children, Second Edition (BASC-2). The BASC-2 (Reynolds & Kamphaus, 2004) is a measure of an individual's behavioral and social-emotional functioning. There are three different forms of the BASC-2: a teacher-report form (TRS), a parent-report form (PRS), and a self-

report form (SRP). The TRS and PRS forms of the BASC-2 present items in a four-response format. The individual responding to the items is required to select from "Never," "Sometimes," "Often," and "Almost Always" when making their ratings.

The original BASC self-report form consisted solely of a two-response format wherein the individual was required to answer either "True" or "False" to each item. Due to concerns that a two-response format, although adequate, presents potential problems (such as decreased reliability and increased susceptibility to response distortion) other response formats were considered for the revision of the BASC self-report form (Reynolds and Kamphaus, 2004). Pre-standardization research conducted by Altmann (2004) suggested that there was an optimal configuration of response format on the revised BASC-2 self-report scales. This research involved comparing parallel forms of the original BASC self-report questionnaire.

In this way, administrations keeping the original self-report BASC binary response format were compared through a mixed two-factor, within subjects, design with BASC administrations using the same four-response-option format as the teacher and parent report forms. This study conducted by Altmann (2004) reported higher reliabilities on BASC scales when a mixture of two- and four-response option items was used. Based upon this research, the current BASC-2 self-report of personality contains a mix of binary-response and four-response items.

Considering the Inclusion of Validity Scales

As has been demonstrated in the preceding sections, some instruments used within the field of affective assessment have validity indices built into them (like the MMPI-2) while others do not (such as the IPIP-50). Some affective measures used in the

schools utilize a two-response option format while others provide a greater number of response options. The following will review some affective instruments that are commonly used in the practice of school psychology, looking specifically at the number of response options provided and the presence or absence of validity indices.

Behavior Assessment System for Children, Second Edition (BASC-2). As previously detailed, the BASC-2 is a measure of an individual's social-emotional and behavioral status. There are separate forms depending on the age of the individual (preschool, child, and adolescent) and the relationship of the respondent to that individual (parent, teacher, or self-report). The current version of the teacher and parent report forms provides respondents with four response options to each item. The rater can respond to each statement with either "Never," "Sometimes," "Often," or "Almost Always." The self-report form, however, provides respondents with only "True" or "False" as available response options for each item on some items and "Never," "Sometimes," "Often," or "Almost Always" on other items.

There are validity scales built into the BASC-2. These scales include an F index, an L index, and a V index. The F index is included in all of the forms of the BASC-2, whereas the L index and the V index are only included in the self-report forms. The F index provides an indication of the level of negativity being reflected in the ratings. The L index, on the other hand, provides an indication that an individual may be "faking good," or portraying themselves more positively than may be warranted. In contrast to these scales, the V index provides a general validity check for the scale by including several nonsensical or highly implausible statements. The items comprising this scale are

ones that an individual, who is paying close attention to, and understanding the content of the items, would rarely endorse.

In addition to these indices, when computer scored the BASC-2 also offers a Consistency index and a Response Pattern index. The Consistency index evaluates the protocol for discrepancies in responding to items that should be answered similarly. The Response Pattern index looks for repeated and cyclical patterns of responding to the items on the scale.

Children's Depression Inventory (CDI). The CDI is a self-report measure of a child's level of depressive symptoms and is intended for use with children ages 7 through 17 years. There are two versions available, a long version and short version. The short version can be used as a screening instrument, whereas the long version is useful to paint a more complete picture of an individual child's current experience of depressive symptoms. Each item on the CDI provides the child with three statements. From these three statements the child is required to choose the one that is the most accurate for them. Each item reflects a different aspect of depressive symptomatology.

The test developers recommend that the individual's responses be examined and considered within the context of their observed behavior so as to make a clinical judgment on their validity (Kovacs, 1992); however, no formal validity indices are included in this instrument.

Piers-Harris Children's Self-Concept Scale, Second Edition (Piers-Harris 2). The Piers-Harris 2 is a self-report measure of how a child views him or her self and is intended for use with children ages 7 through 18. This scale was written in language that was intended to be understood by individuals with at least a second grade reading level.

It is presented in a two-option response format. The individual child responding to the questionnaire must choose a response of either "Yes" or "No" for each statement.

The Piers-Harris 2 has two validity indices. These include an Inconsistent Responding index and a Response Bias index. The Inconsistent Responding index looks for discrepancies in an individual's responding to items that should be answered similarly and provides an indication of random responding. The Response Bias index examines the respondent's tendency to agree (or disagree) with the test items regardless of the actual content of the item.

Revised Children's Manifest Anxiety Scale, Second Edition (RCMAS-2). The RCMAS-2 (Reynolds & Richmond, 2008) is a self-report measure of a child's current experience of anxiety and is normed for use with individuals aged 6 through 19. This instrument provides ratings on three different dimensions of anxiety. The items are presented in a binary choice response format where the child is required to choose an answer of either "Yes" or "No" for each item.

The RCMAS-2 provides two validity indices, an Inconsistent Responding index, which looks for discrepancies in an individual's responding to items that should be answered similarly and provides an indication of whether the individual was attending to the test items, and a Defensiveness index.

Reynolds Adolescent Depression Scale, Second Edition (RADS-2). The RADS-2 is a self-report inventory designed to identify depressive symptoms in adolescents aged 11 to 20 years. In addition to the standard form, there is also a short form available that can be used as a screening instrument for further evaluation. The items on the scale are presented in a four-point response option format.

This instrument does not include any formal validity index.

Reynolds Child Depression Scale, Second Edition (RCDS-2). The RCDS-2 is a self-report inventory designed to identify depressive symptoms in children aged 7 to 13 years. In addition to the standard form, there is also a short form available that can be used as a screening instrument for further evaluation. The items on the scale are presented in a four-point response option format.

This instrument does not include any formal validity index.

Present Study

From the research presented in the preceding sections, it is clear that a factor analytic approach can provide important information regarding the construct validity of a scale. It is also clear that there is great debate in the field regarding the optimal number of response options on affective scales and the impact on the psychometric properties of those scales when different numbers of response options are provided. Additionally, there is clearly room for further investigation of the use of factor analysis in validating the International Personality Item Pool, as well looking into the impact that altering the number of available response options on this scale has on the factor structure. This study adds to this body of literature by investigating these areas. Specifically, the study provides data regarding the factor structure of the IPIP-50 and the impact on that factor structure when item response options are reduced from 5 to 2. The study is outlined in the following chapter.

34

CHAPTER 3

RESEARCH METHODS

Purpose of Study

This study is comprised of two separate but related investigations, each investigating the general hypothesis that the Five Factor Model is evident in the factor structure of the fifty-item scale of the International Personality Item Pool (IPIP-50). This study relies upon secondary analysis of two already existing databases, and uses participants' scores to investigate the factor structure of the IPIP-50 and the impact on that factor structure when the available response options are reduced from five to two.

Study 1

Research Questions

The question guiding the first investigation is: Is the Five Factor Model of personality evident in the factor structure of the IPIP-50?

Hypotheses

This research question was tested using the following null hypothesis: The Five Factor Model of personality is evident in the factor structure of the IPIP-50.

Participants

A de-identified dataset was used in this analysis. The dataset was drawn from a collection of approximately 800 undergraduate and graduate students attending an urban southwestern university in the United States in the spring semester of 2011 and summer semester of 2012. Participants comprising the dataset were volunteers who chose to

complete a particular study from several different IRB approved projects available to meet a research requirement for undergraduate and graduate courses in educational psychology between spring 2011 and early summer 2012. Data on the gender, ethnicity, age, and self-reported grade point average (GPA) of the participants will be reported in the following chapter.

Measures

International Personality Item Pool (IPIP-50). In this study the big five personality characteristics were measured using the 50-item scale from the International Personality Item Pool (http://ipip.ori.org/ipip/). The International Personality Item Pool (Goldberg, et al., 2006) is an open source resource with sample questionnaires for personality constructs including the five-factor model of extraversion (E), neuroticism (N), agreeableness (A), conscientiousness (C), and openness to experience (O).

