

An Investigation into the Applicability of the NEO-PI-R PPM Research Validity Scale to the  
NEO-PI-3 and the Effects of Positive Response Distortion

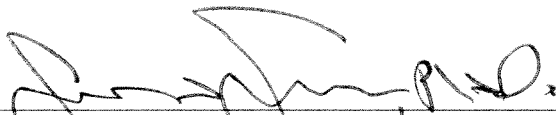
Jolene Young, M.A.

A Clinical Research Project presented to the faculty of the Hawai'i School of Professional  
Psychology at Argosy University, Hawai'i in partial fulfillment of the requirements for the  
degree of Doctor of Psychology in Clinical Psychology.

Honolulu, Hawaii  
April, 2018

An Investigation into the Applicability of the NEO-PI-R PPM Research Validity Scale to the  
NEO-PI-3 and the Effects of Positive Response Distortion


This Clinical Research Project by Jolene Young directed and approved by the candidate's  
Clinical Research Project Committee, was approved by the faculty of the Hawai'i School of  
Professional Psychology at Argosy University, Hawai'i in partial fulfillment of the requirements  
of the degree of Doctor of Psychology in Clinical Psychology.



---


Sean W. Scanlan, PhD, Program Dean  
College of Clinical Psychology

Clinical Research Project Committee



---

Lianne T.S. Philhower, PsyD, MPH  
Chair



---

Lawrie A. Ignacio, PsyD  
Committee Member

April 9, 2018  
Date

© copyright 2018

by

Jolene Young

All Rights Reserved

An Investigation into the Applicability of the NEO-PI-R PPM Research Validity Scale to the NEO-PI-3 and the Effects of Positive Response Distortion

Jolene Young

Hawaii School of Professional Psychology at Argosy University, Hawaii - 2018

This clinical research project investigated the applicability of the Revised NEO Personality Inventory (NEO-PI-R; McCrae & Costa, 2010) Positive Presentation Management Scale (PPM) developed by Schinka, Kinder, and Kremer (1997) to the NEO Personality Inventory-3 (NEO-PI-3), using the Minnesota Multiphasic Personality Inventory – 2 (MMPI-2; Butcher et al., 2001) Validity Scales as a criterion. The sample was extracted from an archival database composed of 303 airline pilot candidates who completed both the MMPI-2 and the NEO-PI-3 as part of their hiring process. Results indicated weak reliability of the PPM research validity scale. However, the PPM scale correlated significantly with the MMPI-2 Validity Scales in the expected directions and significantly added to the MMPI-2 Validity Scales in discriminating between underreporting and valid profiles. The percent and diagnostic agreement of positive response distortion between the two measures were fair. Therefore, this research project provided additional support for the validity and utility of the PPM research validity scale as an indicator of positive response distortion. Finally, applicants who demonstrated positive response distortion on the PPM research validity scale produced significantly different clinical profiles on the NEO-PI-3 and the MMPI-2 than applicants with valid profiles. The clinical implications and limitations of the research are also discussed. In conclusion, developing a psychometrically sound PPM research validity scale is important, as the use of validity scales assists mental health professionals in accurately obtaining information to make informed clinical decisions.

*Keywords:* Positive Response Distortion, Positive Presentation Management, Underreporting, Concurrent Validity, NEO-PI-3, MMPI-2

## Acknowledgements

After four years of intensive learning, professional development, and personal growth, I am proud to put forth my clinical research project (CRP). Thank you very much to Dr. Lianne T. S. Philhower, my CRP chair and a role-model who has mentored me throughout my graduate career. Her kindness, knowledge, and dedication to her students and to the field of psychology are inspiring. I would not have survived graduate school without her guidance and she has helped me to grow into the person I am today. I would also like to thank Dr. Lawrie A. Ignacio, my CRP committee member. Her passion for teaching and her knack for writing have enhanced my knowledge and communication skills. Thank you to Dr. Sean Scanlan for always making time to answer my questions and thank you to the faculty at HSPP for their support and belief in me.

I would like to express my sincerest appreciation to Dr. Marvin W. Acklin, my diagnostic and research supervisor, and mentor. He has challenged me to surpass my limits and to grow as a scientist-practitioner. I would also like to thank him for connecting with his colleagues, Dr. Dan Sass, Dr. James N. Butcher, Dr. R Michael Bagby, and Dr. Martin Sellbom, who have been invaluable consultation resources.

I would like to thank Ricky Itagaki, my best friend and significant other, whose support has been instrumental to the completion of my CRP and to the survival of graduate school. Thank you to my research colleagues, Shannon Pickett and Paige Ramos for their support, collaboration, and sympathetic ears. I would also like to thank Reyn Oyadomori for his assistance and extensive mathematical knowledge. Thank you to Catherine Gallahue, Mei-Lin Lawson, and Lyndsey Tom whose love and friendship have been unwavering and will always be appreciated. Last but not least, thank you to my mom, dad, Aunty Julian, and the Itagaki family for the unconditional love and support they have provided me to pursue my dreams. I am forever grateful for these individuals.

## Table of Contents

	Page
Acknowledgments.....	I
Table of Contents.....	II
List of Tables .....	IV
List of Figures.....	V
I. Introduction .....	1
Rationale for Study .....	2
Review of the Literature .....	6
Clinical Assessment and Validity .....	6
Positive Response Distortion and the Validity Debate .....	8
Developing Validity Scales for the NEO-PI-R.....	20
NEO-PI-R Research Validity Scale Characteristics and the Effects of Positive Response Distortion .....	25
Non-Clinical Samples .....	25
Clinical Samples .....	27
Personnel Samples .....	36
Purpose of the Study .....	41
Research Questions and Hypotheses .....	42
Significance of the Study .....	44
II. Methods.....	46
Participants.....	46
Measures .....	46
Minnesota Multiphasic Personality Inventory-2 (MMPI-2).....	46
NEO Personality Inventory-3 (NEO-PI-3) .....	48

Procedures.....	49
Ethical Considerations .....	50
Statistical Analyses .....	50
Rigor .....	55
III. Results.....	56
Reliability of the PPM Research Validity Scale Applied to the NEO-PI-3.....	56
Assessing the Validity and Utility of the PPM Research Validity Scale, Using the MMPI-2 as a Criterion .....	58
Profile Differences.....	62
IV. Discussion.....	68
Discussion of the Findings.....	68
Clinical Implications.....	78
Limitations .....	81
Recommendations for Future Research.....	83
Conclusions.....	86
References.....	89
Appendices.....	101
A. IRB Letter of Certification.....	102
B. Tables .....	104
C. Figures.....	119

List of Tables

	Page
1. Reliability of the PPM Research Validity Scale (N=303) .....	104
2. Correlations Among PPM Research Validity Scale Items and MMPI-2 Validity Scales .....	105
3. Factor Loadings for Principal Axis Factoring with Varimax Rotation of Two Factor Solution of PPM Research Validity Scale Items .....	106
4. Percentage of Participants Who Obtained Elevated PRD Scores on the L, K, or S, or PPM Scales (N=303) .....	107
5. Means and Standard Deviations for MMPI-2 Validity Scales and the PPM Research Validity Scale .....	108
6. Correlations Between the PPM Research Validity Scale and MMPI-2 Validity Scales .....	109
7. Correlations Among the PPM Research Validity Scale, NEO-PI-3 Factor Scales, and MMPI-2 Clinical Scales (N=303) .....	110
8. Patterns and Structure Matrix for Principal Axis Factoring with Direct Oblimin Rotation of Two Factor Solution of Validity Scales .....	111
9. Agreement Between the PPM Research Validity Scale and the MMPI-2 L, K, and S Validity Scales (N=303) .....	112
10. Classification Estimates for Selected Cutoff Scores Based on ROC Analyses on PPM in Detecting Positive Response Distortion .....	113
11. Factor and Clinical Scores of Participants with Invalid ( $Raw \geq 22$ ) Versus Valid ( $Raw \leq 21$ ) Scores on the PPM Scale .....	114
12. Means and Standard Deviations for MMPI-2 Validity Scales: Published DPG and ARD Studies .....	115
13. Means and Standard Deviations for MMPI-2 Clinical Scales: Published DPG and ARD Studies .....	116
14. Scale Raw Score Means and Standard Deviations for Current Sample, PPM Normative Sample, and the NEO-PI-3 Form S Standardization Sample .....	117
15. Factor and PPM T-Score Means and Standard Deviations for NEO-PI-3 and NEO-PI-R, Current Sample, and Published DPG and ARD Studies .....	118



List of Figures

Page

1. NEO-PI-3 and NEO-PI-R Mean T-Score Profiles by Group.....119

# CHAPTER I

## INTRODUCTION

This study examines positive response distortion on self-report personality inventories in personnel selection. Positive response distortion has been raised as an important factor in personnel selection (McGrath, Mitchell, Kim, & Hough, 2010). Research has indicated that demand characteristics in personnel selection increases the motivation to engage in positive impression management, where applicants highlight the positive attributes and characteristics that match perceived job demands (Detrick, Chibnall, & Call, 2010). Additionally, research has found that response bias reduces the validity of the personality measure (Christiansen, Burns, & Montgomery, 2005; McGrath, Mitchell, Kim, & Hough, 2010; Mueller-Hanson, Heggstad, & Thornton, 2003; Rothstein & Goffin, 2006). With increased use of personality assessments in job applicant settings, the ability to detect positive response distortion and to understand its effects on the validity of self-report inventories is necessary. The issue of positive response distortion and the use of validity scales to address it have been debated for over three decades. Research has found that positive response distortion signifies a response bias and that scoring high on these response bias indicators can be indicative of invalidity (Paulhus, 2002). Opponents argue that the use of validity scales are not substantiated and likely interfere with accurate assessment (Costa & McCrae, 1992). This study examines the use of a positive response distortion scale on a recently revised, commonly utilized self-report personality inventory, the NEO Personality Inventory- 3 (NEO-PI-3), its ability to discriminate under-reporting profiles, and the impact this has on clinical interpretation.

Specifically, the current study examined the applicability of the NEO Personality Inventory-Revised (NEO-PI-R) Positive Presentation Management (PPM) research validity scale

developed by Schinka, Kinder, and Kremer (1997) to the updated NEO-PI-3. Concurrent and criterion-related validity of the NEO-PI-3 were computed and assessed against the established L, K, S, F, F<sub>B</sub>, and F<sub>P</sub> validity scales of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher et al., 2001), along with other MMPI-2 scales of underreporting, including the Edwards Social Desirability Scale (So; Edwards, 1957), the Wiggins Social Desirability Scale (Sd; Wiggins, 1959), the Other Deception Scale (ODecep; Nichols & Greene, 1991). The five-factor model (FFM) of personality assessed by the NEO-PI-3 is a widely accepted theory of personality that has historically helped to stimulate interest in examining the relationship between personality measures and personnel selection (Barrick & Mount, 1995; Rothstein & Goffin, 2006). Currently, the FFM is routinely used in personnel selection contexts (Sellbom & Bagby, 2008). Additionally, the research validity scales developed by Schinka et al. (1997) and their ability to detect response bias in clinical and non-clinical contexts have been empirically supported. However, to the current researcher's knowledge, there have been no studies examining the applicability of Schinka et al.'s PPM research validity scale for the NEO-PI-R to the NEO-PI-3. Given that the NEO-PI-3 is utilized in personnel selection and that research has indicated that response bias significantly affects an individual's profile, decreasing the ability to make accurate interpretations based on the personality measure, it would be important to examine the applicability of the PPM research validity scale of the NEO-PI-R to the NEO-PI-3 and to investigate the effects of positive response distortion on an applicant's profile.

### **Rationale for Study**

One of the primary functions as a psychologist is to use psychological tests to gather information that is pertinent to the provision of clinical services (Ben-Porath & Waller, 1992). Psychologists are called upon to conduct psychological evaluations for various reasons including

helping to determine an individual's diagnoses, cognitive functioning, strengths and weaknesses, and services or accommodations needed. More recently, psychologists have been requested to conduct personality assessments to evaluate the suitability of a job applicant for positions within an organization (Rothstein & Goffin, 2006). Investigators have found that the most prevalent reason for using personality testing for employment purposes is its contribution to reducing turnover rates, estimated to be between 20% (Geller, 2004) and as much as 70% (Wagner, 2000), and to increasing employee fit (Rothstein & Goffin, 2006). After surveying recruiters in 2003, Heller (2005) found that 30% of American companies used personality testing as a part of their hiring process. Hsu (2004) estimated that personality testing is a \$400 million-dollar industry in the United States and grows at an increasing rate of 10% per year.

With the increased popularity of personality assessments in job applicant settings, the quality of the data gathered and its validity in accurately portraying the individual in order to predict personality attitudes, behaviors, and performance are augmented. Research has indicated that in personnel selection, there is an increased motivation to present oneself in a desirable manner, which reduces the validity of personality measures (Christiansen, Burns, & Montgomery, 2005; Mueller-Hanson, Heggstad, & Thornton, 2003; Reid-Seiser & Fritzsche, 2001; Rothstein & Goffin, 2006). Under high-demand conditions, such as personnel selection, applicants are motivated to make a good impression and are likely to select responses that highlight positive attributes and match job-perceived characteristics (Detrick, Chinball, & Call, 2010). This poses a significant problem in accurate personnel prediction, given that research has found that a significant proportion of applicants admittedly engage in response distortion on self-reports collected in the selection process (Donovan, Dwight, & Hurtz, 2003). Specifically, Donovan, Dwight, and Hurtz (2003) found that approximately one-third of recent job applicants

participating in the study misrepresented themselves. Furthermore, the results indicated that about 50% of the participants exaggerated personal characteristics of themselves such as dependability, reliability, and agreeability, and around 60% de-emphasized their negative attributes. Thus, it appears that response bias is a common occurrence in personnel selection. Though positive response distortion can occur across various portions of the application process, including the interview and the self-report assessment, with increased use of personality assessments in job applicant settings, it is important to assess the applicant's degree of honesty and objectivity when completing the measure, as most personality assessments rely on the use of self-reported questionnaires.

Moreover, personality testing in commercial and general aviation settings dates back to the 20<sup>th</sup> century (North & Griffin, 1977). In their literature review summarizing the research on psychological assessments in aviator selection, North and Griffin (1977) noted that considering the characteristics of aviator applicants, above-average intelligence, at least college educated, and having completed numerous tests throughout their academic careers, it is possible for highly motivated aviator applicants to readily be able to determine appropriate and inappropriate responses on measures used for aviator selection. Despite the threat of positive response distortion however, many organizations and companies continue to utilize personality testing as a part of their application process, including pilot selection.

Regarding personality, though there are varying definitions, most define personality as a series of traits, constructs, or patterns for how a person thinks, feels, and behaves that are enduring and constant across situations (American Psychiatric Association, 2013; Fitzgibbons, Davis, & Schutte, 2004; McCrae & Costa, 1994). One widely accepted theory of personality concludes that personality traits can be summarized in terms of five robust factors of personality

(Chapelle, Novy, Sowin, & Thompson, 2010; Costa & McCrae, 1992). The five factors include neuroticism, extraversion, agreeableness, openness, and conscientiousness. The widespread acceptance of the FFM has resurged interest in the use of personality measures for selection purposes due to a series of meta-analytic studies and research in the 1990's that provided support for the relationship between personality and job performance (Barrick & Mount, 1995; Rothstein & Goffin, 2006). For instance, one study found that the personality trait of conscientiousness, along with general mental ability and job experience are central determining variables in job performance (Schmidt & Hunter, 1998). Additionally, after reviewing and summarizing the research on the FFM as a predictor of work outcomes, Barrick and Mount (2005) found that neuroticism (emotional stability) and conscientiousness can be considered as generalizable predictors of trait-oriented work motivation and overall job performance. On the other hand, extraversion, openness, and agreeableness are valid predictors of job performance for specific "niches" or job occupations, where certain jobs place emphasis on one of the traits (Barrick & Mount, 2005). Thus, research has been quantitatively able to demonstrate the importance of personality traits as it relates to job performance.