Respondents to this particular scale of the IPIP (the IPIP-50) are required to respond to a series of 50 item statements by choosing the point on a 5-item scale which they feel best represents them for each item. The available response options for each item are Very Inaccurate, Moderately Inaccurate, Neither Accurate Nor Inaccurate, Moderately Accurate, or Very Accurate.

Procedure

The cases from the dataset that were used in this study include undergraduate and graduate students from whom only an IPIP-50 score was obtained. The study materials were all delivered over the Internet in a web based electronic format and included a demographic questionnaire to obtain information regarding age, gender, and ethnicity of the participants.

Prior to data analysis, cases where participants failed to provide a response to any item on the IPIP-50 were removed. The remaining data were subjected to analysis.

Data Analysis

Descriptive statistics. Frequency counts were used to analyze the gender, age, ethnicity, and self-report GPA of the participants. This information is presented in the following chapter.

Construct validity. An exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA) were run on the data collected for this study to provide evidence of construct validity. Since the IPIP-50 is based upon the five-factor model of personality, it is expected that questions load on five factors. If this is not the case and more or less than these five factors are found, this may indicate a problem with the scale. Finding evidence for the Five-Factor Model in this data would lend support to the construct validity of the scale.

Study 2

Research Questions

The question guiding the second investigation is: What is the impact on the factor structure of the IPIP-50 when response options are reduced from 5 to 2?

Hypotheses

This research questions was tested using the following null hypothesis: Reducing the number of response options on the IPIP-50 from 5 to 2 does not change the factor structure of the scale.

Participants

A separate de-identified dataset was used for analysis in this second investigation. This dataset was similarly drawn from a collection of approximately 422 undergraduate and graduate students attending an urban southwestern university in the United States (208 participants in the 5-alternative condition, and 214 participants in the 2-alternative condition). Participants comprising the dataset were volunteers who chose to complete this study from several different IRB approved projects available to meet a research requirement for undergraduate and graduate courses in educational psychology between late summer 2012 and spring 2013. Data on the gender, ethnicity, age, and self-reported grade point average (GPA) of the participants will be reported in the following chapter.

Measures

International Personality Item Pool (IPIP-50). The IPIP-50 was also used in this investigation. As mentioned previously, the International Personality Item Pool (Goldberg, et al., 2006) is an open source resource with sample questionnaires for personality constructs including the five-factor model of extraversion (E), neuroticism (N), agreeableness (A), conscientiousness (C), and openness to experience (O).

Respondents to this particular scale of the IPIP (the IPIP-50) are required to respond to a series of 50 item statements by choosing the point on a 5-item scale, which they feel best represents them for each item. The available response options for each item are Very Inaccurate, Moderately Inaccurate, Neither Accurate Nor Inaccurate, Moderately Accurate, or Very Accurate.

Procedure

The cases from the datasets that were used in this study include undergraduate and graduate students from whom IPIP-50, CogStyle, reading comprehension, and reasoning ability scores were obtained. Information in this database includes scores on the IPIP-50 when participants were randomly assigned to complete the measure with either 5 response options or 2 response options.

Prior to analysis, the database was examined for any cases to be removed from the analysis. Accordingly, following the procedure used by Jones and Loe (2013), cases with a completion time too short to indicate completion of all study instruments with anything other than random responding was determined by the frequency distribution for the completion time variable. In this way, cases in which the completion time was two standard deviations below the mean completion time for the sample were deleted. Also, just as in the simulation study, cases where a response to any item on the IPIP-50 was omitted were removed from analysis.

It is also important to note that in this study the impact of reducing the number of response options while maintaining a midpoint could not be investigated. Due to limitations with the software used to randomly assign participants to conditions; only two conditions could be programmed, thereby limiting the data to include just two conditions. Accordingly, random assignment data were not provided in the database for a condition reducing the number of response options from 5 to 3 in this investigation.

Data Analysis

      Descriptive statistics. Frequency counts were used to analyze the gender, age, ethnicity, and self-report GPA of the participants. This information is presented in the next chapter.

      Construct validity. An exploratory factor analysis (EFA) and a confirmatory factor analysis (CFA) were run on the data collected for this study to provide evidence of construct validity. The EFA and CFA were run on the data collected for each condition. Since the IPIP-50 is based upon the five-factor model of personality, it is expected that questions load on five factors in each condition. If this is not the case and more or less than these five factors are found in either condition, this may indicate that the number of response options is impacting the factor structure of the scale. Finding evidence for the Five-Factor Model in this data would lend support to the construct validity of the scale in each condition.

CHAPTER 4

RESULTS

Study 1 Results

Descriptive Statistics

The dataset used for the simulation study was comprised of 853 cases. After

listwise deletion of cases omitting a response to any item on the IPIP-50, 801 cases

remained for analysis. Table 1 shows the distribution of the sample based upon different

descriptive criteria. Perusal of Table 1 would indicate that of these remaining cases, 74%

were female ($n = 596$) and 25% were male ($n = 204$). One respondent did not provide an

answer to this question. This unequal distribution is representative of the greater

population of education majors from which this sample was drawn.

Individuals falling within the age range of 18-25 made up 59% of the sample ($n =$

476), 26% fell within the age range of 26-35 ($n = 206$), 8% were in the range of 36-45 (n

$= 62$), 7% fell within the 46 and above range ($n = 55$), while 2 individuals (comprising

0.2% of the sample) did not provide an answer to this question. Self-reported ethnic

backgrounds were Caucasian (63%), Hispanic (14%), African-American (9%), Asian

(7%), Other (4%), Pacific Islander (2%), and Native American (0.5%). Ethnicity data

were not provided by 0.6% of the sample.

Validity Analysis

Exploratory factor analysis. All questions on the IPIP-50 were factor analyzed

using principal components analysis with Varimax (orthogonal) rotation. The Kaiser-

Meyer-Olkin measure of sampling adequacy was .87, which is above the commonly

Table 1

Descriptive Statistics for Study 1

Descriptive Category	Number of Participants	Percentage of Sample
Gender		
Male	204	25
Female	596	74
No Response	1	1
Age Range		
18-25	476	59
26-35	206	26
36-45	62	8
46 and Above	55	7
No Response	2	<1
Ethnicity		
African American	71	9
Asian	57	7
Caucasian	504	63
Hispanic	108	14
Native American	4	<1
Pacific Islander	18	2
Other	34	4
No Response	5	<1

recommended value of .60. Bartlett's test of sphericity was significant (X^2 (1225) =

14222.22, $p < .000$).

When investigating the scree plot (see Figure A1) for this analysis, there appears

to be five strong factors and several weaker factors. Similarly, when Eigenvalues are

considered there are ten factors identified with Eigenvalues greater than 1.00. Using the

Kaiser-Guttman rule (extracting based upon Eigenvalues greater than 1.00) as the

primary method of factor extraction, however, is commonly criticized for overestimating

the true number of underlying factors (Fabrigar et al., 1999). Given this, and considering

the theoretical foundation for the development of the IPIP-50 (detailed in Chapter 2),

only five factors were extracted for the current analysis.

When five factors were retained, 43.2% of the total variance for the entire set of variables was explained. Most of the items have a primary loading above .50 and only five items have a cross loading over .30. Perusal of Table B1 would suggest that when forced to load onto five factors, the majority of items on each scale group together on their expected factor. Since the factor labels proposed through the Five Factor Model of personality suited the extracted factors, they were therefore retained.

Confirmatory factor analysis. To determine the fit of the Five Factor Model of personality to the IPIP-50 data used in this study several different goodness of fit indicators were considered. First, the chi-square value was investigated and found to be significant (X^2 (1165) = 4518.905, $p < .000$). This would indicate rejection of the null hypothesis that the model was a good fit to the data; however, the relative chi-square value was 3.88, which Schumacker & Lomax (2004) deem acceptable. When considering the root mean square error of approximation (RMSEA), the value for this sample was .06, which meets the cutoff of .06 recommended by Hu and Bentler (1999). According to Byrne (1994), a model is regarded as acceptable if the Normed Fit Index (NFI) exceeds .90, the Goodness of Fit Index (GFI) exceeds .90, and the Comparative Fit Index (CFI) exceeds .93. For this condition the NFI, GFI, and CFI values were .69, .79, and .75, respectively.

Study 2 Results

Descriptive Statistics

The dataset for the random assignment study was comprised of a total of 423 cases. After listwise deletion of cases omitting a response to any item on the IPIP-50,

332 cases remained for analysis. Of these remaining cases, 145 were included in the 2-option condition and 186 were in the 5-option condition. Prior to analysis, cases within the two conditions were evaluated and considered for removal from analysis based upon completion time. Cases within the extremes on completion time (more than two standard deviations above and below the mean) for each condition were excluded from analysis. For the 5-option condition (M = 33, SD = 15), this resulted in the removal of six additional cases for a total sample of 180. For the 2-option condition (M = 33, SD = 13), this resulted in the removal of 5 additional cases for a total sample of 140. Average completion time did not differ between the two conditions. Table 2 contains a summary of the descriptive statistics between the two conditions.

2-Option condition. After all suspect cases were removed from analysis, the 2-option condition was comprised of 75% female (n = 105) and 25% male (n = 35) participants. Individuals falling within the age range of 18-25 made up 55.7% of the sample (n = 78), while 28.6% fell within the age range of 26-35 (n = 40), 12.1% were in the range of 36-45 (n = 17), and 3.6% fell within the range of 46 and above (n = 5).

Self-reported ethnic backgrounds were Caucasian (63.6%), Hispanic (13.6%), African-American (7.1%), Asian (7.9%), Other (5.7%), and Native American (0.7%). Ethnicity data were not provided by 1.4% of the sample.

A grade point average (GPA) of 3.5 and above was reported by 49.3% of the sample, a GPA of 3.0-3.4 was reported by 32.1%, and a GPA of 2.5-2.9 was reported by 17.1%. No one reported a GPA of 2.4 and below. Data on reported GPA was not available for 1.4% of the sample.

Table 2

Descriptive Statistics for Study 2

Descriptive Category	5-Option Condition		2-Option Condition	
Gender	*n*	%	*n*	%
Male	45	25.0	35	25.0
Female	135	75.0	105	75.0
No Response	0	0.0	0	0.0
Age Range				
18-25	121	67.2	78	55.7
26-35	39	21.7	40	28.6
36-45	10	5.6	17	12.1
46 and Above	9	5.0	5	3.6
No Response	1	0.6	0	0.0
Ethnicity				
African American	9	5.0	10	7.1
Asian	18	10.0	11	7.9
Caucasian	114	63.3	89	63.6
Hispanic	28	15.6	19	13.6
Native American	0	0.0	1	0.7
Pacific Islander	2	1.1	0	0.0
Other	8	4.4	8	5.7
No Response	1	0.6	2	1.4
GPA				
3.5 and Above	79	43.9	69	49.3
3.0-3.4	63	35.0	45	32.1
2.5-2.9	35	19.4	24	17.1
2.4 and Below	0	0.0	0	0.0
No Response	3	1.7	2	1.4

Note. n = number of participants in the sample choosing that option.

5-Option condition. The 5-option condition was comprised of 75% female (*n* =

135) and 25% male (*n* = 45) participants. Individuals falling within the age range of 18-

25 made up 67.2% of the sample (*n* = 121), while 21.7% fell within the age range of 26-

35 (*n* = 39), 5.6% were in the range of 36-45 (*n* = 10), and 5.0% fell within the age range

of 46 and above (*n* = 9). One participant did not provide information regarding age

(making up 0.6% of the sample).

Self-reported ethnic backgrounds were Caucasian (63.3%), Hispanic (15.6%), Asian (10.0%), African-American (5.0%), Other (4.4%), and Pacific Islander (1.1%). Ethnicity data were not provided by 0.6% of the sample.

A GPA of 3.5 and above was reported by 43.9% of the sample, a GPA of 3.0-3.4 was reported by 35.0%, and a GPA of 2.5-2.9 was reported by 19.4%. No one reported a GPA of 2.4 and below. Data on reported GPA was not available for 1.7% of the sample.

Validity Analysis

Exploratory factor analysis. All questions on the IPIP-50 were factor analyzed using principal components analysis with Varimax (orthogonal) rotation in both the 5-option and 2-option conditions. Within the 5-option condition, the Kaiser-Meyer-Olkin measure of sampling adequacy was .76 which is above the commonly recommended value of .60. Bartlett's test of sphericity was significant (X^2 (1225) = 3939.34, $p < .000$). Within the 2-option condition, the Kaiser-Meyer-Olkin measure of sampling adequacy was .66 which is above the commonly recommended value of .60. Bartlett's test of sphericity was significant (X^2 (1176) = 2748.51, $p < .000$). These results indicate adequate factorability of the data in both conditions.

When investigating the scree plots (see Figure A2 and Figure A3) for these analyses, there appears to be six fairly strong factors and several weaker factors in each condition. This is slightly different than the findings from Study 1. Once again, however, when Eigenvalues are considered for the 5-option condition and the 2-option condition, there are thirteen and fourteen factors, respectively, identified with Eigenvalues greater than 1.00. As stated above, the Kaiser-Guttman rule (extracting based upon Eigenvalues greater than 1.00) as the primary method of factor extraction,

however, is commonly criticized for overestimating the true number of underlying factors (Fabrigar et al., 1999). Given this, and considering the theoretical foundation for the development of the IPIP-50 (detailed in Chapter 2), only five factors were extracted in each condition for the current analysis.

Within the five-option condition, when five factors were retained 43.9% of the total variance for the entire set of variables was explained. Most items have a primary loading over .50 and eleven items have cross loadings above .30. The factor loading matrix for this condition is presented in Table B2. The majority of items had primary loadings on factors as predicted considering the Five-Factor Model underlying the IPIP-50. Consequently, the labels used to identify the scales on the IPIP-50 were retained.

Within the two-option condition, one item was removed from analysis because there was no variance in participant responding. With the remaining 49 items, when five factors were retained 39.7% of the total variance for the set of variables was explained. Most items have a primary loading over .50 and only four items have cross loadings above .30. Two items on the Agreeableness scale failed to load on any of the five factors. The factor-loading matrix for this condition is presented in Table B3. Again, the majority of items had primary loadings on the factor that would be expected based upon the Five-Factor Model; accordingly, the labels of the IPIP-50 were retained.

Perusal of Table B2 and Table B3 indicates that the items within each scale on the IPIP-50 tend to load on the same factor. This lends some support to the Five Factor Model upon which this scale of IPIP was based. It is interesting to note, however, that in both conditions there were a significant number of factors with initial eigenvalues over

1.00. In the 5-option condition there were thirteen factors meeting this criterion, and in the 2-option condition there were fourteen.

  Confirmatory factor analysis. To determine the fit of the Five Factor Model of personality to the IPIP-50 data used in this study several different goodness of fit indicators were again considered within each condition. First, for the 5-option condition, the chi-square value was investigated and found to be significant (X^2 (1165) = 2279.966, $p < .000$). This would indicate rejection of the null hypothesis that the model was a good fit to the data; however, the relative chi-square value was 1.96, which is considered acceptable (Kline, 1998; Ullman, 2001). When considering the root mean square error of approximation (RMSEA), the value for this sample was .07, which is greater than the cutoff of .06 recommended by Hu and Bentler (1999). According to Byrne (1994), a model is regarded as acceptable if the Normed Fit Index (NFI) exceeds .90, the Goodness of Fit Index (GFI) exceeds .90, and the Comparative Fit Index (CFI) exceeds .93. For this condition the NFI, GFI, and CFI values were .48, .65, and .65, respectively. These values indicate a poor fit of the Five Factor Model of personality to the data in the sample.

  Secondly, for the 2-option condition, the chi-square value was investigated and found to be significant (X^2 (1117) = 1786.366, $p < .000$). This would indicate rejection of the null hypothesis that the model was a good fit to the data; however, the relative chi-square value was 1.60, which is considered acceptable (Kline, 1998; Ullman, 2001). When considering the root mean square error of approximation (RMSEA), the value for this sample was .07, which is greater than the cutoff of .06 recommended by Hu and Bentler (1999). According to Byrne (1994), a model is regarded as acceptable if the

Normed Fit Index (NFI) exceeds .90, the Goodness of Fit Index (GFI) exceeds .90, and the Comparative Fit Index (CFI) exceeds .93. For this condition the NFI, GFI, and CFI values were .43, .67, and .66, respectively. These values indicate a poor fit of the Five Factor Model of personality to the data in the sample.