The NEO Personality Inventories, including the NEO-PI-3, are the most frequently used instrument in assessing the Five Factor Model of Personality and are popular personality measures utilized in pilot selection, and (Bagby & Marshall, 2003). The NEO Personality Inventory-3 (NEO-PI-3; Costa & McCrae, 1985), began its development in 1978 as the NEO Inventory aimed at measuring traits related to neuroticism, extraversion, and openness (McCrae & Costa, 2010). In 1983, facet scales measuring conscientiousness and agreeableness were established and in 1985, the NEO Personality Inventory (NEO-PI) was published (McCrae & Costa, 2010). After some modifications and completion of the agreeableness and conscientiousness

facet scales, the NEO-PI-R was created. Then in 2005, after revising 37 out of the 240 items on the NEO-PI-R that either had weak psychometric properties or had wording that was unfamiliar to respondents, the NEO-PI-3 was developed (McCrae & Costa, 2010).

The NEO-PI-R has been utilized in numerous studies investigating the relationship between personality and pilot personnel selection (Callister, King, Retzlaff, & Marsh, 1999; Chapelle et al., 2010; Fitzgibbons, Davis, & Shutte, 2004). To date, however, no studies have been found investigating job-applicant response distortion, specifically positive response distortion and its effects on the NEO-PI-3. Though the developers of the NEO-PI-3 did not include validity scales into the measure, research validity scales measuring positive presentation management, along with negative presentation management and inconsistent responding have been developed for use with the NEO-PI-R. However, the application of the research validity scales to the NEO-PI-3 has not been investigated. Thus, given that the FFM as measured by the NEO Personality Inventories is utilized in pilot selection, that positive response distortion is prevalent in the application process, and that pilot applicants have the intelligence and education to determine which responses are appropriate and inappropriate to pilot selection affecting the validity of the personality measure, a means of measuring positive response distortion is needed.

## **Review of the Literature**

### **Clinical Assessment and Validity**

When conducting any clinical assessment, the first essential step in interpretation is to determine the quality of the data that serves as the foundation for the evaluation (Ben-Porath & Waller, 1992). Ben-Porath and Waller list several pertinent questions useful to evaluating the quality of data gathered, including: “[1] Did the individual cooperate with the evaluation? [2]

Was he or she capable of reading and understanding the test items? [3] Did he or she attempt to portray himself or herself in an overly positive or overly negative manner?" (p.15).

One way test developers and users of clinical measures assess the question of validity outlined by Ben-Porath and Waller (1992) is through measures of response bias. Response bias has generally been defined as a consistent tendency to distort one's responses on a questionnaire in such a way that it interferes with the accuracy of the self-report (McGrath, Mitchell, Kim, & Hough, 2010; Paulhus, 2002). There are a variety of ways in which response bias occurs on a self-report measure. Response bias can occur without consideration of item content. *Inconsistent responding*, also known as *random responding* or *content-nonresponsiveness*, occurs when an individual responds in an unsystematic, non-purposeful manner that is not related to the content of the items (Graham, 2012; Jackson, 1970; McGrath et al., 2010). *Acquiescence* is a tendency to endorse the most positive alternative regardless of item content (McGrath et al., 2010). In contrast, *negativism* refers to the opposite, selecting the most negative response option. When scales have more than two selections, *extreme* or *neutral response bias* can occur. *Extreme bias* occurs when responders endorse the endpoints (lowest or highest) of the item scale; *neutral responders* choose the middle option (McGrath et al., 2010).

When considering item content (i.e., content responsiveness), other forms of response bias may occur that affects the way in which the respondent is portrayed. These types of response biases are negative presentation management and positive presentation management. *Negative response distortion* also referred to as *faking bad*, *negative impression management*, or *overreporting*, occurs when responders endorse an excessive number of negative attributes and/or deny virtues with the intention of portraying themselves as having more problems or symptoms than they really are (Detrick et al., 2010; Graham, 2012; McGrath et al., 2010).



*Positive response distortion* also known *faking good*, *positive impression management*, *socially desirable responding*, *impression management*, or *underreporting* is either the failure to report negative tendencies or the over-endorsement of positive self-descriptions (Detrick et al., 2010; McGrath et al., 2010; Paulhus, 2002). It should be noted that though the concepts listed above tend to be used interchangeably, there are subtle differences in the literature regarding their definition and use. Nevertheless, the focus of this literature review and the current study will be on investigating the effects of positive response distortion, that is, presenting “oneself in a more favorable light than is actually the case, including over reporting of basic virtues and under reporting of faults/dysfunction...in the context of attempting to achieve some advantage by appearing...ideal” on a self-report profile (Detrick et al., 2010, p. 410).

### **Positive Response Distortion and the Validity Debate**

Loevinger (1959) wrote, “proliferation of tests of high sounding psychological constructs in disregard of response bias is conspicuous waste of research” (p.306). Nevertheless, the issue of positive response distortion and the use of validity scales to detect and address response bias has been debated for over three decades. Opponents against the use of validity scales, such as McCrae and Costa (1983), contend that scales used to detect positive impression management have one common defect, which is the inability to differentiate between individuals who are falsely presenting themselves as having desirable characteristics and those who accurately report socially desirable characteristics. McCrae and Costa state, “An individual who is in fact highly conscientious, well-adjusted, and cooperative would appear to be high in social desirability. Paradoxically, it is the most honest and upstanding citizen that these scales would lead us to accuse of lying!” (p. 883). Costa and McCrae (1997) further note that scales used to detect socially desirable responding consist of items that have a desirable response. Thus, creators of

socially desirable responding scales attempt to infer that respondents who endorse such items are attempting to falsely present themselves in a positive manner. However, Costa and McCrae (1997) note that there are other reasons for endorsing such items, including that the items are accurate self-descriptions.

McCrae and Costa (1983) also conclude that within the domain of self-reports it is not possible to disentangle personality substance from response style. Swanson and Ones (2002) echo the remarks of McCrae and Costa, suggesting that measures of positive response distortion are more strongly related to substance and stable personality traits rather than situational factors associated with response distortion (as cited in Goffin & Christiansen, 2003).

With respect to the NEO-PI-3, Costa and McCrae (1992) and McCrae and Costa (2010) note purposefully omitting scales measuring social desirable responding. Though the developers acknowledge that the NEO-PI-3 is not immune to positive response distortion, they chose not to include measures of positive response distortion for two general reasons. First, there is good reason to believe that most respondents in volunteer or clinical contexts tend not to distort their responses. Thus, they believe that the patient's self-reports are generally trustworthy (Costa & McCrae, 1992). Second, scales intended to assess or correct for positive response distortion tend not to work or possibly interfere with accurate assessment (Costa & McCrae, 1997; McCrae & Costa, 1983; McCrae et al., 1989).

McGrath et al. (2010) find support for such assertions; however, they are more conservative in concluding that response bias indicators should be omitted from self-report assessments. McGrath et al. reviewed 41 studies investigating the use of bias indicators as either suppressors or moderators of substantive indicators in applied assessment in various settings. When examining 22 studies that investigated the use of response bias indicators in the general

population, McGrath et al. (2010) found little empirical support for their use as either suppressors or moderators. This suggests that in the general population, when there is no motivation to distort, members are not likely to do so. However, the researchers caution that this finding does not mean that individuals only provide distorted responses in the context of an external incentive. Rather, consistent with Paulhus (1984), response bias may occur because respondents may be “characteristically incapable of perceiving themselves accurately (McGrath et al., 2010, p. 455).

Further, when analyzing 11 studies examining response bias in personnel settings, McGrath et al. found that use of response bias indicators did not substantially enhance the effectiveness of substantive predictors (though use of inconsistent responding scales may be an exception). McGrath et al. identified three possible explanations to account for the limited support, which included: (1) popular response bias indicators such as the MMPI K correction and the Obvious Subtle Index may not be effective indicators of response bias or may reflect substantive aspects of personality; (2) the base rate of response bias may be misestimated, or; (3) the inclusion criteria utilized in the studies examined may have been too unreliable to sufficiently detect bias. In their discussion, the researchers discussed the effects of utilizing response indicators in applied assessment settings, including the effects of false negative (i.e., an airline pilot is successfully able to distort his extremely suicidal or aggressive proclivities) and false positive (financial and effort costs associated) results. Finally, McGrath et al. questioned the justification for using response bias indicators in applied settings due to the lack of empirical support. Rather, the researchers emphasized that if identifying response bias is essential, then “perhaps the best strategy would be to require convergence across multiple methods of assessment before it is appropriate to conclude that faking is occurring” (p.465).

On the other hand, Lowmaster and Morey (2012) found that positive response distortion moderated the effectiveness of predictor variables when examining the predictive validity of the full-scale and subscale scores of the Personality Assessment Inventory (PAI; Morey, 1991) on supervisor ratings in the areas of job performance, integrity problems, and abuse of disability status in a sample of 85 law enforcement officer candidates. Specifically, Lowmaster and Morey found that the relationship between the PAI's full-scale and subscales and job performance outcomes were significantly stronger when participants engage in low positive response distortion in comparison to participants who engage in high positive response distortion. Further, in response to McGrath et al. (2010), Morey (2012), a proponent for the use of response bias indicators, wrote an article critically evaluating McGrath et al. in order to provide a rationale supporting his recommendations to use response bias indicators. Morey critiques the stringent inclusion criteria utilized by McGrath et al. where the researchers had excluded approximately 99% of the scientific literature directly examining whether indicators can detect bias. Morey states that the literature reviewed by McGrath et al. was too restricted in order to truly address the question of validity of response bias indicators. Furthermore, Morey critiqued McGrath et al.'s use of examining only the indirect effect using multiple regression analysis as representative of sufficiently evaluating evidence for validity. After evaluating the regression model put forth by McGrath et al., Morey argues that the formula includes the response bias indicator as a main effect rather than as a moderator. Thus, Morey contends that response bias indicators have utility in predicting a criterion as a main effect. Overall, Morey asserts that there is much evidence as it pertains to the PAI, demonstrating that the PAI validity indicators have utility in detecting the presence of response bias, as both a main effect and as a moderator. Additionally, Morey concludes that "based upon the very narrow focus of the McGrath et al. review and also based

upon the very limited statistical power of all of the studies that they did review, the probability that their conclusions represent a type II error in failing to reject the null hypothesis is appreciable” (p. 160).

However, in another study conducted by McCrae and Costa (1983) investigating the utility of correcting for response bias, the researchers compared 215 men’s and women’s scores on the NEO Inventory domains to the external criterion of spousal ratings of the domains. In an attempt to improve correspondence between the self-report and spousal ratings, the self-report scores were adjusted using the Marlow-Crowne SD scale and the Lie scale of the Eysenck Personality Inventory (EPI). However, after correcting for SD, McCrae and Costa found a decreased in correspondence between self-report and spousal ratings. McCrae and Costa conclude that attempts to correct for SD do not enhance validity.

Nevertheless, reiterating the limitations acknowledged by McCrae and Costa (1983), the researchers examined SD responding using only 2 measures of positive presentation management. Results may have been different if the researchers had examined more established scales of positive presentation management, including the K scale of the MMPI-2. Furthermore, the sample used in McCrae and Costa’s study consisted of a group of volunteers who generally do not have a motivation to distort their responses on the questionnaire. Thus, as emphasized by McCrae and Costa, their findings that “conscious falsification may not be a problem in such contexts as personnel assessments and psychiatric interview” is not substantiated (p. 886).

Proponents for the use of validity scales to detect positive response distortion, such as Ben-Porath and Waller (1992) in their article examining the application of the NEO-PI as a clinical measure, critically evaluated Costa and McCrae’s (1992) claim that any attempts to correct protocol invalidity will not succeed. Ben-Porath and Waller state that Costa and

McCrae's perception of validity tends to be viewed dichotomously. However, Ben-Porath and Waller state that validity should be viewed as a spectrum, where there are different meanings applied to different levels of validity elevations. Thus, rather than viewing response distortions as a means of determining categorically, validity and invalidity of a profile, validity scales should be used to assist in interpretation, where moderate elevations indicate that profile interpretation may proceed but accounting for the likelihood of exaggeration.

Further, rather than viewing positive response distortion as an indicator of dichotomous classification, understanding the conceptualized constructs that underlie positive response distortion is also important. Paulhus (1984) proposed that response bias stems from two different types of motivation. One type is self-deception, which occurs when the respondent is unaware of the truth; meaning that the respondent believes his or her self-reports as it is a dispositional tendency to view oneself favorably. The other type is impression management, where the respondent purposefully deceives on the self-report measure in order to mislead the test administrator. Bagby and Marshall (2004) conducted a study investigating positive response distortion using the MMPI-2 standard validity scales: Lie (L), Correction (K), and Superlative (S), and other developed MMPI-2 validity scales Edwards Social Desirability Scale (Esd; Edwards, 1957), the Wiggins Social Desirability Scale (Wsd; Wiggins, 1959), the Other Deception Scale (Od; Nichols & Greene, 1991), and the Positive Mental Health Scale (PMH4; Nichols, 1991) in an archival sample of university students, a sample of reality television show applicants and parents involved in a family custody evaluation, and a sample of participants who were given either standard instructions or instructed to fake good. The purpose of the study was to investigate the underlying structure of the MMPI-2 validity scales as measures of self-deception and impression-management across various samples.

In their study, Bagby and Marshall (2004) found that the applicants instructed to fake good, reality television show applicant, and parents involved in a family custody evaluation generally scored lower on the MMPI-2 clinical scales in comparison to the sample responding under standard instructions. This finding indicates that positive response distortion is associated with minimization of psychopathological symptoms. Further, the researchers conducted a factor analysis to assess the underlying components of the MMPI-2 validity scales. The results of the principal components analysis on the MMPI-2 validity scales suggested a two-factor solution, where K, S, Eds, and PMH4 had factor loadings  $\geq .87$  on the first factor and  $\leq .25$  on the second factor, and where L, Wsd, and Od had factor loadings  $\geq .75$  on the second factor and  $\leq .28$  on the first factor. The findings obtained by Bagby and Marshall are similar to that of Paulhus (1984), who reported that Wsd and L loaded primarily on an impression management factor and Esd loaded primarily on a self-deceptive factor. Nichols (2001) also reported that the S scale is more related to self-deception than with impression management (as cited in Bagby & Marshall, 2004). Moreover, Bagby and Marshall (2004) found that impression management and self-deception are related to specific assessment contexts, where impression management was more likely to be elevated in analog designs under deliberate instructions to fake good. Thus, it is important to view positive response distortion not only on a spectrum but also to understand the components that make-up positive response distortion.