CHAPTER 5

DISCUSSION

Discussion of Study 1 Results

The findings of Study 1 provide some support for the factor structure of the IPIP-50. In this condition the five strongest factors found through exploratory factor analysis corresponded to the five factors of the Five Factor Model of personality as expected. In this study, however, the exploratory factor analysis did identify more than five factors. Although the remaining factors tended to be weak when compared to the first five, this is still an interesting and unexpected finding.

Results of the confirmatory factor analysis in this study support the notion that the IPIP-50 may be measuring more than just the five factors intended. In this condition, the null hypothesis that the Five Factor Model was a good fit for the data was largely rejected. All but two of the goodness of fit indices investigated did not meet the recommended thresholds to indicate a good fit of the model to the data. It would appear that while the five main factors found through exploratory factor analysis matched the Five Factor Model, the remaining weaker factors impacted the fit of this data to the model when investigated through confirmatory factor analysis. Further investigation of the items loading on the weaker factors found through exploratory factor analysis could help to shed light on the issue of what factors are really being measured by the IPIP-50.

It is important to note, however, that this pattern of results is consistent with other research conducted with this scale (see Chapter 2). Previous studies found less support for the Five Factor Model through CFA than EFA.

Discussion of Study 2 Results

Just as in Study 1, results of the exploratory factor analysis in Study 2 provide some support for the Five-Factor Model underlying the development of the IPIP-50. With the exception of only a few items, the items that were identified on the Extraversion scale tended to load on one factor, items on the Agreeableness scale tended to load on a separate factor, and the same was found for items on the Conscientiousness, Neuroticism, and Openness scales as well. This pattern was found in both the five-option condition as well as the two-option condition. In this way, it does not appear that the reduction in response options negatively impacted the factor structure of the IPIP-50 with this particular sample. Indeed, there were fewer items with cross loadings greater than .30 when the data from the 2-option condition were analyzed. However, it is interesting to note that Item 43 had to be removed from analysis in the 2-option condition. All participants selected the "Accurate" response to that item.

Results of the confirmatory factor analysis paint a different picture regarding the factor structure of this scale. Goodness of fit indices in both conditions supports the notion that the IPIP-50 may be measuring more than just the five factors intended. In both the 5-option and the 2-option conditions, the null hypothesis that the five-factor model was a good fit for the data was rejected. This could be due to the relatively small sample size in each condition. While there are differing opinions regarding the ideal sample size for conducting factor analysis (Merenda, 1997; Fabrigar et al., 1999; MacCullum, Widaman, Zhang, & Hong, 1999), it is commonly accepted that some measures of model fit used in confirmatory factor analysis tend to underestimate the fit of the model when sample sizes are small (Ullman, 2001).

It is important to note that the same pattern of results was found for both the 5-option condition and the 2-option condition. The same factors were evident in the exploratory factor analysis for each condition, and the same likelihood of rejecting the null hypothesis was found when confirmatory factor analysis results were analyzed in each condition. In this way, the potential problems with the factor structure indicated through these analyses do not seem to be unique to one condition over the other. This important point lends some support to the notion that reducing the number of available response options on an affective measurement scale is not necessarily detrimental to the psychometric properties of that scale. Again, further investigation of the items loading on the weaker factors found through exploratory factor analysis could help to shed light on the issue of what factors are really being measured by the IPIP-50.

Implications for the Field

The findings of the studies included in this dissertation can be applied to the practical realm of school psychology and psychological assessment. As discussed in previous chapters, school psychologists often rely upon affective measures to obtain information about the students whom they serve. The teachers, parents, and students who are often asked to provide this information have other demands placed upon their time and do not always have the amount of time that long questionnaires with many response options for each question require. Shortening questionnaires by reducing the number of response options provided for each option has not typically been practiced as the common belief and understanding is that to do so would be detrimental to the psychometric properties of the scale.

The findings of the presented line of research in this dissertation do not definitively support this common notion. As far as construct validity goes, no significant changes to the factor structure of the IPIP-50 were observed when 5-response options were provided as opposed to when only 2-response options were provided. Additionally, there do not appear to be drastic differences in the factor structure of the IPIP-50 between the samples comprising the databases used for the two studies.

Given the limitations addressed above, these studies are not sufficient to instigate a complete overhaul in the construction of affective measurement scales. The data presented, however, do add another bit of support to the notion that the inclusion of more response options is not inherently the better choice. Further research addressing the limitations of this dissertation and more studies adapting other affective measures for administration with a decreased number of response options is needed.

Cautions for Generalization

The generalizability of this sample to other participant samples is cautioned. This is a common caution in research studies, and is particularly relevant to the studies presented here. The samples comprising the databases used for analysis in both Study 1 and Study 2 consisted solely of university students who participated in the research to fulfil a class requirement. There is potential with this study, as with all subject pool studies, for participants to have suspect motivation (Jones and Loe, 2013). As previously detailed, the subjects participated in this study as part of a course requirement. As such, participants may have been tempted to complete the study materials without giving careful consideration to their responses in an attempt to finish quickly. Study 2 addressed

this limitation by removing from the database subjects who completed the study materials in significantly less time than the other study participants. This was not able to be addressed in Study 1 as completion time for each participant was not reported.

In addition, with regards to Study 2, the computer software that was used to randomly assign participants to each condition limited the investigation. The program used only allowed for random assignment between two conditions, resulting in a database allowing for limited comparisons to be made. Due to this, investigation of the impact on the factor structure of the scale when response options are reduced but a midpoint is maintained (ie. going from 5-options to 3-options) could not be conducted. Future research with more diverse participant samples, and additional studies using random assignment of participants to investigate more varied response option conditions, is needed.

Directions for Future Research

More Diverse Samples

As alluded to above, future research in this area could be aimed at replicating these results with more diverse participant samples. The databases used for this dissertation were drawn from college students participating as a class requirement and in all conditions were composed of predominantly white, female participants. While this is representative of the population of education majors in the institution from which the sample was drawn, it is not representative of the greater population. Research utilizing samples including a greater number of male participants and individuals from more

diverse ethnic and collegiate backgrounds would add to the research and understanding in this area.

Random Assignment into Additional Groups

Additionally, future studies utilizing random assignment to more than just the two conditions investigated in Study 2 detailed in this dissertation would greatly add to the body of literature in this area. Study 2 was limited to simply investigating differences in the factor structure between a 5-option condition and a 2-option condition. This does not address the effects on responding when response options are reduced, but a midpoint retained. Future investigations including a randomly assigned group completing the IPIP-50 with only 3 response options would be an interesting addition to the research base.

Dependent Sample Investigation

Another interesting variation to the procedure, and possible avenue for future research, would be to use a single participant sample to directly investigate differences in responding between differing response-option conditions. To accomplish this, each participant would serve as their own control group from which to compare responding as the number of response options is varied. For example, participants would be randomly assigned to first complete the IPIP-50 with 5-response options, and then complete it with only 2-response options, or vice versa. This would allow for direct investigation of the impact on the factor structure of the IPIP-50 when differing numbers of response options are provided. This would provide an interesting perspective, but care would need to be taken in the design of the study and interpretation of results as practice effects may be evident from repeated exposure to the same measure.

Relationships between the IPIP-50 and Related Measures

In addition to the above possibilities for future research, there is the potential to investigate the criterion validity of the IPIP-50 by investigating the relationship between scores on the IPIP-50 and related outside measures. One potential example is the CogStyle. CogStyle is a measure of vocational personality traits providing direct assessment of the six vocational personality traits identified by Holland (1973, 1997): Realistic, Investigative, Artistic, Social, Enterprising and Conventional. The scale uses paired comparisons of the following six adjectives to represent each of the vocational personality traits: practical, curious, flexible, sympathetic, ambitious, and efficient. The scale then uses paired comparisons of the following six action verbs to represent each of the six traits: fixing things, investigating things, designing things, assisting others, persuading others, and organizing things. Previous research (Jones & Loe, 2013) indicates split-half reliability estimates of .72, .64, .53, .62, .69, and .77 for the Realistic, Investigative, Artistic, Social, Enterprising and Conventional scales, respectively.

The six Holland vocational personality traits (measured by the CogStyle) have been found through a large-scale meta-analysis to have a predictable relationship between certain of the big five personality traits (which are measured by the IPIP-50) (Larson, Rottinghaus, & Borgen, 2002). Through their analysis the researchers found consistent positive relationships between the following traits:

- Big five Extraversion trait and Social vocational personality trait
- Big five Extraversion trait and Enterprising vocational personality trait
- Big five Agreeableness trait and Social vocational personality trait

- Big five Conscientiousness trait and Conventional vocational personality trait

- Big five Openness trait and Artistic vocational personality trait

- Big five Openness trait and Investigative vocational personality trait

Research using this, or another similar instrument, to investigate the established relationships between the Five Factor Model (as evidenced by scores on the IPIP-50) and the Holland vocational personality traits would greatly add to the body of research validating the IPIP-50.

Impact of Reduced Response Options on Other Psychometric Properties

Another area requiring further investigation that was somewhat touched upon in this dissertation is the impact on the psychometric properties of the IPIP-50 when the number of response options is reduced. Study 2 looked at the impact on the factor structure of the IPIP-50 when response options are reduced, but it would be interesting to investigate the impact that doing so has on the psychometric properties other than construct validity (such as reliability and other forms of validity) of the scale as well.

As addressed in Chapter Two, an individual's profile on the MMPI is considered to be valid when they have responded consistently and accurately to the items on the measure, without distorting their responses in an identifiable way (Greene, 2000). This same concept of validity can be applied to the IPIP-50 in future investigations. Since validity scales have not been theoretically and empirically built into the IPIP-50 the same way that they were for the MMPI, the consistency of individual responding will need to be separately analyzed for the participants in this study. Johnson (2005) suggests a

method that was proposed by Douglas Jackson for doing just this. Johnson (2005) describes this method in the following terms:

> In Jackson's method, items within each of the standard scales are numbered sequentially in the order in which they appear in the inventory and then divided into odd-numbered and even-numbered subsets. Scores are computed for the half-scale subsets, a product moment correlation is computed between the odd- and even-numbered half-scale scores across all scales, and corrected for decreased length by the Spearman–Brown formula. Jackson refers to this value as an "individual reliability" coefficient. (p. 110)

In this way, each individual protocol could be evaluated and checked for consistency in responding. Following the method used when interpreting the MMPI, IPIP-50 protocols exhibiting higher levels of consistency would be interpreted as being more valid representations of the individual. Further, future investigations could look at the difference in consistency of protocols between different response-option conditions to see if a decrease in available response options on the IPIP-50 results in any significant change in the proposed internal validity indices.

Using consistency of responding as an indicator of validity for the IPIP-50 makes sense as an individual could reasonably be expected to respond consistently to items assessing the same personality construct. Consistency and reliability in responding to items on a scale, however, regardless of the number of response options provided, does not inherently equate to accuracy. An individual can be consistent in rating themselves high on any given personality trait, but not exemplify that trait at all in their day-to-day life. Also, when response options are reduced an individual may not feel like they have

enough options to accurately portray themselves. Future research investigating not only the impact on the reliability and validity of a scale when response options are reduced, but also the impact on the respondent's ability to accurately portray themselves would be interesting.

Conclusion

The investigations presented in this dissertation looked at the factor structure of the IPIP-50 as well as the impact on that factor structure when response options were reduced from five to two. Reducing the number of response options provided to a respondent on an affective measure may prove advantageous as it may effectively decrease the amount of time and effort required by the respondent to complete the measure. This reduction in time and effort required for completion could potentially increase the practical utility of such affective rating scales within the realm of psychological assessment and, more specifically, the field of school psychology.

It is of great importance, however, that affective measures retain their psychometric properties when the number of response options is reduced in order for them to retain their practical utility. It does not matter if a questionnaire takes only half the time to complete if it is no longer producing valid results. Given these considerations, this dissertation investigated not only the general factor structure, of the IPIP-50, an open source personality assessment, but also the impact on that factor structure when the number of available response options was reduced from five to two.

Exploratory factor analyses provided some preliminary evidence of the Five Factor Model underlying the IPIP-50. In Study 1, as well as in both conditions of Study

2, EFA results indicated the presence of five factors. In all conditions, items on the same subscale of the instrument tended to group together on the same factor. Confirmatory factor analyses, however, painted a slightly different picture. Goodness of fit indices in both Study 1, as well as in both conditions of Study 2, supported rejection of the null hypothesis that the Five Factor Model was a good fit for the data. Further research is clearly needed in this area, and possible avenues to continue this investigation were discussed.