Additionally, positive response distortion has also been investigated to determine whether it signifies a response bias or a personality dimension. For example, Morey et al. (2002) investigated whether the NEO-PI-R research validity scales were “measuring something substantive (such as psychopathology or its absence) or something stylistic (either effortful distortion or something less conscious such as exaggeration or lack of insight)” in a clinical

sample (p.587). Using a multimethod-multitrait approach, the researchers explored the construct validity in a sample consisting of 668 participants diagnosed with personality disorders (schizotypal, borderline, obsessive-compulsive, or avoidant) or major depressive disorder with no personality disorder. After participants completed the Structured Clinical Interview for DSM-IV (SCID-IV), the Diagnostic Interview for Personality Disorders-IV, the Global Assessment of Functioning (GAF) Scale, the NEO-PI-R, and the Schedule for Nonadaptive and Adaptive Personality (SNAP), researchers used descriptive and correlational statistics to examine the characteristics of the NPM and PPM scale. Additionally, Morey et al. conducted four confirmatory factor analyses to determine whether the research validity scales indicate substantive or stylistic features of responding. The results of the analyses revealed that the NPM and PPM scales were significantly correlated with measures of both global functioning and response validity. These associations held true for both self-report and interviewed-based assessments. However, the CFA revealed that the best fitting model relative to the data was one in which the NEO-PI-R validity scales were considered as part of the stylistic variables but associated with substantive trait variables. Thus, the results of Morey et al. support the NEO-PI-R validity scales as measures of presentation style in clinical populations. Nevertheless, due to the significant correlations among the various substantive and stylistic variables in the study, Morey et al. cautioned against interpreting the NEO-PI-R validity scales solely as measures of response distortion. On the other hand, due to the naturalistic nature of the research, there appeared to be very little incentive for any form of distortion. Thus, the relationship between stylistic and substantive factors may have been more closely intertwined and it would be a mistake to evaluate positive response distortion as simply indicative of a dichotomous outcome that should or should not be included on a personality measure. Further, Morey et al. suggested



that it would be important to examine the relationship between global functioning and response validity in samples where there is more motivation to distort.

All in all, the research suggests that positive response distortion should not be viewed as a dichotomous, all or nothing, variable, but rather it should be viewed along a spectrum, and there are at least two constructs that underlie positive response distortion. Additionally, though positive response distortion is related to substantive characteristics, there is research to suggest that it represents more of a stylistic variable.

Research investigating positive response distortion in high demand conditions, where there is proposed motivation to distort suggests that respondents indeed engage in positive response distortion (Detrick et al., 2010; McGrath et al., 2010). This negates Costa and McCrae's (1992) claim that most respondents in volunteer or clinical contexts tend not to distort their response assertions. Moreover, research indicates that positive response distortion occurs in the context of both self-deception and impression management (Detrick et al., 2010; McGrath et al., 2010). For example, Lautenschlager and Flaherty (1990) found that in high-demand conditions, impression management increased; whereas under anonymous conditions, impression management decreased (as cited in Detrick et al., 2010).

Specifically, using the NEO-PI-R, Detrick et al. (2010) investigated the nature of positive response distortion by police officer applicants in a high demand condition, completing the NEO-PI-R during their pre-employment evaluation, and in a low demand condition, completing the NEO-PI-R after being hired and completing police training academy. The results of the study found that police officers do engage in positive response distortion in the high-demand context of personnel selection. In particular, the researchers found that positive response distortion was primarily manifested on the NEO-PI-R through significantly lower scores on Neuroticism and

high scores on Agreeableness and Conscientiousness when comparing the individual's profiles when completed in a high demand versus a low-demand condition. Thus, though it is aspirational to think that respondents will not engage in positive response bias, research indicates that positive response distortion indeed occurs, especially in high demand conditions where the increased motivation to present oneself positively.

Some research however, has suggested that engagement in positive response distortion does not affect the validity of the personality scales in personnel settings. For instance, Marshall, De Fruyt, Rolland, and Bagby (2005) investigated the factor structure, a form of construct validity, of the NEO-PI-R in job applicant, career counseling, and normative samples in two populations, a French sample and a Belgium sample. In the French sample, the applicants were divided into five subgroups based on their engagement in positive response distortion as measured by Schinka et al.'s (1997) Positive Presentation Management (PPM) research validity scale. The degree of socially desirable responding in the Belgium sample was assumed to vary across the job applicant, career counseling, and normative samples. After conducting an Orthogonal Procrustes rotation, the results of the study found no evidence that the factor structure of the NEO-PI-R significantly differed across groups engaging in socially desirable responding or groups assumed to elicit such responding based on context. This suggests that the construct validity of the NEO-PI-R is sustained at both the domain and facet scales across a varying degree of positive response distortion. However, Marshall, De Fruyt, Rolland, and Bagby found significant differences in applicant profiles depending on their level of positive response distortion. Elaborating further, participants who were categorized in the very high PPM subgroup obtained the significantly highest scores on the Extraversion and Conscientiousness scales and the significantly lowest scores on the Neuroticism scale. These results indicate that

though the factorial validity remains stable, positive response distortion significantly affects scale elevations.

Other research has supported this finding that assessing for response bias on the NEO is important because response distortion affects the validity and utility of a respondent's profile (Sellbom & Bagby, 2008). Holden and Jackson (1981) also found that profiles of individuals with elevated validity scales tend to differ significantly from those individuals less motivated to distort responses. As pertaining to the NEO inventories, Furnham (1997) conducted a study investigating the effects of response distortion on the NEO-FFI in a sample of 70 participants. The participants were divided into three groups, one given instructions to "fake good," one to "fake bad," and one to respond honestly (control group). The results of the study indicated that not only is the NEO-FFI susceptible to faking, but that the profiles of the response groups significantly differ. Specifically, when comparing fake good to control profiles, the control group had significantly high scores on Neuroticism and lower scores on Agreeableness and Conscientiousness. The profile differences found in Furnham (1997) are similar to that of Detrick et al. (2010). Several other studies investigating the NEO-PI-R, which is comparable to the NEO-PI-3, given only modest differences found between the two (McCrae & Costa, 2010), have found similar results, which will be discussed further. Overall, the results of the studies suggest that the NEO-PI-R is not only susceptible to faking, but also that positive presentation management significantly distorts NEO-PI-R profiles (Ballenger, Caldwell-Andrews, & Baer, 2001; Caldwell-Andrews, Baer, & Berry, 2000; Morasco, Gfeller, & Elder, 2007). Specifically, as mentioned previously, research has found that individuals who engage in positive presentation management tend to score significantly high on Extraversion, Agreeableness, and Conscientiousness and lower on Neuroticism and those who engage in negative presentation

management demonstrate a reverse pattern (Detrick et al., 2010; Furnham, 1997; Sellbom & Bagby, 2008). Moreover, distortion in profiles makes it difficult for the clinician to make accurate interpretations. Thus, studies finding profile distortions associated with positive response distortion warrant the continued use of positive presentation management scales and investigating the effects of response distortion on profile interpretation.

Furthermore, Jackson (1970) reports that though at times responding in a desirable manner may contain valid variance, if socially desirable responding is saturated in every scale, it becomes difficult to successfully measure certain traits. High scores on validity scales can affect many statistical analyses, skewing not only the results but also the interpretations made. For instance, high scores on response bias indicators can account for a major portion of the variance. As such content scales may be more highly correlated with one another than should be, interfering with an examiner or investigator's ability to reliably distinguish an individual on distinct dimensions (Jackson, 1970).

Paulhus, Bruce, and Trapnell (1995) conducted a study specifically examining the effects of self-presentation on the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1984, 1989) and the NEO Five Factor Inventory (FFI; Costa & McCrae, 1989) profiles. 370 undergraduate students were asked to respond to the questionnaires as if they were applying for an unspecified job. The participants were randomly assigned to one of seven groups asking them to *fake best, fake good, play up, fake modest, fake bad, fake worst*, or to respond honestly. Not only did the profiles of the six groups differ on all big five personality measures and impression management in comparison to the honest group, but also under all self-presentation conditions, there were substantial intercorrelations between the Big Five dimensions. Furthermore, the correlations between socially desirable responding scales were also inflated for the positive

faking groups. Thus, the researchers demonstrated how response distortion can induce inflated correlations. However, follow-up analyses revealed that convergence of dimensions was not the source of inflation, rather it is the variability in response strategy within a group that leads to inflated intercorrelations. Nevertheless, the results of Paulhus et al.(1995) warn, “a distorted correlational structure may be indicative of self-presentation effects” (p. 107).

Therefore, in accordance with proponents for the use of validity scales, the current study argues that positive response distortion is a response bias that affects the psychologist’s ability to accurately interpret the results of an individual’s self-report measure. Thus, this current study will investigate the underlying construct and effects of positive response distortion in a sample of pilot applicants who were administered the NEO-PI-3 as a part of their application process. Since the measure was completed to assist in determination of applicant selection, given the demands of the situation, it is assumed that pilots are motivated to distort. Research indicates that due to the transparency of item content on the NEO-PI-R, which is highly similar to the NEO-PI-3, it is possible for applicants to engage in positive response distortion if they choose, which compounds the problem (Barrick & Mount, 1996; Marshall, De Fruyt, Rolland, Bagby, 2005). Finally, it is important to examine the effects of positive response distortion on an applicant’s profile as prior research has demonstrated that engaging in positive response distortion is related to significantly different profiles in comparison to individuals who do not engage in response distortion which impacts the predictive validity as it relates to job performance.

### **Developing Validity Scales for the NEO-PI-R**

Response bias in assessment has been addressed in multiple ways. For some, stand-alone assessments measuring response bias have been included as part of the evaluation battery. According to McGrath et al. (2010), the most popular of the free-standing response bias

instruments are the Balanced Inventory of Desirable Responding (BIDR; Paulhus, 1998) and the Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlow, 1960). Others, when developing psychological measures, have eliminated items that correlate too highly with desirable and undesirable responding (Jackson, 1970). The most popular option for managing response bias however, has been to incorporate validity scales within an assessment measure, aiding in the detection and minimization of its effects (McGrath et al., 2010). Examples of such measures include the Minnesota Multiphasic Personality Inventory -2 (MMPI-2; Butcher, Graham, Ben-Porath, Tellegen, Dahlstrom, & Kaemmer, 2001) and the Personality Assessment Inventory (PAI; Morey, 1991). Though McCrae and Costa (2010) omitted scales assessing social desirability from the NEO Inventories, including the NEO-PI-3, NEO-PI-R, and NEO-FFI-3, Schinka, et al. (1997) created validity scales that clinicians could use to assess positive response distortion on the NEO-PI-R.

In their effort to construct self-report validity scales for the NEO, Schinka et al. (1997) conducted a series of studies with the purpose of developing a set of validity scales in order to detect response distortion on the NEO-PI-R. The first study was aimed at using the items on the NEO-PI-R to develop three validity scales: positive presentation management (PPM), negative presentation management (NPM), and inconsistency (INC). The goal for the PPM was to identify respondents who, “in a statistically uncharacteristic manner, claimed uncommon virtues and/or denied common faults” (Schinka et al., 1997, p. 129). The purpose of the NPM scale was to identify respondents who reported uncommon faults and/or denied common virtues. The aim of the INC scale was to identify inconsistent or random responding.

For the first study, 140 men and women were randomly selected from a larger pool of working adults who were participating in a national study of job performance. The participants

completed a biodata form and an array of psychological inventories, including the NEO-PI-R. The 240 items on the NEO-PI-R were utilized to develop PPM, NPM, and INC validity scales. The items chosen for the PPM and NPM scores were derived using empirical and content analysis methods. Empirically, Schinka et al. (1997) selected items with extreme scores (.5 standard deviations above or below the item score midrange) for analysis. Based on the item means, corrected-item total scale correlations, item  $R^2$  values, and domain scale membership, 10 items were each selected for the NPM and PPM scales. Schinka et al. developed the INC scale using strictly empirical methods, using a cutoff score of  $r > .40$ . 10 pairs of items were selected for the INC scale, where 5 were scored in the “normal direction” (Schinka et al., 1997, p. 130), and 5 were scored in the reverse direction. Thus, in the first study, Schinka et al. developed the PPM, NPM, and INC validity scales.

In the second study, Schinka et al. (1997) normed the validity scales on a separate group of 400 men and women. Data from the normative sample were used to examine the viability of the validity scales by analyzing scale scores and interrelationships, specifically the internal consistency of the validity scales and their relationships to the NEO-PI-R domain scales. The participants were proportionately selected from the same pool of working adults in the national study of job performance to match the U.S. 1995 census projections. The researchers found that the PPM scale had a mean of 20.25 and a SD of 4.66; the NPM scale had a mean of 8.78 and a SD of 3.48; the INC scale had a mean of 6.54 and a SD of 2.69. The alpha coefficients for the PPM ( $\alpha=.56$ ) and NPM ( $\alpha=.67$ ) scales, and the correlations between the validity scales and domain scales were comparable to those reported by Morey (1991).

For the third study, Schinka et al. (1997) focused on assessing the effects of response manipulation on the PPM, NPM, and INC scales. The participants were categorized into 5

different groups. The first group consisted of a separate group of 200 participants selected from the same working pool of adults in the previous 2 studies. The next three groups of participants consisted of undergraduate students who were randomly assigned to three different instructional conditions. One set involved standard instructions, where participants were asked to respond honestly. The second set of students were asked to respond to the NEO-PI-R by portraying themselves in a favorable manner. The third set of instructions asked the students to answer the NEO-PI-R by feigning to be an individual with a mental disorder. The fifth group of data was gathered using a computer algorithm to generate 100 NEO-PI-R protocols with random item responses. The data from the five groups were scored following standard NEO-PI-R scoring procedures. Raw scores for the PPM, NPM, and INC were transformed to T-scores using the means and standard deviations derived from the normative sample in study 2. Analyses of the NEO-PI-R protocols of the five different groups revealed that the validity scales as well as the NEO-PI-R were sensitive to the group differences in the expected direction. Therefore, the validity scales established for the NEO-PI-R were able to distinguish positive impression management responding, negative response bias, and inconsistent responding.

Thus, the three studies conducted by Schinka et al. (1997) developed, normed and assessed the usefulness of the NEO-PI-R validity scales in detecting response distortion, which is beneficial for psychologists as it assists in making accurate interpretations about an individual based on the respondent's profile. Nonetheless, there are limitations to the NEO-PI-R validity scales research. As mentioned by Schinka et al., the normative sample in comparison to the projected 1995 US Census, overrepresents young adults ages 21-29, and underrepresents White individuals. Thus, though the researchers attempted to match the normative sample to the US population, the sample was not able to fully achieve a census-matched sample. Additionally,



because the sample was normed on the US population, it would be difficult to generalize the characteristics of the PPM, NPM, and INC scales to populations outside of the general US population. Furthermore, although the researchers outlined the analyses utilized in order to create the PPM and NPM scales, because the researchers use both empirical and subjective methods based on item content analysis, it is difficult to evaluate the methodological rigor used to generate the scales. Additionally, if new validity scales needed to be created based on NEO-PI assessment revisions, it would be difficult to replicate the analyses performed by Schinka et al. (1997) without contacting them directly. Along those limitations, as applied to this current research, there have been no validity scales established for the NEO-PI-3. As mentioned previously, the NEO-PI-3 is likely susceptible to positive response distortion and such distortion has an effect on an individual's profile, as demonstrated in studies assessing faking on the NEO-PI-R, making accurate interpretation about an individual based on their self-report profile difficult. Thus, it would be important to determine whether or not the validity scales developed for the NEO-PI-R are applicable to the NEO-PI-3 (i.e., does the scale demonstrate appropriate psychometric properties, concurrent and discriminant validity?) despite the changes made to the items on the NEO-PI-3, or if new validity scales need to be created for the NEO-PI-3.