APPENDIX A

Figure A1. Scree plot of factor eigenvalues in Study 1.
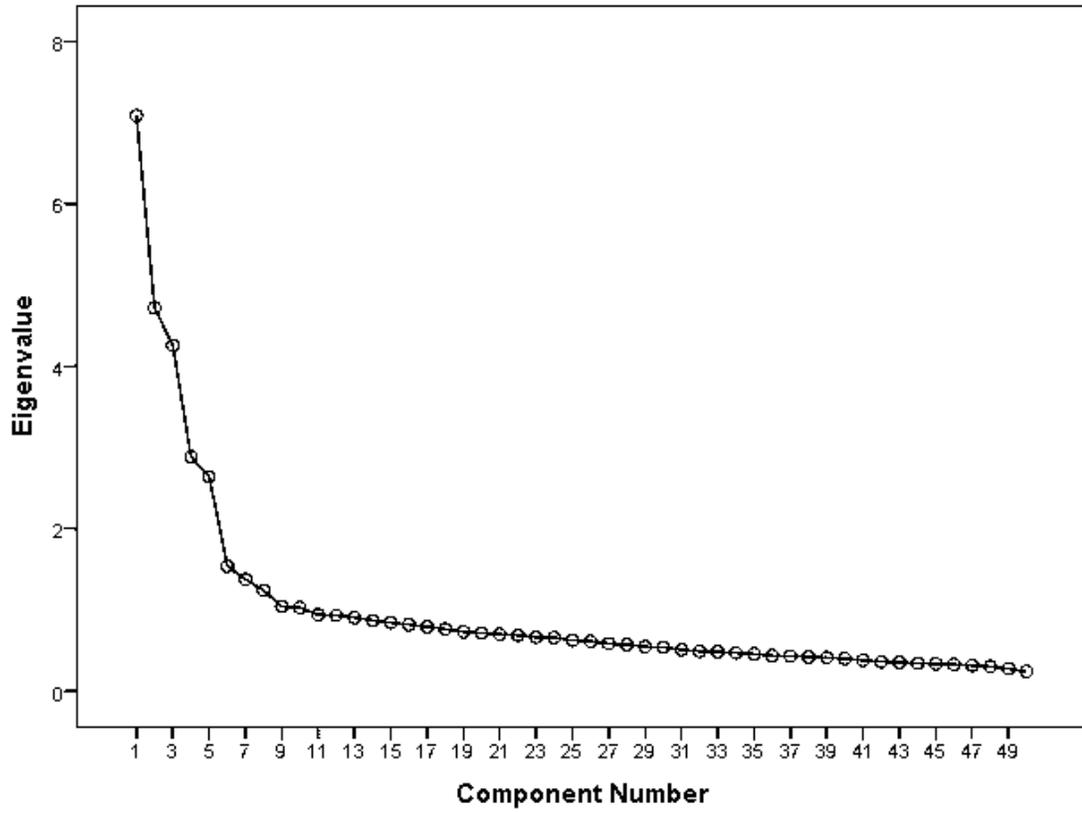
Figure A2. Scree plot of factor eigenvalues in Study 2, 5-option condition.
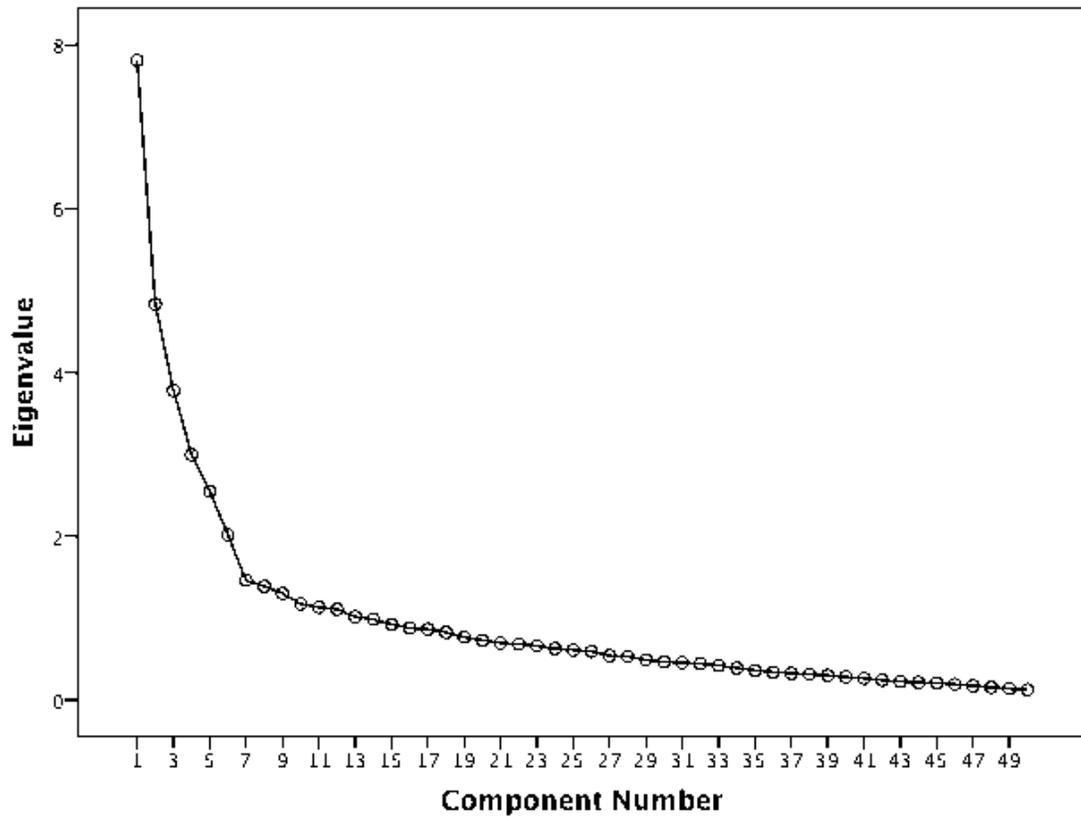
Figure A3. Scree plot of factor eigenvalues in Study 2, 2-option condition.
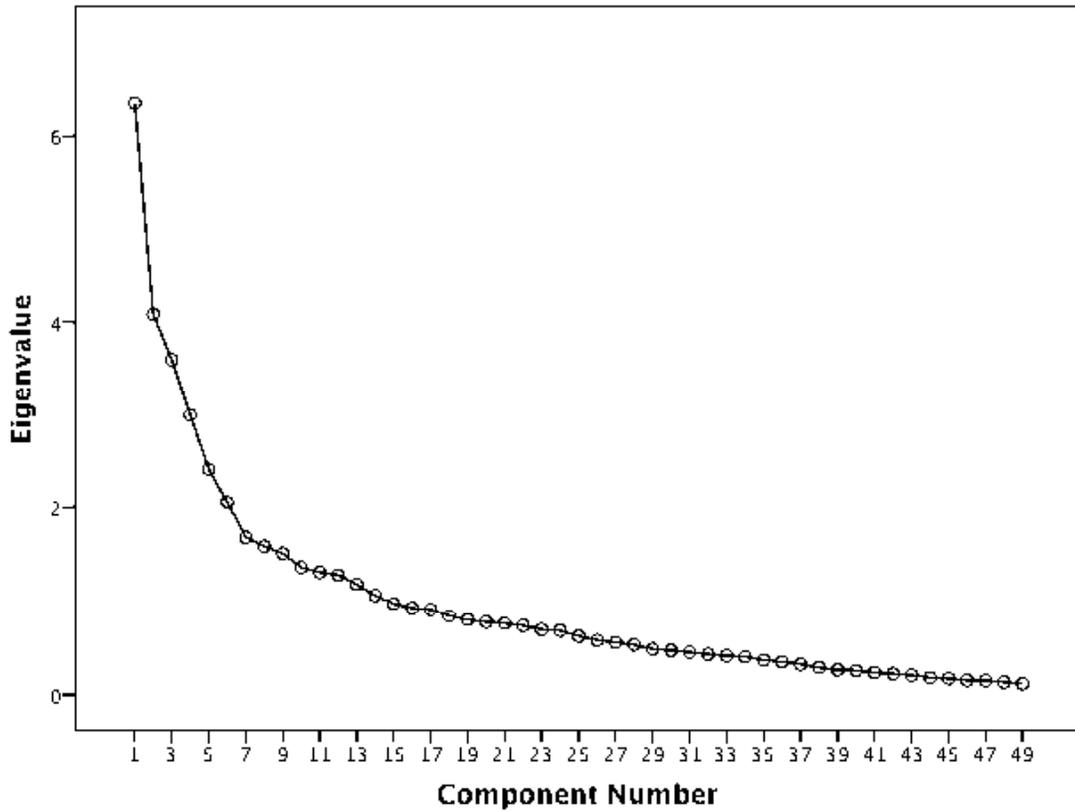
Table B1

Factor Loadings of the International Personality Item Pool (IPIP-50) in Study 1 (N=801)

Item	E	N	C	A	O
Q1	.62				
Q6	.72				
Q11	.62				
Q16	.73				
Q21	.72				
Q26	.49				
Q31	.77				
Q36	.64				
Q41	.72				
Q46	.74				
Q2				.44	
Q7	.34			.47	
Q12				.40	
Q17				.77	
Q22				.50	
Q27				.63	
Q32				.61	
Q37				.58	
Q42				.69	
Q47	.36			.43	
Q3			.65		
Q8			.65		
Q13			.47		.32
Q18		.31	.62		
Q23			.68		
Q28			.67		
Q33			.55		
Q38			.47		
Q43			.65		
Q48			.51		
Q4		.73			
Q9		.58			
Q14		.67			
Q19		.46			
Q24		.56			
Q29		.77			
Q34		.71			
Q39		.72			
Q44		.73			
Q49		.61			
Q5					.62
Q10					.65
Q15					.56
Q20					.50
Q25					.60
Q30					.42
Q35					.53
Q40					.65
Q45				.31	.31
Q50					.68

Note. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism. Extraction Method: Principal Component Analysis. 5 components extracted. Factor loadings below .30 have been suppressed.

Table B2

Factor Loadings of the International Personality Item Pool (IPIP-50) in Study 2 with 5 Response Options (N=180)

Item	E	N	A	C	O
Q1	.69				
Q6	.65				
Q11	.52		.38		
Q16	.72				
Q21	.70		.30		
Q26	.41				.33
Q31	.78				
Q36	.54		-.42		
Q41	.69				
Q46	.71				
Q2			.40		
Q7	.44		.43		
Q12			.31		
Q17			.70		
Q22			.40		
Q27			.56		
Q32			.44		
Q37			.67		
Q42			.69		
Q47	.47		.49		
Q3				.60	
Q8				.62	
Q13				.49	.30
Q18				.69	
Q23				.66	
Q28				.66	
Q33				.58	
Q38				.49	
Q43			.31	.52	
Q48			.36	.52	
Q4		.65			
Q9		.55			
Q14		.55	-.38		
Q19		.52			
Q24		.47			
Q29		.64			
Q34		.74			
Q39		.76			
Q44		.74			
Q49		.63			
Q5					.55
Q10					.62
Q15					.51
Q20					.55
Q25					.50
Q30					.54
Q35					.51
Q40					.64
Q45			.31		.34
Q50					.64

Note. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism. Extraction Method: Principal Component Analysis. 5 components extracted. Factor loadings below .30 have been suppressed.

Table B3

Factor Loadings of the International Personality Item Pool (IPIP-50) in Study 2 with 2 Response Options (N=140)

Item	N	E	C	A	O
Q1		.59			
Q6		.68			
Q11	.38	.57			
Q16		.79			
Q21		.71			
Q26		.55			
Q31		.68			
Q36		.50			
Q41		.66			
Q46		.69			
Q2					
Q7				.62	
Q12					
Q17				.68	
Q22				.50	
Q27				.51	
Q32				.47	
Q37				.60	
Q42				.60	
Q47				.34	
Q3			.68		
Q8			.60		
Q13			.45		.37
Q18			.58		
Q23			.68		
Q28			.68		
Q33			.37		
Q38			.46		
Q48			.64		
Q4	.72				
Q9	.76				
Q14	.62				
Q19	.40				
Q24	.68				
Q29	.80				
Q34	.55				
Q39	.72				
Q44	.75				
Q49	.58				
Q5					.56
Q10					.49
Q15					.57
Q20					.41
Q25					.48
Q30					.37
Q35					.35
Q40					.46
Q45		-.39		.31	.35
Q50					.62

Note. E = Extraversion; A = Agreeableness; C = Conscientiousness; N = Neuroticism. Extraction Method: Principal Component Analysis. 5 components extracted. Factor loadings below .30 have been suppressed.

REFERENCES

Albaum, G., Wiley, J., Roster, C., & Smith, S.M. (2011). Visiting item non-response in

    internet survey data collection. *International Journal of Market Research, 53*(5),

    687-703.

Altmann, R. (2004). Response format types for student self-report inventories: A

    comparative study. Poster session presented at the annual meeting of the

    American Psychological Association, Honolulu, HI.

Anastasi, A, & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River,

    NJ: Prentice Hall.

Baker, S.R., Victor, J.B., Chambers, A.L., & Halverson, C.F. (2004). Adolescent

    personality: A Five-Factor Model construct validation. *Assessment, 11*(4), 303-

    315.

Baumgartner, H., & Steenkamp, J.E.M. (2001). Response styles in marketing research:

    A cross-national investigation. *Journal of Marketing Research, 38*(2), 143-156.

Borgers, N., Hox, J., & Sikkel, D. (2004). Response effects in surveys on children and

    adolescents: The effect of number of response options, negative wording, and

    neutral mid-point. *Quality and Quantity: International Journal of Methodology,*

    38(1), 17-33.

Brace, I. (2004). *Questionnaire design: How to plan, structure and write survey*

    material for effective market research. London, UK: Kogan Page.

Butcher, J.N, Dahlstrom, W.G., Graham, J.R., Tellegen, A., & Kaemmer, B. (1989).

    MMPI-2: Manual for administration and scoring. Minneapolis, MN: University

    of Minnesota Press.

Byrne, B. M. (1994). Structural equation modeling with EQS and EQS/Windows.

    Thousand Oaks, CA: Sage Publications.

Churchill, G.A., & Peter, J.P. (1984). Research design effects on the reliability of rating

    scales: A meta-analysis. *Journal of Marketing Research, 21*, 360-375.

Cooper, A.J., Smillie, L.D., & Corr, P.J. (2010). A confirmatory factor analysis of the

    Mini-IPIP five-factor model personality scale. *Personality and Individual*

    Differences, 48(5), 688-691.

Costa, P.T., & McCrae, R.R. (1992). *Revised NEO Personality Inventory (NEO PI-R)*

    and NEO Five-Factor Inventory (NEO-FFI) professional manual. Odessa, FL:

    Psychological Assessment Resources.

Courtenay, B.C., & Weidemann, C. (1985). The effects of a don't know response on

    Palmore's facts on aging quizzes. *The Gerontologist, 25*, 177-181.

Cox, E.P. (1980). The optimal number of response alternatives for a scale: A review.

    Journal of Marketing Research, *17*, 407-422.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests.

    Psychometrica, 16, 93–96.

Dillman, D. (2000). *Mail and internet surveys: The tailored design method.* New York,

    NY: John Wiley & Sons, Inc.

Fabrigar, L., Wegener, D., MacCallum, R.C., & Strahan, E. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272-299.

Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 363-373.

Fink, A. (2003). *How to ask survey questions.* Thousand Oaks, CA: Sage.

Foddy, W. (1993). *Constructing questions for interviews and questionnaires: Theory and practice in social research*. Cambridge, UK: Cambridge University Press.

Forbey, J.D., & Ben-Porath, Y.S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI–2 computerized adaptive version. *Psychological Assessment, 19,* 14-24.

Goldberg, L. R. (1992). The development of markers for the big-five factor structure. *Psychological Assessment, 4*, 26-42.

Goldberg, L.R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In: I. Mervielde, I. Deary, F. De Fruyt, F. Ostendorf (Eds.), *Personality psychology in Europe, Vol. 7* (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press. (1999), pp. 7–28.

Goldberg, L.R., Johnson, J.A., Eber, H.W., Hogan, R., Ashton, M.C., Cloninger, C.R., & Gough, H.C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.

Gow, A.J., Whiteman, M.C., Pattie, A., & Deary, I.J. (2005). Goldberg's 'IPIP' Big-Five factor markers: Internal consistency and concurrent validation in Scotland. *Personality and Individual Differences, 39*(2), 317-329.

Greene, R.L. (2000). *The MMPI-2: An interpretive manual* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Hathaway, S.R., & McKinley, J.C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Clinical Psychology, 10,* 249-254.

Helgeson, J.G, Voss, K.E., & Terpening, W.D. (2002). Determinants of mail-survey response: Survey design factors and respondent factors. *Psychology & Marketing, 19*(3), 303-328.

Hills, M. (1977). A review of "An introduction to multivariate techniques for social and behavioral sciences." *Applied Statistics, 26*, 339-340.

Hughes, L., Jones, W.P., & Loe, S.A. (2011). *Measuring reading speed and comprehension online*. Poster presented at the annual convention of the National Association of School Psychologists, San Francisco, CA.

Holland, J. L. (1973). *Making vocational choices*. Englewood Cliffs, NJ: Prentice Hall.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment Resources.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

International personality item pool: A scientific collaboratory for the development of

    advanced measures of personality traits and other individual differences

    (http://ipip.ori.org/). Internet Web Site.

Jacoby, J., & Matell, M.S. (1971). Three-point Likert scales are good enough. *Journal*

    of Marketing Research, 8, 495-500.

Johnson, J.A. (2005). Ascertaining the validity of individual protocols from Web-based

    personality inventories. *Journal of Research in Personality, 39*, 103-129.

Jones, W.P., & Loe, S.A. (2013). Optimal number of questionnaire response categories:

    More may not be better. *SAGE Open.* doi: 10.1177/2158244013489691

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New

    York, NY: Guilford Press.

Kovacs, M. (1992). *Children's Depression Inventory manual.* North Tonawanda, NY:

    Multi-Health Systems, Inc.

Krosnick J.A. (1991). Response strategies for coping with the cognitive demands of

    attitude measurement in surveys. *Applied Cognitive Psychology, 5*(2), 213–236.

Larson, L.M., Rottinghaus, P.J., & Borgen, F.H. (2002). Meta-analyses of big six

    interests and big five personality factors. *Journal of Vocational Behavior, 61*,

    217-239.

Lietz, P. (2010). Research into questionnaire design: A summary of the literature.

    International Journal of Market Research, 52(2), 249-272).

MacCullum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor

    analysis. *Psychological Methods, 4*, 84-89.

Madden, T.M., & Klopfer, F.J. (1978). The "cannot decide" option in Thurstone-type

    attitude scales. *Educational and Psychological Measurement, 38*, 259-264.

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

    (2009). The effect of varying the number of response alternatives in rating scales:

    Experimental evidence from intra-individual effects. *Behavior Research*

    Methods, 41(2), 295-308.

Merenda, P.F. (1997). A guide to the proper use of factor analysis in the conduct and

    reporting of research: Pitfalls to avoid. *Measurement and Evaluation in*

    Counseling and Development, 30(3), 156-164.

Miller, M.D., Linn, R.L., & Gronlund, N.E. (2009). *Measurement and assessment in*

    teaching (10th ed.). Upper Saddle River, NJ: Pearson.

Murphy, G., & Likert, R. (I938). *Public opinion and the individual.* New York, NY:

    Harper and Brothers Publishers.