Following, the focus of the literature review and the current research will be on the NEO-PI-R research validity scales developed by Schinka et al. (1997). Though other methods using the NEO-PI-R have been developed to detect presentation style (for example, Ross, Bailey, and Millis (1997) developed four facet scales from the NEO-PI-R to detect response distortion), much of the research examining validity using the NEO-PI-R has focused on the NEO-PI-R research validity scales developed by Schinka et al. (Ballenger, Caldwell-Andrews, & Baer, 2001; Morasco, Gfeller, & Elder, 2007; Morey et al., 2002; Sellbom & Bagby, 2008; Young &

Schinka, 2001). Further, while Scandell (2000) has developed validity scales for the NEO-Five Factor Inventory (NEO-FFI, Costa & McCrae, 1992), a brief 60-item version of the NEO-PI-R, this CRP will focus on the research conducted on the NEO-PI-R due to the purpose of the current study, which is to examine the applicability of the NEO-PI-R research validity scales to the NEO-PI-3.

### **NEO-PI-R Research Validity Scale Characteristics and the Effects of Response Distortion**

**Non-Clinical Samples.** Following the development of the Schinka et al. (1997) NEO-PI-R research validity scales, many studies have been conducted investigating the validity and utility of the research validity scales against an external criterion in varying populations, including non-clinical, clinical, and personnel samples. For instance, Caldwell-Andrews, Baer, and Berry (2000) investigated the effects of response distortion on the criterion-related validity of the NEO-PI-R in a sample of 135 undergraduate students. Participants were asked to complete the NEO-PI-R twice, once under standard conditions (Time 1) and once after being randomly assigned to one of three groups (Time 2). One group was given standard instructions, another group was given instructions to create a positive impression, and the last group was given instructions to create a negative impression. Participants also completed the Interpersonal Adjective Scale-Revised-B5 (IASRB5; Wiggins & Trapnell, 1997) under standard instructions, and parents of participants also completed the NEO-PI-R Form R using standard instructions for observer ratings. Using a multivariate analysis of variance (MANOVA) to analyze group differences in NEO-PI-R factor scores during the second administration, results of the analysis revealed that the fake-good group scored significantly higher than the standard group on Agreeableness and Conscientiousness, and significantly lower on Neuroticism. Whereas the fake-bad group scored significantly high on Neuroticism and significantly lower on all other

factors in comparison to the standard group. Furthermore, examining correlations between the NEO-PI-Rs at Time 1 and Time 2 and the external criterion measures (IASRB5 and NEO-PI-R Form R) using a nonindependent-samples significance test revealed that correlations between NEO-PI-R completed after given instructions to distort responses and external criterion measures were significantly lower in comparison to correlations between NEO-PI-Rs completed after standard instructions and external criterion measures. Thus, the results of these analyses empirically support the importance of validity scales, as the findings indicate significant differences in profiles among individuals who engage in response distortion and that response distortion affects the correlations amongst domain scores, both of which affect accurate clinical interpretation. Additionally, Caldwell-Andrews et al. found support for the Schinka et al. (1997) research validity scales, where those who were instructed to fake-bad at Time 2 had very high T scores on the NPM scale and those instructed to fake-good at Time 2 had moderately high scores on the PPM scale. Using a cut-off score of 22 on the PPM, revealed a hit rate of 79% in the scales ability to discriminate fake-good participants from the standard participants at Time 2. A cut-off score of 16 on the NPM yielded an overall hit rate of 85%. Finally, the researchers compared the correlations between NEO-PI-R profiles at Time 2 and IASRB5 score before and after eliminating all participants who scored above the cut-off scores on either the PPM or NPM scale. The results indicated high correlations for most domains after removing the participants who were identified as engaging response distortion using the PPM and NPM scales. Therefore, the findings of Caldwell-Andrews et al. indicate that use of the PPM and NPM scales can help to improve the criterion-related validity of the NEO-PI-R.

On the other hand, when investigating the utility of the NEO-PI-R validity scales along with the validity scales on the Multidimensional Personality Questionnaire (MPQ; Tellegen,

1982), Piedmont, McCrae, Riemann and Angleitner (2000) did not find support for the utility of the NEO-PI-R research validity scales in two different volunteer samples. The researchers compared the NEO-PI-R and MPQ protocols of 178 American students and over 1700 individuals from Germany participating in a twin study against an external criterion of observer ratings using the NEO-PI-R, Form R for the student sample and the NEO-FFI peer-report version for the twin sample. Using suppressor and moderated regression analyses, the researchers found that neither the NEO-PI-R or MPQ validity scales functioned as suppressor or moderating variables, meaning that both the NEO-PI-R and MPQ validity scales failed to enhance the validity of personality assessments. The researchers hypothesized that the utility of the NEO-PI-R and MPQ validity scales may have failed for various reasons, including the unanticipated ways in which the validity scales function or a limited amount of biased or careless responding within the sample. However, as mentioned previously, Morey (2012) argues that response bias indicators have utility in predicting a criterion as a main effect rather than as a moderator. Additionally, prior research has noted that response style is not a consistent moderating factor of agreement between self-reports and observer ratings (McCrae, Stone, Fagan, & Costa, 1998). Thus, investigating response distortion as a moderator of self- and other ratings may not have been a suitable analysis. Nonetheless, the findings of Piedmont et al. do suggest that for the future research investigating the utility of response bias indicators in non-clinical samples should be continued.

**Clinical Samples.** Though the NEO-PI-R and its predecessors were not originally designed to assess disordered personality functioning or psychopathology, clinicians and researchers have applied and investigated the use of the NEO-PI-R in such manner (Costa & McCrae, 1992). Though researchers such as Ben-Porath and Waller (1992) have argued against

the use of the NEO-PI and five-factor model as a stand-alone measure in clinical assessment, they do not oppose the use of normal personality measures, such as the NEO-PI to augment clinical measures in a clinical evaluation.

Numerous research studies have been conducted on the NEO-PI, its variants and revisions and its utility in clinical evaluations. Widiger, Costa, Gore, and Crego (2012) in their chapter, "Five Factor Model Personality Disorder Research," attempted to review and summarize the research that had been conducted on the FFM and personality disorder, as they had done in the previous edition of the text by Widiger and Costa (2002). However, due to a drastic increase in research being conducted on the FFM of personality disorders, Widiger et al. (2012) ceased to create a comprehensive list after identifying more than 200 studies. Additionally, clinicians and researchers have also suggested that the NEO-PI and the NEO-PI-R have diagnostic and treatment-related use in clinical settings (Piedmont & Ciarrochhi, 1999; Quirk, Christiansen, Wagner, & McNulty, 2003). Thus, the reliability and the validity of the NEO-PI-R research validity scales have also been investigated in clinical populations.

Ballenger, Caldwell-Andrews, and Baer (2001) investigated the effects of response distortion using the NEO-PI-R PPM and NPM validity scales in a clinical sample of 60 outpatient psychotherapy patients. The 60 outpatient participants were randomly assigned to one of two groups, where one group was given standard instructions for the NEO-PI-R, and the second group was instructed to fake good. Specifically, the second group was given a scenario where the participants were to imagine they were going through a child custody battle and they needed to make the best possible impression. The outpatient participants' NEO-PI-R profiles were compared to external criteria, which included the Interpersonal Adjective Scale-Revised, Big Five (IASRB5, Wiggins & Trapnell, 1997) profile, and their NEO-FFI Form R (ratings by

others) profile completed by their therapists. Finally, the outpatient participants' NEO-PI-R scores were compared to an archival nonclinical sample of 30 undergraduate students' NEO-PI-R profiles that were completed under standard instructions.

After conducting a multivariate analyses of variance (MANOVA) to compare group differences, the researchers found that the three groups, outpatient-standard, outpatient-fake-good, and student-standard, produced significantly different NEO-PI-R profiles. Follow-up univariate ANOVAs revealed that the patient-standard group scored significantly higher on the Neuroticism scale and significantly lower on the Extraversion and Conscientiousness scales than the other two group and the patient-fake-good group scored significantly higher on the Agreeableness scale in comparison to the other two groups. This indicates that the patients engaging in positive presentation management produce significantly different factor scores on the NEO-PI-R.

When comparing the outpatient NEO-PI-R, IASRB5 profiles, the standard group produced significant correlations among all five corresponding scales. The outpatient-fake-good group produced only two out of five significant correlations among the corresponding scales, one of which (Conscientiousness) was negative. When compared to their therapist's ratings on the NEO-FFI, the outpatient-standard group produced significant correlations for Neuroticism, Extraversion, and Openness. For the outpatient-fake-good group, only one factor (Conscientiousness) was correlated between self-report and therapist ratings in the negative direction. These results indicate that criterion related validity is reduced when respondents are faking good.

Finally, analysis revealed that research validity scales were moderately accurate in detecting faking good from standard patients, and marginally accurate in detecting faking good

from the non-clinical student sample. These results indicate that it may be difficult to detect whether or not a person who achieves an average functioning profile is truly functioning in that range. Though the study made strides in examining the validity of the NEO-PI-R PPM and NPM validity scales in a clinical population in comparison to both external criteria and a non-clinical sample, the study was not without its limitations. These include, the homogeneity of the sample, limiting the finding's generalizability, and the fact that the "standard" instructions presented to the outpatient-standard group for the NEO-PI-R were altered. The outpatient-standard group was also given the instructions/incentive that "those who were honest and paying attention while completing the questionnaires would be included in a drawing for three prizes" (Ballenger et al., 2001, p. 256). These added instructions may have influenced the way in which the outpatients responded to the NEO-PI-R test items. Some participants may have changed their responses from what they would have originally chosen to a response that they think is "more honest." Thus, it is difficult to determine the accuracy of differences found amongst the profiles given the additional instructions.

Researchers have also examined the convergent and discriminant validity of the NEO-PI-R research validity scales in clinical samples using only standard instructions. Following the study conducted by Schinka et al. (1997) where the researchers investigated the NEO-PI-R research validity scales characteristics in a sample of working adults and undergraduate students, Young and Schinka (2001) conducted a study examining the characteristics of the PPM and NPM validity scales in a clinical sample. The clinical sample consisted of 118 males diagnosed with alcohol dependence disorder voluntarily seeking treatment in a Veterans Affairs substance abuse treatment program. Further, Young and Schinka examined the relationship between the NEO-PI-R research validity scales and the Personality Assessment Inventory (PAI) validity

scales. The PAI consists of four validity scales, Infrequent Responding, Inconsistent Responding (ICN), Positive Impression Management (PIM), and Negative Impression Management (NIM). Research supports the utility of validity scales in detecting response distortion (Morey, 1991; Morey & Lanier, 1998). Young and Schinka analogized the NEO-PI-R PPM scale to the PAI PIM scale, the NEO-PI-R NPM scale to the PAI NIM scale, and the NEO-PI-R INC scale to the PAI ICN scale. The researchers hypothesized that the NEO-PI-R PPM and NPM scales would show significant concurrent and discriminant correlations consistent to corresponding PAI validity scales. Young and Schinka (2001) also postulated that the NEO-PI-R INC scale would not be highly correlated with the PAI ICN scale because the item content differs between the two inventories.

For their analyses, Young and Schinka (2001) considered NEO-PI-R and PAI profiles with validity scale T scores  $\geq 70$  invalid. For their study, Young and Schinka found satisfactory internal consistency, using alpha coefficients, for the PPM ( $\alpha=.70$ ) and for the NPM ( $\alpha=.75$ ). Correlations between the NEO-PI-R and PAI validity scales supported Young and Schinka's hypotheses and revealed a significant positive correlation between the PPM and PAI PIM and a significant negative correlation between the PPM and PAI NIM. Additionally, the NPM was found to have a significant positive correlation with the PAI NIM and a significant negative correlation with the PAI PIM. The INC scale was not significantly correlated with any of the NEO-PI-R or PAI validity scales consistent with Young and Schinka's hypothesis that the "scale is a measure of content-less response style and not of an underlying construct" (p. 416). When assessing the agreement between the respective PAI and NEO-PI-R validity scales, Young and Schinka needed to restrict their examination to the NPM scale due to the lack of invalid profiles based on the PPM and PAI PIM. Young and Schinka attributed these findings to the voluntary



nature of the participants. Using a T score of  $\geq 70$  as the classification criterion for valid and invalid profiles, the study found 70 percent agreement for the NPM and PAI NIM scales. Finally, after conducting a multivariate analysis of variance (MANOVA), the study found significant differences in PAI profiles depending on the validity of their NPM scales. Follow-up analyses revealed that individuals with invalid NPM scale scores had significantly higher scores on the following PAI validity and clinical scales: Infrequency, NIM, Somatic Complaints, Anxiety, Anxiety-Related Disorders, Depression, Paranoia, Schizophrenia, Borderline, and Drug Problems.

Overall, the research study found that the NPM and PPM scales have internal reliability, consistent to those reported previously in a non-clinical sample (Schinka et al., 1997). Furthermore, the findings support the concurrent and discriminant validity of the NEO-PI-R research validity scales as compared to the PAI validity scales. Finally, analyses revealed that the PAI profiles of invalid NPM scale score significantly differ from profiles of valid NPM scale scores on various clinical and validity indexes. Thus, the research conducted by Young and Schinka (2001) support the concurrent validity of the NEO-PI-R PPM and NPM validity scales. Therefore, these findings support the importance of incorporating validity scales into measures as there is evidence that response distortion affects profile scales which in turn affects the Psychologists' ability to make accurate interpretations and recommendations. In the future, it would be important to select a sample where both NPM and PPM are evident in order to examine the effects of both types of response distortion on a respondent's profile. Additionally, to improve the strength of the findings, it would have also been important to assess the level of agreement between cut-off scores used to determine invalid profiles using the NPM and the NIM.

Morasco, Gfeller, and Elder (2007) examined the usefulness of the NEO-PI-R validity scales in detecting response distortion by comparing the scales with the MMPI-2 validity scales. The researchers specifically chose to compare the NEO-PI-R validity scales with the MMPI-2 validity scales because the MMPI-2 validity scales have been reliably shown to detect random responding, malingering, and positive impression management (Bagby et al., 1997; Gallen & Berry, 1996; Graham, Watts, & Timbrook, 1991; Morasco et al., 2007). The data was drawn from a sample of clients who completed comprehensive psychological evaluations at a university-based psychological services center. The referral questions ranged from identifying psychological factors that may be impacting educational functioning in order to determine whether or not special accommodations were needed, to neuropsychological evaluation and psychiatric differential diagnoses. The researchers compared the NEO-PI-R research validity scales to the MMPI-2 validity scales, specifically, the L, F, K, F—K, F(b), and Variable Response Inconsistency (VRIN) scales. Additionally, the researchers evaluated the diagnostic overlap of invalid responding, where the NEO-PI-R PPM score was compared with T scores > 65 on the MMPI-2 K scale as indicative of attempting to present oneself in an overly positive manner. The NEO-PI-R NPM scale was compared with a T score > 80 on the MMPI-2 F-scale to suggest malingered psychopathology.