Overall, J.E. (1964). Note on the scientific status of factors. *Psychological Bulletin, 61*,

    270-276.

Popham, W.J. (2011). *Classroom assessment: What teachers need to know* (6th ed.).

    Upper Saddle River, NJ: Pearson.

Rappaport, R.B. (1982). Sex differences in attitude expression: A general explanation.

    Public Opinion Quarterly, 57, 305-313.

Reynolds, C.R., & Kamphaus, R.W. (2004). *Behavior Assessment System for Children*

    manual (2nd ed.). Circle Pines, MN: AGS Publishing.

Reynolds, C.R., & Richmond, B.O. (2008). *Revised Children's Manifest Anxiety Scale*

    manual (2nd ed.). Los Angeles, CA: Western Psychological Services.

Saris, W.E. & Gallhofer, I. (2007). Estimation of the effects of measurement characteristics on the quality of survey questions. *Survey Research Methods, 1*(1), 29–43.

Schmidt, F.L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*, 206–224.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Siefert, C.J., Stein, M., Sinclair, S.J., Antonius, D., Shiva, A., & Blais, M.A. (2012). Development and initial validation of a scale for detecting inconsistent responding on the Personality Assessment Inventory–Short Form. *Journal of Personality Assessment, 94,* 601-606.

Singer, E., Couper, M.P., Raghunathan, T.E., Antonucci, T.C., Burmeister, M., & Van Hoewyk, J. (2010). The effect of question framing and response options on the relationship between racial attitudes and beliefs about genes as causes of behavior. *Public Opinion Quarterly, 74*(3), 460-476.

Skinner, H.A. (1980). Factor analysis and studies of alcohol. *Journal of Studies on Alcohol, 41*, 1091-1101.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201-293.

Spearman, C. (1927). *The abilities of man*. New York, NY: Macmillan.

Thompsan, B, & Daniel, L.G. (1996). Factor analytic evidence for the construct validity

of scores: A historical overview and some guidelines. *Educational and*

Psychological Measurement, 56(2), 197-208.

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell

(2001). *Using Multivariate Statistics* (4th ed; pp 653- 771). Needham Heights,

MA: Allyn & Bacon.

Weems, G.H., & Onwuegbuzie, A.J. (2001). The impact of midpoint responses and

reverse coding on survey data. *Measurement and Evaluation in Counseling and*

Development, 34(3), 166-176.

Leslie Forbush
HUGHESL5@UNLV.NEVADA.EDU

Education
Doctor of Philosophy: Educational Psychology, School Psychology Specialization Strand
University of Nevada, Las Vegas
December 2014
Dissertation Title: Validation of a Personality Scale: A Factor Analytic Approach

Master of Science: Educational Psychology
University of Nevada, Las Vegas
August 2009

Bachelor of Arts: Psychology (*magna cum laude)*
University of Nevada, Las Vegas
May 2008

Practical and Professional Experiences in School Psychology

August 2014 – Present
: School Psychologist
Clark County School District
Supervisor: Dr. Don Blagg
Duties: Conduct psychoeducational assessments; prepare multidisciplinary team reports; lead multidisciplinary team meetings; participate in the special education eligibility determination and reevaluation process; consult with teachers and teaching staff regarding the psychological, behavioral, and educational needs of students; participate in the Response to Intervention process.

August 2013 – June 2014
: Internship in School Psychology
Clark County School District
Supervisors: Dr. Don Blagg, Dr. Marie Schrader
Duties: Conduct psychoeducational assessments; prepare multidisciplinary team reports; conduct multidisciplinary team meetings; participate in the special education eligibility determination and reevaluation process; consult with teachers and teaching staff regarding the psychological and educational needs of students; participate in the Response to Intervention process.

April 2012 – March 2014	Advanced Clinical Practicum and Internship Experience Dr. W. Paul Jones, Ltd. Supervisor: Dr. Paul Jones Duties: Conduct background interviews and mental status exams with individuals applying for Social Security Disability Benefits.
August 2012 – May 2013	Advanced Doctoral Practicum in School Psychology University of Nevada, Las Vegas Supervisor: Dr. Paul Jones Duties: Conduct comprehensive psychoeducational assessments; prepare psychological reports integrating assessment data; conduct intake and feedback sessions with clients.
August 2009 – May 2010	Practicum Experience in School Psychology University of Nevada, Las Vegas – partnered with Clark County School District Supervisors: Dr. Rhiannon Rager, Dr. Michelle Nathan Duties: Administer cognitive, academic, and affective assessments; participate in the Response to Intervention process; conduct classroom observations; participate in multidisciplinary team meetings to determine initial and continued eligibility for special education services; conduct small study skills group counselling sessions.

Professional Teaching and Research Experience

August 2011 – May 2013	Part-time Instructor Graduate Assistant, EPY 451 Foundations of Educational Assessment, and EPY 452 Counseling/Consultation Skills for Classroom Teachers University of Nevada, Las Vegas Supervisor: Dr. Paul Jones Duties: Prepare lectures and instruct education students at the undergraduate level on the fundamentals of educational assessment and basic counseling skills with a focus on practical classroom applications.
August 2009 – May 2013	Part-time Instructor Graduate Assistant, EPY786 Applied Assessment in Educational and School Psychology University of Nevada, Las Vegas Supervisor: Dr. Paul Jones Duties: Prepare lessons and instruct school psychology students in administration, scoring, and interpretation of standardized intelligence and academic achievement tests.

August 2008 – August 2009   Project Manager, Empowerment Schools Study and UNLV
                            Center for Evaluation and Assessment
                            University of Nevada, Las Vegas
                            Supervisor: Dr. Ralph Reynolds
                            Duties: Assist in the analysis of interview transcripts;
                            engage in scheduling of data collection; ensure general
                            organization of projects; oversee administration of online
                            course evaluations; prepare and distribute evaluation
                            reports to the proper departments on campus.

August – December 2008   Teaching Assistant, EPY757 History and Philosophy of
                            Educational Psychology
                            University of Nevada, Las Vegas
                            Supervisor: Dr. Ralph Reynolds
                            Duties: Attend class sessions; organize student
                            assignments; aide in general instructor preparation for
                            class.

Presentations and Publications
Parriott, D., Jones, W.P., Loe, S.A., **Hughes, L.**, (2012, November). *Adapting neuropsychological function scales for online administration: Study 2*. Poster presented, Annual meeting of the National Academy of Neuropsychology, Nashville, TN.

Hughes, L., Jones, W.P., & Loe, S.A. (2012, March). *Adapting reasoning scale for online administration: Pilot study*. Poster presented, Annual conference of the American College of Professional Neuropsychology, Las Vegas, NV.

Hughes, L., Bjork, K.A., Jones, W.P., & Loe, S.A. (2011, March). *Measuring reading speed and comprehension online*. Poster presented, 2011 University of Nevada, Las Vegas Graduate and Professional Student Research Forum, Las Vegas, NV.

Hughes, L., Bjork, K.A., Jones, W.P., & Loe, S.A. (2011, February). *Measuring reading speed and comprehension online*. Poster presented, Annual convention of the National Association of School Psychologists, San Francisco, CA.

Bjork, K.A., **Hughes, L.**, Jones, W.P., & Loe, S.A. (2011, February). *Conscientiousness: A predictor of online reading performance?* Poster presented, Annual convention of the National Association of School Psychologists, San Francisco, CA.

Reynolds, R. E. & the Empowerment Research Team* (2009, November). *The influence of school culture on teaching and learning: Challenges and opportunities.* Paper presented, Annual meeting of the Southwest Consortium of Innovative Psychology in Education, Las Vegas, NV. (*Hall, G., Marchand, G., Parker, R., Putney, L., Bjork, K., Hamilton, N., **Hughes, L.**, Khilnani, T., Kuch, F., & Surwill, V.).

Honors, Awards, and Memberships
Recipient, Dean's Doctoral Fellowship for Engaged Scholarship and Research,
University of Nevada, Las Vegas, College of Education (2012-2013)
Member, National Association of School Psychologists (2008-present)
Student Leader, National Association of School Psychologists (2009-2011)

References
References available upon request.