The results of the correlations revealed that the NPM scale and the MMPI-2 validity scales assessing exaggerated symptomatology, which included the F, F—K, and F(b) scales were all significantly and positively correlated. Further, the NPM scale was significantly negatively correlated with the MMPI-2 L-scale, but not the K-scale. The PPM scale was significantly and positively correlated with the MMPI-2 L and K-scales, which measure positive impression management. Moreover, the PPM scale was significantly negatively correlated with the MMPI-2

F, F–K, and F(b) validity scales. The NPM–PPM scale was positively and significantly correlated with the MMPI-2 F–K scale, both measures of exaggerated psychopathology. However, the INC scale was not significantly correlated with the MMPI-2 VRIN scale. Univariate analyses also revealed significant differences between participants with valid and invalid PPM and NPM profiles. In comparison to individuals with valid PPM profiles, individuals with invalid PPM profiles scored higher on the NEO-PI-R Conscientiousness scale, and lower on the NEO-PI-R Neuroticism scale and MMPI-2 Scales 2, 7, and 0. Individuals with invalid NPM scale scores achieved significantly lower NEO-PI-R Extraversion scale scores and higher MMPI-2 Scale 0 scores in comparison to valid NPM responders.

Though the researchers attempted to select clients with a range of referral questions, a majority of the referral questions were geared towards assessing for problems that may be impacting an individual's educational functioning. Additionally, because the participants were selected through a university clinic, the sample was moderate in size, and only a few participants (10/74) produced an invalid MMPI-2 profile, the power and generalizability of the findings of the research are limited. Furthermore, the internal consistency scores for the NEO-PI-R research validity scales were rather low (.43 for PPM and .60 for NPM) in comparison to those reported by Young and Schinka (2001), indicating caution when using these scales. Thus, more research should be done investigating the psychometric properties of the NEO-PI-R research validity scales and their use as validity scales in clinical populations.

Sellbom and Bagby (2008) conducted a similar study investigating the validity and utility of the PPM and NPM research validity scales for the NEO-PI-R in a sample of 360 psychiatric patients, using the MMPI-2 as a referent. The researchers divided the patients into three groups, invalid underreporting, invalid overreporting, and valid responding using the L, K, S

(underreporting validity scales),  $F$ ,  $F_B$ , and  $F_P$  (overreporting validity scales) of the MMPI-2. The researchers classified patients in the invalid underreporting group if they scored  $\geq 65T$  on at least two of the three MMPI-2 underreporting validity scales. They identified patients as invalid overreporting if they scored  $\geq 100T$  on  $F$  and  $\geq 110T$  on  $F_B$  or  $\geq 100T$  on  $F$  and  $\geq 100T$  on  $F_P$ .

To examine whether response bias posed any threat to the internal validity of the NEO-PI-R, Sellbom and Bagby (2008) first utilized a covariance matrix. The results of the analyses indicated that the correlations between the personality domains were larger for the underreporting than the valid responding group, but correlations were generally equivalent between the personality domains of the overreporting and valid responding group. To assess if response bias affected the clinical interpretability of the NEO-PI-R, the researchers conducted an ANOVA. The findings indicated that relative to the valid responding group, the underreporting group score significantly lower on Neuroticism and significantly higher on Agreeableness and Conscientiousness, whereas the overreporting group score significantly higher on Neuroticism and significantly lower on Extraversion, Agreeableness, and Conscientiousness. The researchers did not examine the effects of response bias on clinical interpretability of the MMPI-2. Assessing the concurrent validity of the NEO-PI-R PPM and NPM research validity scales, the researchers found: moderate and positive correlations between the PPM and the MMPI-2 underreporting scales, small to moderate negative correlations between the PPM and overreporting scales, moderate and positive correlations between the NPM and MMPI-2 overreporting scales, and small negative correlations between the NPM and MMPI-2 underreporting scales. Therefore, the results of Sellbom and Bagby also find support for the concurrent validity of the NEO-PI-R research validity scales, and results add to the current literature indicating that response bias significantly affects a respondent's profile, likely leading to misinterpretation.

Overall, the current research establishes the convergent and discriminant validity of the NEO-PI-R research validity scales in detecting response bias in clinical samples. The PPM and NPM were significantly correlated with validated measures of underreporting and overreporting in the expected directions. Additionally, the results highlight the importance of assessing for validity as those participants who responded in an invalid manner on the NEO-PI-R research validity scales generated significantly different profiles than those participants with valid protocols. This supports findings that positive response distortion affects criterion related validity and profile interpretability. However, research in clinical samples is inconsistent regarding the psychometric properties of the PPM research validity scale. Thus, more research needs to be conducted investigating the reliability of the PPM research validity scale as well as the effects of positive response distortion on clinical interpretability in other samples.

**Personnel Samples.** As discussed earlier in this literature review, given the high demand conditions of personnel selection, increasing an applicant's motivation to distort, Schinka et al.'s (1997) research validity scales have been investigated in the context of employment settings. For instance, when investigating the use of PPM and NPM research validity scales in employment, normative, and clinical samples, Blanch, Aluja, Gallart, and Dolcet (2009) found a large difference in PPM research validity scores between employment and clinical samples. Specifically, they found that employment samples score higher on PPM than clinical samples. This suggests that individuals in employment contexts tend to engage more in positive response distortion than in clinical samples. However, it is noted that the researchers emphasize caution when interpreting the results of their study due to a number of limitations, including a small number of studies analyzed. Nevertheless, Birkeland, Manson, Kisamore, Brannick, and Smith (2006) conducted a meta-analytic study investigating the extent to which job applicants

demonstrate positive response distortion on personality measures assessing the Big Five personality constructs. The results of their study indicated that job applicants tend to present themselves more positively in comparison to non-job applicants, but to a lesser extent in comparison to individuals who were specifically instructed to fake. Further analyses indicated that job applicants may distort responses to match the perceived desired characteristics relevant to the job. Moreover, individuals who scored high on measures of social desirability also scored high on big five personality dimensions that assess for emotional stability and conscientiousness. Blanch, Aluja, Gallart, and Dolcet (2009) highlight the fact that this is a problem in employment contexts, as conscientiousness and emotional stability are important predictors of job performance.

Juhel, Brunot, and Zapata (2012) investigated the effects of positive response distortion using the PPM research validity scale in 974 candidates taking the NEO-PI-R as part of their entrance examination to the National School of Civil Aviation (ENAC) in France. Analyses indicated that the candidates scored significantly higher on PPM and obtained higher scores on the Conscientiousness, Extraversion, and Agreeableness domains and lower scores on the Neuroticism domain in comparison to a reference sample. Thus, the researchers found support for the utility of Schinka et al.'s (1997) PPM research validity scale in personnel selection, specifically as applied to aviation applicants.

On the other hand, Reid-Seiser and Fritzsche (2001) investigated the PPM research validity scale, first using archival data from 90 participants. The participants were customer service representatives at a national insurance company. When assessing the relationship between personality and performance, the results indicated that PPM did not moderate the relationship between any of the five NEO-PI-R factors and performance ratings. Additionally,

when Reid-Seiser and Fritzsche examined the relationship between personality and performance in 150 psychology majors, using grade point average as the performance measure, they found that PPM again, did not moderate the relationship. Thus, Reid-Seiser and Fritzsche concluded that positive response distortion does not negatively impact the criterion-related validity of personality measures in a personnel setting. On the other hand, due to the significant relationship between PPM and performance ratings and between PPM and four of the FFM dimensions in sample of customer service participants, Reid-Seiser and Fritzsche suggest that PPM may be more aligned with self-deceptive enhancement rather than impression management. Particularly, in their second study involving students, Reid-Seiser and Fritzsche found that PPM was significantly related to self-deceptive enhancement and not to impression management. Therefore, the PPM may not be useful in employment contexts in detecting individuals who are attempting to deliberately distort their responses in order to be viewed positively (impression management) from individuals who have a more stable view of oneself in positive terms (self-deceptive enhancement).

The researchers acknowledged the small sample size, limiting their power to detect a small effect. Additionally, the methodological rigor of the studies is also questionable. Though the aim of the study was to investigate the ability of the PPM scale to detect response distortion in employment contexts, real-world job applicants were not used in the study. Moreover, only five participants from the customer service sample obtained PPM scores that were 2 standard deviations above the mean for the normative sample. Therefore, the number of individuals engaging in positive response distortion on the PPM was low, limiting the studies ability to detect significant differences. Moreover, their second study was a simulation study where university students were provided instructions to engage in positive response distortion in order

to maximize their chances of being hired. Though important to investigate positive response distortion in a simulated sample, it is difficult to generalize the findings to real-world personnel selection samples completing psychological measures as part of their hiring process, who likely experience a great more distress in comparison to simulation samples.

Stress may be an important factor to consider, as high stress levels have found to negatively impact performance in personnel selection (Can, 2011). Specifically, applicants for pilot selection are likely experiencing an immense amount of stress due to the desire to be hired. Traditionally, if an applicant has applied to an airline selection program at a major air carrier, if the applicant is being routed to take the psychological examination, he or she has passed the flight experience, simulator check ride, and aviation knowledge exam and has been extended a provisional offer of employment (Butcher, Gucker, & Hellervik, 2012). Therefore, it is likely the applicant is experiencing an increase in pressure, as he or she has a conditional offer and the psychological examination is an obstacle to overcome. Thus, motivation to pass and to “look good” on the assessment increases. Moreover, Minter (2015) found that pilots usually acquire a \$200,000 debt prior to employment. Thus, there is likely the added stress of obtaining a job in order to pay off one’s loans. Therefore, it is important to investigate the ability of the PPM scale to detect response distortion in real-world personnel selection samples.

Bagby and Marshall (2003), conducted a study comparing the test results from the NEO-PI-R between real-world participants attempting to win a competition, where positive response distortion was likely to occur (differential prevalence group; DPG) and participants who were provided instructions in an experimental context designed to imitate the same test-taking condition (analog research design; ARD) The ARD sample completed the NEO-PI-R twice, once under standard instructions, and a second time after being given instructions to fake-good. The



results of the analyses indicated that the ARD group under fake-good instructions obtained significantly lower scores on the Neuroticism scale and significantly higher on the Extraversion and Conscientiousness scale in comparison to responding under standard instructions. Further, the ARD group under standard instructions when compared to the DPG sample, scored significantly higher on the Neuroticism scale and significantly lower on the Extraversion scale. Additionally, the ARD group under fake good instructions scored significantly higher on the Extraversion scale and significantly lower on the Agreeableness scale than did the DPG sample. Moreover, the ARD group under standard instructions scored significantly higher than the DPG sample, and there was no significant difference on PPM between the ARD group under fake good instructions and the DPG.

The results of the analyses indicated that the findings from ARD studies investigating the effects of positive response distortion on the clinical interpretability of self-report inventories, such as the NEO-PI-R could be generalized to real-world personnel selection samples. However, the findings also indicated significant profile differences between the ARD group under fake-good instructions and the DPG sample. Therefore, as previously discussed, it is important to continue to investigate positive response distortion in real-world personnel selection samples to obtain a better understanding of its effects. Of surprise, the findings of Bagby and Marshall (2003) regarding the PPM research validity scale is inconsistent with previous research that suggest that the PPM research validity scale is able to distinguish between fake-good and valid profiles (Juhel et al., 2010; Schinka et al., 1997; Young and Schinka, 2001). Thus, more research investigating the PPM research validity scale in employment contexts is needed.

Overall, research has demonstrated that there is a significant difference in profiles between individuals who demonstrate and do not demonstrate positive response distortion on

self-report measures. However, the ability to detect such differences using the NEO-PI-R PPM research validity scale has been inconsistent. Moreover, as in other samples investigating the validity and utility of the NEO-PI-R PPM research validity scale, no studies were found investigating the applicability of the NEO-PI-R PPM research validity scale to the NEO-PI-3 in employment contexts. This is troublesome, as the NEO-PI-3 has been applied in various personnel selection settings, including that of pilot selection. The ability to accurately detect positive response distortion in pilot selection is imperative as it affects the clinical interpretability of personality measures, which in turn affects the mental health professional's ability to accurately assess an applicant's personality. An applicant's personality has found to be pertinent to a pilot's successful performance. Foushee and Helmreich (1988) and Sells (1955) have indicated that the effectiveness of the flight crew is rooted in the combination of three components: technical skills, attitudes, and personality characteristics (as cited in Lochner & Nienhaus, 2016). Therefore, investigating the applicability of the NEO-PI-R PPM research validity to the NEO-PI-3 and the effects of positive response distortion on the clinical interpretability of self-report personality measures will assist mental health professionals in making for informed decisions regarding pilot selection.

### **Purpose of the Study**

Given the increased popularity of using personality assessments in personnel selection, assessing the validity of self-report measures is vital in order to accurately interpret an applicant's profile. Thus, the purpose of the study is to examine positive response distortion in pilot applicants who completed the NEO-PI-3 and the MMPI-2 as a part of their application process. Specifically, given that the NEO-PI-3 is used in personnel selection but that no validity scales have been developed to assist in interpretation, the aim of this study is to determine the

applicability of the PPM research validity scale developed for the NEO-PI-R to the NEO-PI-3. In order to do so, the PPM research validity scale for the NEO-PI-R as applied to the NEO-PI-3 will be calculated for the present study's sample and compared to the L, K, S, Sd, So, ODecp, F, FB, and F<sub>p</sub> scales of the MMPI-2. Thus, the concurrent validity, or the extent to which test scores correspond to an established measure of the same construct (Carducci, 2009) of the PPM research validity scale, will be assessed. Additionally, due to prior research indicating that positive response distortion on the NEO-PI-R significantly distorts an individual's profile, the current study will also be examining the effects of positive presentation management on an applicant's NEO-PI-3 and MMPI-2 profiles.

### **Research Questions and Hypotheses**

The current study will examine the applicability of the NEO-PI-R PPM research validity scale developed by Schinka et al. (1997) to the updated NEO-PI-3. The concurrent validity of the NEO-PI-3 PPM scale will be computed and assessed against the established L, K, and S scales of the MMPI-2 and other MMPI-2 validity scales of underreporting, including the Sd, So, and ODecp scales. The following research questions will be addressed in this proposal: (1) *Does the empirically derived PPM research validity scale demonstrate acceptable psychometric properties;* (2) *Is the PPM research validity scale developed for the NEO-PI-R applicable to the NEO-PI-3 in detecting and classifying positive response distortion using the MMPI-2 validity scales as criterion measures,* (3) *Is the PPM research validity scale as applied to the NEO-PI-3 a measure of impression management, and;* (4) *Does positive response distortion have an effect on a respondent's profile?*

Investigating the applicability of the PPM research validity scale of the NEO-PI-R to the NEO-PI-3 by assessing the research validity scale's concurrent validity with the established

validity scales on the MMPI-2 will aid in assessing whether or not the validity scale originally developed for the NEO-PI-R is useful in detecting positive response distortion for NEO-PI-3 users. Additionally, examining the profile differences among participants who demonstrated positive response distortion on the PPM and those participants who did not will not only add to the literature about the use of validity scales but also has the potential of aiding mental health professionals, specifically those who use personality assessments to predict certain outcomes, in understanding how positive response distortion may affect profile interpretation. Based on the literature review, this proposal postulates the following hypotheses:

Hypothesis 1: The PPM research validity scale of the NEO-PI-R as applied to the NEO-PI-3 will demonstrate acceptable psychometric properties.

Hypothesis 2: There will be a positive, significant correlation at the  $p < .05$  level among the calculated PPM research validity scales of the NEO-PI-R as applied to the NEO-PI-3 and the L, K, S, Sd, So and ODecp scales of the MMPI-2

Hypothesis 3: There will be a negative, significant correlation at the  $p < .05$  level among the calculated PPM research validity scale of the NEO-PI-R as applied to the NEO-PI-3 and the F, F<sub>B</sub>, and F<sub>P</sub> scales of the MMPI-2

Hypothesis 4: There will be two factors that underlie positive response distortion and the PPM research validity scale will load strongly on the factor measuring impression management.

Hypothesis 5: The PPM research validity scale of the NEO-PI-R as applied to the NEO-

PI-3 will accurately discriminate between profiles identified as valid and invalid based on the PPM cut-off score.

Hypothesis 6: There will be at least a moderate agreement ( $Kappa \geq .5$ ) between profiles identified as invalid by the PPM research validity scale of the NEO-PI-R as applied to the NEO-PI-3 and profiles identified as invalid by the K scale of the MMPI-2. The agreement will be significant at  $p < .05$ .

Hypothesis 7: There will be a significant difference at the  $p < .05$  level in profiles between participants who demonstrated positive response distortion on the PPM validity scale and participants who did not demonstrate positive response distortion on the PPM validity scale.

### **Significance of the Study**

A current review of the literature found no studies investigating the use of the PPM research validity scale of the NEO-PI-R to the NEO-PI-3 to assist in detecting positive response distortion. By applying the PPM research validity scale to the NEO-PI-3 and examining its concurrent validity against established validity scales measuring underreporting for the MMPI-2, it is hoped that this study will be useful in not only adding to the literature about the importance of assessing for validity in personnel selection and investigating the usefulness of the PPM research validity scale as applied to the NEO-PI-3, but also to assist psychologists in profile interpretation. As stated earlier in the literature review, one of the primary and unique functions of a psychologist is to use psychological evaluations to assist in the provision of clinical services. Psychologists are called upon to conduct assessments for a multitude of reasons and in order to increase the accuracy of interpretations, predictions, or hypotheses made from the evaluation, an

assessment of the validity of data, which serves as the foundation of the evaluation, is vital.

Thus, in utilizing personality measures in personnel selection, it is hoped that we can develop a workforce that not only has essential knowledge, skills, and abilities, but also work-related personality characteristics that will contribute to success (Cascio, 1995; Reid-Seiser & Fritzsche, 2001).

## CHAPTER II

### METHODS

#### Participants

A sample of 303 airline pilot candidates who completed both the MMPI-2 and the NEO-PI-3 as part of their hiring process was extracted from an archival database. The archival database was obtained from a licensed psychologist's independent practice in Hawaii who is a contracted with a commercial airline company to conduct personnel selection psychological evaluations. The sample represents all airline pilot applicants who completed both the MMPI-2 and NEO-PI-3 as part of their hiring process from the years 2014 to 2018. No participant was excluded from the sample as all of them had completed a MMPI-2 and a NEO-PI-3 and all participants' protocols fell within the normal range for number of items missing on the MMPI-2 and NEO-PI-3. The mean age of the sample ( $n=303$ ) was 35.5 years ( $SD = 6.8$ , range = 23-57). A majority of the participants were male ( $n=270$ , 89.1%) and the sample's mean education level ( $n=159$ ) was 16 years of education ( $SD=.3$ , range =14-18).

#### Measures

**Minnesota Multiphasic Personality Inventory-2 (MMPI-2).** The Minnesota Multiphasic Personality Inventory-2 (MMPI-2) is a broad-band instrument designed to assess a number of personality patterns and psychopathology. The inventory is a 567-item true-false questionnaire. The MMPI-2 consist of multiple scales including Validity, Clinical, Restructured-Clinical, Harris-Lingoes, Content, and Supplementary Scales. Butcher, Graham, Ben-Porath, Tellegen, Dahlstrom, and Kaemmer (2001) found good internal consistency ranging from .37 to .90. A median interval of one-week test-retest reliabilities have ranged from .54 to .93 (Butcher et al., 2001). The MMPI-2 validity scales have been extensively researched, and studies have indicated that the validity scales are useful in detecting response distortion (Bagby et al., 1997;

Gallen & Berry, 1996; Graham, Watts, & Timbrook, 1991; Morasco et al., 2007).

The specific validity scales of the MMPI-2 that will be used in this study are the L, K, S, Edwards Social Desirability Scale (So; Edwards, 1957), the Wiggins Social Desirability Scale (Sd; Wiggins, 1959), the Other Deception Scale (ODecp; Nichols & Greene, 1991) as measures of underreporting, and the F, F<sub>B</sub>, and F<sub>P</sub> scales as measures of overreporting. The L scale was developed to identify respondents who may distort their responses by denying various minor faults that most individuals are willing to acknowledge (Butcher et al., 2001). The K scale is a more subtle index that assesses a respondent's level of defensiveness on the MMPI-2 items (Butcher et al., 2001). The S scale measures the tendency to present oneself on the MMPI-2 as highly virtuous, responsible, and free of psychological problems (Graham, 2012). Though not a part of the standard scoring for the MMPI-2, the Sd, So, and ODecp scales were calculated. The Sd scale assesses for socially desirable responding and there is support for including the Sd scale in future research studies (Graham, 2012). The So scale too assesses for socially desirable responding and item content for the scale reflects freedom from psychological problems, comfortability with others, and good attention and concentration skills (Greene, 2000). The ODecp scale was updated from the Positive Malingering (Mp) scale for the MMPI-2. Item content for the ODecp scale reflects self-confidence and denial of psychological problems (Greene, 2000). The F scale was developed to identify individuals who responded to the MMPI-2 in a deviant or atypical manner, whereas the F<sub>B</sub> scale is used to detect test takers who may have responded in an invalid manner in the second half of the test booklet (Butcher et al., 2001). The F<sub>P</sub> scale consists of items that were answered infrequently by both psychiatric inpatients and persons in the MMPI-2 normative sample (Butcher et al., 2001).

In order to measure the ability of the PPM research validity scale to detect positive



response distortion from valid profiles, using the MMPI-2 validity scales as a criterion measure, a cut-off score  $T \geq 65$  on the K, L, and S scales will be indicative of positive response distortion. The cut-off score utilized is similar to the study conducted by Morasco et al. (2007) and Sellbom and Bagby (2008) and in accordance with the cut-off scores reported by Butcher et al. (2001).

**NEO Personality Inventory – 3 (NEO-PI-3).** The NEO Personality Inventory-3 (NEO-PI-3) is a concise measure of the FFM of personality, including neuroticism, extraversion, openness, agreeableness, and conscientiousness, and the major traits that define each factor (McCrae & Costa, 2010). The applicants completed Form S, the self-report item booklet of the NEO-PI-3. The inventory consists of 240-items using a 5-point Likert scale, ranging from 0 to 4: *Strongly Agree, Agree, Neutral, Disagree, or Strongly Disagree* with the item statement. The items comprise five factors represented by six scales that measure facets of the factor (McCrae & Costa, 2010). McCrae and Costa (2010) report good internal consistency for Form S of the NEO-PI-3, with values ranging from .89 to .93 for the five domains and from .54 to .83 for the 30 facets. McCrae & Costa do not report values for test-retest reliability for the NEO-PI-3. However, using a study conducted by Kurtz and Parish (2001), McCrae and Costa estimated test-retest reliability values for the NEO-PI-3 based on Kurtz and Parish's administration of the NEO-PI-R. McCrae and Costa rationalize that this is possible because of the near-equivalence between the NEO-PI-R and the NEO-PI-3. The estimated values range from .91 to .93 for the five domains and from .70 to .91 for the 30 facets.

Standard scoring of the NEO-PI-3 does not produce validity scale scores. McCrae and Costa (2010) purposefully omitted scales measuring social desirable responding based on the notions that a patient's self-reports are generally trustworthy and scales intended to assess or correct for positive response distortion tend not to work. Therefore, the PPM research validity

scale was calculated using the procedures outlined by Schinka et al. (1997). Six out of the 10 items on the PPM research validity scale are negatively keyed. Schinka et al. (1997) reported a coefficient alpha of .60 for the PPM research validity scale. Internal consistency for the sample was calculated for the PPM and is discussed in the results section. In accordance with previously conducted research (Caldwell-Andrews, Baer, & Berry, 2000; Morasco et al., 2007; Sellbom & Bagby, 2008), a cut-off score of  $\geq 22$  on PPM would be indicative of positive response distortion, in order to measure the ability of the PPM scale to detect positive response distortion from valid profiles

### **Procedures**

After receiving approval from the University's Institutional Review Board, archival data was gathered, coded, and analyzed. The archival data was exported from Pearson and PARinc databases into Microsoft Excel. A codebook was created that includes a list of variables and their corresponding coding instructions including demographic variables and variables classifying the sample into dichotomous groups of profiles that demonstrated positive response distortion and normal range profiles based on the discussed MMPI-2 K, L, and S validity scale cut-off scores of  $T \geq 65$  and the NEO-PI-R PPM research validity scale as applied to the NEO-PI-3 cut-off raw score of  $\geq 22$ . Items for the PPM research validity scale for the NEO-PI-R and its corresponding NEO-PI-3 item were first identified. Nine of the original ten PPM research validity scale items on the NEO-PI-R were identical to the items on the NEO-PI-3. Only one of the ten items wording had changed, but content was similar. The raw score and T-score for the PPM research validity scale was then calculated and transformed for the NEO-PI-3 according to the procedures outlined by Schinka et al. (1996). Once the data was error checked, the data was transferred and analyzed in IBM SPSS V25.0.

**Ethical Considerations.** The original data used in this study was archival data gathered from an independent psychology practice located in Hawaii. The owner of the data gave permission to use the archival data and was extensively involved in the research project. The data was obtained from airline pilot applicants from a commercial airline company. Prior to the administration of the psychological testing, the participants (airline pilot applicants) signed a consent form outlining their rights to confidentiality and allowing use of the data gathered for future consultation, research, or educational purposes. Tests were administered by airline personnel according to the standard instructions using online platforms (Pearson Q Global and PARIConnect). The current researcher did not have a conflicting role with the participants; the researcher did not administer any of the assessment measures. The participant's name was used only to assign a random code to the participant's data. Once the random code was assigned, all personal identifying information was removed from the data. The original test data is stored on password protected online databases on a computer in a secure facility, inside a locked area at the independent psychology practice, which also may be locked. The coded data is stored as a file on a password protected computer in a locked office in a secure facility. Access to the coded data was limited to only those researchers involved in the study who had signed confidentiality agreement forms as well. The use of archival data, where the researcher had no human participant interaction qualified as a IRB Level 1 certification and was certified on June 23, 2017 for one year. The coded data for the proposed study will be held for at least 3 years, from December 1, 2017 to December 1, 2020.

**Statistical Analyses.** Items selected for the PPM research validity scale for the NEO-PI-3 were selected based on similar content to the PPM research validity scale for the NEO-PI-R, not on item correspondence. Subsequently, the PPM research validity scale for NEO-PI-R was

calculated for the NEO-PI-3 pilot applicant profiles using the scoring procedures, mean, and standard deviations ( $M=20.25$ ,  $SD=4.66$ ) that were provided by Schinka et al. (1997).

Participant MMPI-2 and NEO-PI-3 profiles were then dichotomized into four groups. Two groups were created that classified all profiles as engaging and not engaging in positive response distortion if profiles obtained  $T \geq 65$  on any of the L, K, S scales of the MMPI-2. Additionally, all profiles were classified as engaging in positive response distortion using a cut-off raw score  $\geq 22$  on the PPM research validity scale on the NEO-PI-3. The Sd, So, and ODecp scales were calculated using the procedures outlined by Butcher et al., (2001).

Once the data were imported and coded, a data integrity check was conducted, assessing for errors and for outliers. One participant was removed from the mean differences analyses due to extreme scores. Once the data were determined to be error-free, a descriptive analysis was conducted examining the characteristics of the sample, including the range, mean, median, standard deviation, skewness, and kurtosis of the data. The variables were examined to check that the assumptions underlying the statistical techniques utilized in the study were not violated.

Following, a reliability analysis was conducted to assess whether the PPM research validity scale was reliable when applied to the NEO-PI-3. Furthermore, a principle components analysis and an exploratory factor analysis using varimax rotation were used to assess the underlying structure of the PPM research validity scale. A varimax rotation was used, as it was hypothesized that the factors that underlie the PPM research validity scale are independent of one another as the items that make-up the PPM research validity scale were intentionally designed to cut across the NEO-PI-R domains. Finally, follow-up correlations among the PPM research validity scale items and the MMPI-2 validity scales and an exploratory factor analysis among the PPM research validity scale and the MMPI-2 validity scales using a direct oblimin rotation were

conducted to further understand the PPM research validity scale. A direct oblimin rotation was used as the PPM research validity scale demonstrated positive, significant moderate to large correlations with the MMPI-2 validity scales measuring underreporting.

The inferential statistics utilized in the study included Pearson product moment correlations to explore the relationship among continuous variables, which included the validity and clinical scales of the MMPI-2, the factor and facet scales of the NEO-PI-3, and the PPM research validity scale of the NEO-PI-3. To further understand the PPM research validity as an indicator of positive response distortion, a principal component analysis using a direct oblimin rotation was conducted using the PPM research validity scale and the MMPI-2 validity measures of underreporting (L, K, S, Sd, So, and ODecp) to assess the latent factors of positive response distortion. A direct oblimin rotation was utilized as the underlying factors of positive response distortion, impression management and self-deception, which are hypothesized to be related. Additionally, an independent samples t-test comparing mean score differences on the PPM research validity scale between individuals who demonstrated positive response distortion and individuals who responded in a valid manner was conducted.

Though a discriminant function analysis was originally planned to examine the classification accuracy of the PPM scale, using the MMPI-2 validity scales as a criterion, examination of the groups revealed unequal group sizes. Thus, in accordance with the recommendations by Tabachnick and Fidell (2001), a logistic regression was conducted to account for the differences. Again, using the cut-off score of  $\geq 22$  on the PPM, the logistic regression assessed the utility of the PPM in discriminating individuals who demonstrated positive response distortion from individuals who responded validly. A logistic regression was appropriate because the study explored the predictive ability of the PPM scale on a categorical

dependent measure, profiles that demonstrated positive response distortion and valid profiles as classified by the MMPI-2 L, K, and S scales.

To examine the extent of agreement between cutoff scores for the NEO-PI-R PPM research validity scale as applied to the NEO-PI-3 and MMPI-2 K validity scale, Cohen's kappa was conducted. In order to conduct a kappa measure of agreement, the participants' profiles were coded as valid or invalid based on the cut-off scores of raw score  $\geq 22$  on PPM and  $T \geq 65$  on the K, L, or S scales. A kappa statistic was an appropriate analysis because there are two categorical variables, classification of valid profiles from the NEO-PI-3 and the MMPI-2, and each have an equal number of categories, valid or not valid.

A receiver operating characteristic (ROC) analysis was also conducted to assess the optimal cutoff score for the PPM research validity scale. The ROC curve was implemented to examine the sensitivity and specificity of the PPM research validity scale at various cutoff points. Sensitivity reflects the proportion of cases that were correctly identified as having the condition a test is measuring. Specificity represents the proportion of cases without the condition that were correctly classified by the measure. A ROC analysis was suitable for the current study, as previous research has indicated that a raw cutoff score of 22 on the PPM provides the best balance between sensitivity and specificity (Caldwell-Andrews et al., 2000; Morasco et al., 2007; Sellbom & Bagby, 2008)

In order to assess if positive response distortion had an effect on applicant profiles, a multivariate analysis of variance (MANOVA) was utilized. A MANOVA was chosen because it allowed the researcher to examine differences in NEO-PI-3 and MMPI-2 profiles between participants who demonstrated positive response distortion on the PPM research validity scale and participants who did not demonstrate positive response distortion on the PPM research

validity scale. Differences were examined using a variety of scales, including the Clinical and Content scales of the MMPI-2 and the Factor and Facet scales of the NEO-PI-3. This parametric test was utilized because there was one categorical independent variable, with two levels, those who demonstrated positive response distortion on the PPM research validity scale and those who did not, and multiple related continuous dependent variables. When significant multivariate relations were found to exist, follow-up univariate tests were conducted to determine specifically what dependent measures differed between applicants who demonstrated positive response distortion on the PPM and those with valid scores on the PPM. Follow-up analyses included applying a Bonferroni adjustment in order to reduce the probability of a Type 1 error.

Finally, to assess if and what differences existed between our sample, the normative sample used to derive the PPM research validity scale, the NEO-PI-3 Form S standardization sample, and other published DPG and ARD samples, means comparisons were conducted. Specifically, multiple independent samples t tests were conducted comparing the Factor and Facet scale means obtained between the current sample and the NEO-PI-3 Form S standardization sample. Further comparisons among our current sample, other published DPG and ARD samples, and the NEO-PI-3 standardization sample were conducted. A Bonferroni adjustment was applied in order to control for multiple comparisons. This parametric test was utilized because there was one categorical variable, with two levels, the current sample and the NEO-PI-3 Form S standardization sample, and one continuous dependent variable (the Factor and Facet scales being compared individually). Though a MANOVA could have also been implemented, there was not enough data to make such a comparison because only means and standard deviations were available for the NEO-PI-3 Form S standardization sample. Additionally, due to the unavailability of the NEO-PI-3 Form S standardization sample

individual data, the t distribution comparison of means was computed using a statistical calculator, which was cross-referenced by hand calculations.

**Rigor.** As methodological strengths, the current study utilized empirically validated measures of psychopathology and personality. Moreover, the current research employed a DPG methodological design incorporating actual personnel applicants where positive response distortion is suspected to have occurred rather than an ARD (Bagby & Marshall, 2003). The current research examined positive response distortion in a real-world setting rather than examining the distortion in a simulated setting, where respondents are prompted to fake-good. As discussed following review of Bagby and Marshall (2003), it is important to investigate positive response distortion in DPG samples, as there are significant differences between ARD samples instructed to fake good and DPG samples who are motivated to engage in positive response distortion.

Furthermore, though the archival data used in the study was administered over a period of approximately 4 years, the tests were administered online using group administrations, according to standardized instructions from test manuals. Therefore, there was consistency across administration, accounting for extraneous variables that would possibly have impacted the results of the study if the procedures used during each administration were not identical. Finally, because the data was exported from a database rather than hand-coded, it reduces the likelihood of human error. However, there are limitations of this research that are acknowledged in the limitations section of this clinical research project.



## CHAPTER III

### RESULTS

#### **Reliability of the PPM Research Validity Scale Applied to the NEO-PI-3**

A reliability analysis was conducted on the PPM research validity scale as applied to the NEO-PI-3. In the current study, the Cronbach alpha coefficient was .56. The mean, standard deviation, alpha coefficient, range and mean of inter-item correlations for the PPM research validity scale are presented in Table 1. The results in Table 1 suggest a low degree of internal consistency for the PPM research validity scale. Table 2 presents the correlations among the items. The heterogeneity and range of correlations ( $r = -.17$  to  $r = .42$ ) indicates that the items that make-up the PPM research validity scale are not measuring the same construct.

Principal components and principal axis factor analyses were also conducted to further examine the reliability of the PPM research validity scale. The first analysis included a principal component analysis (PCA) on the 10 items of the PPM research validity scale to examine how the items clustered. Prior to performing a PCA, the suitability of the data was assessed. The Kaiser-Meyer-Olkin value was .67, exceeding the recommended value of .6 (Kaiser, 1970) and Bartlett's Test of Sphericity (Bartlett, 1954) reached statistical significance supporting the factorability of the correlation matrix.

The principal component analysis revealed the presence of three components with eigenvalues exceeding 1, explaining 22.2%, 15.3%, and 11.3% of the variance respectively. An inspection of the screeplot revealed a break after the second component. Using Catell's (1966) scree test, it was decided to retain two components for further investigation. This was further supported by the results of the Parallel Analysis, which showed only two components with eigenvalues exceeding the corresponding criterion values for a randomly generated data matrix

of the same size (10 variables x 303 respondents). The two-component solution explained a total of 37.5% of the variance, with Component 1 contributing 22.2% and Component 2 contributing 15.3%. To aid in the interpretation of these two components, varimax rotation was performed. The highest items loading on Component 1 were items that corresponded to assertiveness and were made up of items from the Assertiveness, Vulnerability, and Self-Consciousness Facet scales of the NEO-PI-3. Fantasy Facet scale items loaded strongly on Component 2.

Next, a principal axis factor analysis with varimax rotation was conducted to assess the underlying structure for the ten items on the PPM research validity scale. The tests of assumptions were met indicating the suitability of the data for factor analysis. Principal axis factoring again, revealed the presence of three factors with eigenvalues exceeding 1. An inspection of the screeplot revealed a break after the second component. Using Catell's (1966) scree test, it was decided to retain two components for further investigation. This was again, further supported by the results of the Parallel Analysis, which showed only two components with eigenvalues exceeding the corresponding criterion values for a randomly generated data matrix of the same size (10 variables x 303 respondents). After rotation, the first factor accounted for 14.5% of the variance and the second factor accounted for 9.5% of the variance. However, after rotation, only the first factor had an eigenvalue greater than 1 and contained 6 of the 10 PPM research validity scale items. Follow-up reliability analyses of the first factor revealed a Cronbach alpha coefficient of .59. The second factor had large loadings from only 2 of the 10 PPM items. The other two items that make-up the PPM did not appear to load strongly onto either of the two factors. The initial communalities for the variables are rather low having a small amount of variance in common with one another (5% to 22%), indicating that the variables are weakly related to one another. The extraction communalities also indicate that the proportion

of the variable's variance that can be accounted for by the retained factors is generally low, except for item 9 (54%). Table 3 demonstrates the items and factor loadings for the rotated factors and the communalities.

Finally, a post hoc correlation analysis was also conducted on the items that make-up the PPM research validity scale and the MMPI-2 L, K, S, So, Sd, and ODecp validity scales, which are presented in Table 2. The results of the analysis indicated significantly small, positive correlations with four PPM research validity scale items and the L scale, and 3 significantly moderate, positive correlations with three PPM research validity items and the MMPI-2 L scale ( $p < .05$ ). The K scale had small, positive correlations with nine PPM research validity scale items ( $p < .05$ ). There were significantly small, positive correlations with six PPM research validity scale items and the S scale, and significantly moderate, positive correlations with three PPM research validity scale items and the S scale. ( $p < .05$ ). In addition, the ODecp scale had small, positive significant correlations with seven PPM research validity scale items and moderate, positive correlations with three PPM research validity scale items ( $p < .05$ ). There were small, significant positive correlations with seven PPM research validity scale items and the Sd scale, and one moderate, significant correlation ( $p < .05$ ). Finally, there were small, positive significant correlations with the So scale and six PPM research validity scale items, and moderate, positive correlations with the So scale and three PPM research validity scale items ( $p < .05$ ).

#### **Assessing the Validity and Utility of the PPM Research Validity Scale, using the MMPI-2 as a Criterion**

Classifying the pilot applicant profiles, 159 (52.5%) of participants had elevated T-Scores of  $T \geq 65$  on either the L, K, or S scales and 209 (69%) of pilot applicant profiles fell at or above the cut-off raw score  $\geq 22$  on the PPM research validity scale. Table 4 presents the percentage of

applicants who elevated on the L, K, S, or PPM scales separately. The means and standard deviations for the MMPI-2 validity scales and the PPM research validity scale is provided in Table 5.

Zero-order, one-tailed correlations were calculated between the PPM research validity scale as applied to the NEO-PI-3 and the MMPI-2 validity scales, which are presented in Table 6. Table 7 displays zero-order, one-tailed correlations among the PPM research validity scale, NEO-PI-3 factor scales, and MMPI-2 Clinical scales. The PPM research validity scale and all MMPI-2 validity scales measuring underreporting were positively and significantly correlated. Specifically, the PPM demonstrated significantly moderate, positive correlations with the L, K, Sd, and So scales ( $r > .40, p < .001$ ) and significantly strong, positive correlations with the S and ODecp scales ( $r > .50, p < .001$ ). In addition, PPM was significantly negatively correlated with the MMPI-2 F and F<sub>B</sub>, but not the F<sub>P</sub> scale. In particular, the PPM research validity scale demonstrated a significantly small, negative correlation with the F<sub>B</sub> scale ( $r = -.24, p < .001$ ), and a significantly moderate negative correlation with the F scale ( $r = -.34, p < .001$ ). Regarding measures of inconsistent responding, the PPM research validity scale as applied to the NEO-PI-3 had a significantly small, negative correlation with TRIN ( $r = -.22, p < .001$ ) and a significantly moderate, negative correlation with VRIN ( $r = -.45, p < .001$ ).

To further examine the PPM research validity scale as a reliable indicator of positive response distortion and the factors that underlie positive response distortion, a principal axis factor analysis using a direct oblimin rotation was conducted on the PPM and the MMPI-2 validity indicators. The Kaiser-Meyer-Olkin value was .81, exceeding the recommended value of .6 (Kaiser, 1970) and Bartlett's Test of Sphericity (Bartlett, 1954) reached statistical significance supporting the factorability of the correlation matrix. Principal axis factoring revealed the

presence of two factors with eigenvalues exceeding 1. An inspection of the screeplot revealed a break after the second component. Using Catell's (1966) scree test, it was decided to retain two components for further investigation. This was further supported by the results of the Parallel Analysis, which showed only two components with eigenvalues exceeding the corresponding criterion values for a randomly generated data matrix of the same size (seven variables x 303 respondents). After rotation, the total variance accounted for was 78%, where the first factor accounted for 60.6% of the variance and the second factor accounted for 17.4% of the variance. Table 8 demonstrates the items and factor loadings for the rotated factors and the communalities. K, S, and So loaded strongly on the first factor, and ODecp, Sd and L loaded strongly on the second factor. The PPM research validity scale loaded more strongly on the second factor. The interpretation of the two components is consistent with previous research on the underlying scales Edwards Social Desirability Scale (Esd; Edwards, 1957), the Wiggins Social Desirability Scale (Wsd; Wiggins, 1959), the Other Deception Scale (Od; Nichols & Greene, 1991), wo factor structure of positive response distortion. As indicated in previous research, K, So, and S loaded primarily on self-deception, and Sd, L, and ODecp loaded primarily on impression management (Bagby & Marshall, 2004; Paulhus, 1984). Therefore, Factor 1 was labeled self-deception and Factor 2 was labeled impression management. There was a strong positive correlation between the two factors ( $r = .50$ ).

An independent samples t-test was conducted to compare the mean scores for the PPM research validity scale between individuals who demonstrated positive response distortion on the L, K, or S scales and individuals who responded in a valid manner. The results revealed a significant difference in scores for individuals who demonstrated positive response distortion ( $M = 60.53$ ,  $SD = 8.1$ ) and individuals who responded in a valid manner ( $M = 52.90$ ,  $SD = 7.8$ ;  $t$

(301) = -8.31,  $p = .0005$ , two-tailed). The magnitude of the differences between the means (mean difference = -7.6, 95% *CI*: -9.4 to -5.8) was large ( $\eta^2 = .186$ ).

A logistic regression was also conducted to assess the ability of the PPM research validity scale, using a cutoff score of  $\geq 22$ , to correctly classify profiles as valid and invalid, based on the MMPI-2 validity scales L, K, and S. The model containing the PPM research validity scale was statistically significant, indicating that it reliably distinguished between valid and invalid profiles ( $\chi^2 = 34.45$ ,  $p < .0005$  with  $df = 1$ ). The model as a whole explained 10.7% (Cox and Snell R Square) and 14.3% (Nagelkerke R squared) of the variance in profile validity. Exp (B) value indicates that when PPM is raised by one unit (one raw score), the odds ratio is 4.6.

To determine the extent of agreement between the recommended cutoff score for the PPM research validity scale and MMPI-2 validity scales in identifying positive response distortion, a Cohen's kappa statistic and the overall classification accuracy were examined. Table 9 presents the results. The MMPI-2 L, K, and S scales were used as the criterion for comparison with the PPM research validity scale. The scales yielded the same classification accuracy in 66.3% of the cases. The PPM research validity scale demonstrated a sensitivity value of 83.6 percent (133/159) and a specificity value of 47.2 (68/144). There was fair diagnostic agreement between the PPM scale and the K, L, and S scales,  $K = .31$ ,  $p < .0005$ .

Further examining PPM's diagnostic efficiency, a ROC analysis was conducted to assess its classification accuracy and to detect the best cut point for positive response distortion. Predictive performance was assessed using the area under the curve (AUC), which is presented in Table 10. Additionally, Table 10 displays the cutoff score values derived from the ROC analysis and its sensitivity and specificity. The AUC in differentiating the positive response distortion and valid responding groups were .75 ( $SE = .028$ ). According to Youngstrom (2014),

traditional interpretation of this AUC is considered to be fair, meaning that the PPM research validity scale does fairly well at differentiating individuals who engage in positive response distortion versus those who respond in a valid manner. The corresponding Cohen's  $d$  for the AUC is  $d = .96$ , which is considered large (Rice & Harris, 2005). A cutoff score of 22 or greater on the PPM scale had the best balance for sensitivity (sensitivity = .83, specificity = .47), whereas a cutoff score of 23 or greater improved specificity (sensitivity = .77, specificity = .60). The base rate for the current sample is .52. The positive predictive power (PPP), or the accuracy rate of positive test results, using a cutoff score of 22 or greater, is .64. The negative predictive power (NPP), or the accuracy of negative test results, using a cutoff score of 22 or greater, is .72. Post hoc analyses using a PPM cutoff score of 23 did not significantly improve diagnostic agreement between the PPM research validity scale and the MMPI-2 L, K, and S scales,  $K = .37$ ,  $p < .0005$ .

### **Profile Differences**

Mean differences for the NEO-PI-3 factor and facet scales and the MMPI-2 clinical and content scales of applicants who demonstrated positive response distortion on the PPM scale and applicants who responded in a valid manner on the PPM were examined using a one-way between-groups multivariate analysis of variance (MANOVA). Preliminary assumptions testing was conducted to check for normality, linearity, and multicollinearity. Univariate and multivariate outlier testing resulted in one participant being excluded from the analyses. The assumption of homogeneity of variance-covariance matrices was violated when assessing mean differences between groups for the NEO-PI-3 factor and facet scales and the MMPI-2 clinical and content scales. According to Salkind (2007), parametric statistics, including the MANOVA, are robust enough, even when there is a violation to one of the assumptions using sample sizes of

30 or more. The current sample included 93 participants in the valid responding group and 209 participants in the positive response distortion group. Nevertheless, due to the unequal sizes between groups, and that the valid responding group produced larger variances and covariances, Pillai's Trace was used as the overall test statistic for comparing mean differences, as this statistic is more robust (Pallant, 2013; Warner, 2008). A value of  $p < .001$  was used, as the current sample includes a large total  $N$ , creating greater statistical power for the Levene's Test of Equality of Error Variances (Warner, 2008). The assumption for equality of error variances was violated at  $p < .001$  for 2 Clinical scales and 5 Content scales for the MMPI-2. Thus, a more conservative alpha level was used when determining significance for those variables in the univariate  $F$  test. Modifications to the alpha levels to detect significance on these comparisons were made using the Bonferroni adjustment to control for the risk of Type 1 error.

Table 11 summarizes NEO-PI-3 Factor scale scores and MMPI-2 Clinical scale scores of applicants who engaged in positive response distortion on the PPM scale compared with valid PPM responders. The results of the MANOVA revealed that participants who engaged in positive response distortion had significantly different NEO-PI-3 profiles,  $F(35, 266) = 6.22, p = .0005$ ; Pillai's Trace = .45; partial eta squared = .45 and MMPI-2 profiles,  $F(25, 276) = 5.27, p = .0005$ ; Pillai's Trace = .32; partial eta squared = .32. Univariate tests using a Bonferroni adjusted alpha level of .0014 indicated that invalid PPM responders had significantly lower scores on the NEO-PI-3 Neuroticism scale and significantly higher scores on the NEO-PI-3 Extraversion, Agreeableness, and Conscientiousness scales. Further, invalid PPM responders obtained significantly lower scores on all six of the Neuroticism facet scales, higher scores on five Extraversion facet scales (all except E5), lower scores on one Openness facet scale (O1) and higher scores on 2 Openness facet scales (O4 and O5), higher scores on four Agreeableness facet



scales (A1, A2, A3, A6), and significantly higher scores on all six Conscientiousness facet scales.

Univariate tests using a Bonferroni adjusted alpha level of .002 for variables that met the assumption for equality of error variance and an alpha level of .001 for those that did not meet the assumption for equality of error variance revealed that invalid PPM responders had significantly lower scores on MMPI-2 Clinical Scales 2 and 0 and a significantly higher score on Scale 3. Scale 8 approached a statistically significant difference at  $p = .002$ . Respondents that demonstrated positive response distortion on the PPM scale also scored significantly lower on 13 Content scales including, ANX, OBS, DEP, HEA, ANG, CYN, ASP, TPA, LSE, SOD, FAM, WRK, and TRT. A review of the effect sizes in these comparisons, expressed as partial eta squared, ranged from small to large.

Finally, post-hoc comparisons on the NEO-PI-3 Factor and Facet scale raw and T score means and standard deviations and on the MMPI-2 clinical and validity scales were calculated. Table 12 displays the T scores and standard deviation for the MMPI-2 Validity scales and Table 13 presents MMPI-2 T scores and standard deviations for Clinical scales for the current sample of airline pilots, the MMPI-2 standardization sample, other published studies of ARD and DPG samples (Bagby & Marshall, 2004; Butcher, 1989; Butcher, 1994, Detrick, Chibnall & Rosso, 2001; King, Schroeder, Manning, Retzlaff, & Williams, 2008). There does not appear to be substantial differences on the MMPI-2 Validity and Clinical scales between the means obtained by the current sample, and other DPG and ARD samples. Table 14 presents the raw score means and standard deviations for NEO-PI-3 Factor scales for the current sample of predominately males, the 1995 census-matched sample used to normalize the PPM research validity scale which included 200 men and 200 women (means and standard deviations of the normative sample were

presented separately for men and women), and the standardization sample for the NEO-PI-3 Form S, which consisted of 279 males and 356 females (McCrae & Costa, 2010; Schinka et al., 1997). Table 15 displays the T score means and standard deviations for the NEO-PI-3 Factor scales for the standardization sample for the NEO-PI-3 Form S and for the current sample, and T score means and standard deviations for the NEO-PI-R Factor scales for a DPG sample of 288 police officer applicants (Detrick & Chibnall, 2013), a DPG sample of 370 psychiatric patients, of whom 20 were classified as underreporting using a cutoff score of  $\geq 65T$  on at least two of the three standard MMPI-2 underreporting scales (L, K, and S; Sellbom & Bagby, 2008), and an ARD sample of 30 outpatient psychotherapy participants who were given specialized instructions to fake good when completing the NEO-PI-R (Ballenger et al., 2001). Comparisons between the NEO-PI-3 and the NEO-PI-R factor scales were permitted due to the near equivalence between the NEO-PI-R and the NEO-PI-3 (McCrae & Costa, 2010).

Examination of the means between the current sample and the PPM research validity scale normative sample indicates differences on all factor scales, where the current sample scored lower on the Neuroticism scale and higher on all four of the other factor domains. Multiple independent samples t tests were conducted analyzing the differences between profile means of the current sample and the NEO-PI-3 Form S standardization sample using a Bonferroni adjusted alpha level of .0014. The results of the analyses revealed that the means obtained on the factor scales for the current sample were significantly different from the means for the NEO-PI-3 Form S standardization sample at  $p < .0001$ . Specifically, the current sample scored significantly lower on the Neuroticism scale ( $M = 58.2$ ,  $SD = 19.3$ ) than the NEO-PI-3 Form S standardization sample ( $M = 82.7$ ,  $SD = 22.3$ ),  $t(936) = 16.41$ ,  $p < .0001$ , CI [21.57, 27.43]. The current sample obtained significantly higher means on the Extraversion

( $t(936)=11.63, p<.0001, CI[12.38, 17.4]$ ), Openness ( $t(936)=5.35, p<.0001, CI [4.30, 9.30]$ ), Agreeableness ( $t(936)=11.17, p<.0001, CI[11.05, 15.75]$ ), and Conscientiousness ( $t(936)=11.76, p<.0001, CI[13.08, 18.32]$ ) scales in comparison to the NEO-PI-3 Form S standardization sample. Regarding the facet scales, the means for the current sample were significantly different from the means for the NEO-PI-3 Form S standardization sample for 26 out of the 30 facet scales at  $p<.0001$ . The current sample obtained significantly lower means on all Neuroticism Facet scales, significantly lower means on two Openness Facet scales (Fantasy and Aesthetics), significantly higher means on three Openness Facet scales, and on five Extraversion, Agreeableness, and Conscientiousness Facet scales in comparison to the NEO-PI-3 Form S standardization sample. The results indicate that our current sample significantly differs from the NEO-PI-3 standardization sample.

Further independent samples t-test analyses using a Bonferroni adjusted alpha level of .0025 revealed similar differences among the current sample and published DPG and ARD samples when compared to the NEO-PI-3 Form S standardization sample. Figure 1 presents NEO-PI-R and NEO-PI-3 Mean T-Score profiles by group. The current sample of airline pilot applicants, police officer applicants (Detrick & Chibnall, 2013), DPG sample of underreporting psychiatric participants (Sellbom & Babgy, 2008), and ARD sample of faking good outpatient psychotherapy participants (Ballenger et al., 2001) scored significantly lower on the Neuroticism domain in comparison to the NEO-PI-3 Form S standardization sample at  $p\leq.0009$ . The current sample and the police officer applicant sample scored significantly higher on the Extraversion factor scale than the NEO-PI-3 Form S standardization sample, while the DPG sample of underreporting psychiatric participants scored significantly lower than the NEO-PI-3 Form S standardization sample at  $p\leq.0001$ . The current sample scored significantly higher on the

Openness Domain in comparison to the NEO-PI-3 Form S standardization sample and the police officer applicant and the DPG sample of underreporting psychiatric participants scored significantly lower on the Openness domain compared to the NEO-PI-3 Form S standardization sample at  $p \leq .0017$ . The current sample, DPG sample of underreporting psychiatric participants, and ARD sample of outpatient psychotherapy participants scored significantly higher on the Agreeableness domain in comparison to the NEO-PI-3 Form S standardization sample at  $p < .0001$ . Finally, the current sample, pilot applicant sample, and the DPG sample of underreporting psychiatric participants scored significantly higher on the Conscientiousness domain than the NEO-PI-3 Form S standardization sample at  $p < .0001$ . Thus, though our current sample significantly differs from the NEO-PI-3 Form S standardization sample, it displays similar characteristics to other published ARD and DPG samples classified according to demand characteristics and instructions to fake good.

## CHAPTER IV

### DISCUSSION

#### **Discussion of the Findings**

This study examined the PPM research validity scale developed for the NEO-PI-R and its application to the NEO-PI-3 and its ability to accurately identify individuals who are demonstrating positive response distortion in a high demand personnel selection context. In order to determine whether the PPM research validity scale developed for the NEO-PI-R was applicable to the NEO-PI-3, a reliability analysis was conducted on the PPM research validity scale for the current sample of airline pilot applicants who were administered the NEO-PI-3 as part of their hiring process. The results of the reliability analysis did not support the first hypothesis, according to traditional views regarding the psychometric properties of scale (Clark & Watson, 1995).

In the current study, the alpha coefficient of .56 for the PPM research validity scale applied to the NEO-PI-3 is roughly comparable to prior research reports. According to Schinka et al. (1997), the NEO-PI-R PPM research validity scale demonstrated good internal consistency, with a Cronbach alpha coefficient reported of .60. The alpha coefficient obtained in this study using the NEO-PI-3 was comparable to other studies investigating the NEO-PI-R PPM research validity scale. Reid-Seiser and Fritzsche (2001) reported NEO-PI-R PPM alpha coefficients of .70 for an ARD sample of students, .50 for a sample of students given standard instructions, and .52 for a DPG sample of customer service representatives. In a review of the PPM research validity scale, Blanch et al. (2009) examined 15 studies using the NEO-PI-R PPM and NPM research validity scales in normative, personnel selection, and clinical samples. Of the 15 studies examined, 8 reported PPM alpha coefficients ranging from .43 to .70. Three of the 8 studies were

conducted in employment contexts. The PPM alpha coefficients for those studies ranged from .50 to .60.

In their study, Schinka et al. (1997) acknowledged that the alpha coefficient for the PPM research validity scale was lower than those of personality scales that measure a single construct. Schinka et al. accounted for such differences by stating that the items that make up the PPM research validity scale were purposefully selected to cut across various domains of behavior and emotion so as not to be contaminated or saturated by a true personality domain or facet. Thus, a lower Cronbach alpha coefficient was expected as Cronbach alpha usually assumes unidimensionality (Raubenheimer, 2004). To address these issues analyses were conducted to examine the underlying component and latent factor structure of the PPM research validity scale. The analyses revealed that the PPM research validity scale is composed of two factors. Thus, on the surface, the results appear to support the multidimensionality of the PPM research validity scale and confirm Schinka et al.'s derivation of the PPM research validity scale. However, the reliability of the factor analysis is questionable due to the low inter-item correlations, which will be discussed further, and low initial and extraction communalities. Further examination of the factors indicated that the PPM research validity scale is functioning as a unidimensional scale, as only one factor retained an eigenvalue greater than 1 following rotation. The factor had moderate item loadings from six PPM research validity scale items. However, this first factor had low reliability. The second factor had generally weak factor loadings, where only two items loaded strongly. Therefore, the second factor did not have enough items to create a stable second factor. To this researcher's knowledge, the current study was the first to report a principal component and principal axis factor analyses on the PPM research validity scale. Thus, more research needs to be conducted investigating the underlying structure of the PPM research validity scale in other

samples, as prior research has also indicated that participant selection criteria can influence the factor solutions obtained (Gaskin, Lambert, Bowe, & Orellana, 2017).

Further examination of the reliability and internal consistency of the PPM research validity scale as applied to the NEO-PI-3 revealed a low mean inter-item correlation. Though Clark and Watson (1995), recommend an acceptable range of .15-.50 for the inter-item correlation, Kline (1979) asserts that an optimal range for item inter-correlations should be lower than .30 to avoid redundancy (as cited in Boyle, 1991). Further, Kline (1979) stated, "...if one constructs items that are virtually paraphrases of each other, the results would be high internal consistency and very low validity" (p. 3, as cited in Boyle, 1991). According to Kline (1986), "maximum validity...is obtained where test items do not correlate with each other, but where each correlates positively with the criterion. Such a test would have only low internal-consistency reliability" (p.3, as cited in Boyle, 1991). To address this issue, a post hoc correlation analysis on the items that make-up the PPM research validity scale and the MMPI-2 L, K, S, So, Sd, and ODecp validity scales indicated small to moderate relationships among many of the PPM research validity scale items and the MMPI-2 validity scales measuring underreporting. Thus, in accordance with Kline (1979), though the items of the PPM research validity scale, applied to the NEO-PI-3, demonstrated low internal-consistency reliability, examination of the correlations between the PPM items and the MMPI-2 validity scales indicated that most of the items were positively related to established and other MMPI-2 validity scales used to assess underreporting. Caution regarding these findings may be warranted due to the large sample size, which strongly influences rho, where small correlations were statistically significant, though accounting for only a small amount of variance (Pallant, 2013). Rosenthal and Rubin (1982) argue however, that Pearson  $r^2$  grossly underestimates the importance of a

finding. Thus, the significant correlations among the PPM research validity scale items and the MMPI-2 validity scale measures of underreporting support the items' criterion-related validity, despite the poor correlations they have with one another.

Consistent with traditional views of the psychometric properties of a scale, the reliability analysis suggests that the PPM research validity scale as applied to the NEO-PI-3 is not a reliable measure and has low internal consistency for the current sample of airline pilot applicants despite the fact that it demonstrates robust criterion-related validity. This issue deserves an explanation. Analysis of the PPM research validity scale items indicates that airline pilot applicants are endorsing the PPM research validity scale items differentially. The median score for the items for 7 of the 10 PPM research validity scale items was a 3, while two items had a median score of 2, and one item had a median score of 1. Also, two items on the PPM research validity scale were being endorsed almost equally in opposing directions. Examination of the factor analysis also indicated that 4 of the 10 PPM research validity scale items accounts for little variance and does not appear to fit well with the other items of the PPM research validity scale. The findings therefore suggest that the PPM research validity scale may not be a reliable measure when applied to the NEO-PI-3. These conclusions are further supported as the means comparisons analyses among the current sample of airline pilot applicants, published DPG and ARD samples, and the NEO-PI-3 Form S standardization sample revealed that the current sample significantly differed from the NEO-PI-3 Form S standardization sample in ways that were similar to other published DPG and ARD samples. This is not surprising. The DPG studies were typically conducted in high demand situations. Examination of MMPI-2 clinical and validity scale means among the current sample of airline pilot applicants, the MMPI-2 standardization sample, and other published DPG and ARD samples also indicated that the current sample did not appear to



substantially differ from the other samples. Therefore, it seems unlikely that the current sample of airline pilot applicants is a cause for the PPM research validity scale's poor reliability.

Altogether, this suggests again, that the PPM research validity scale may not be a reliable scale when applied to the NEO-PI-3.

It is noted that comparisons between the current sample and other published DPG and ARD samples were made between the NEO-PI-3 and the NEO-PI-R, which may have impacted the findings of the study. It is unlikely however, as McCrae and Costa (2010) acknowledge the near equivalency between the NEO-PI-R and the NEO-PI-3. On the other hand, it was observed that the respondents were endorsing the PPM items differentially, which may have accounted for the low inter-item correlations. Demographically and qualitatively, the current sample differs from that of the general population. These differences may have contributed to the poor reliability of the PPM research validity scale. However, due to the current research findings which include, a low alpha coefficient and low inter-item correlations, poor factorability of the PPM research validity scale, that the current sample is not uniquely influencing the reliability of the scale, and prior research reporting lower than expected alpha coefficients for the PPM in various samples (Blanch et al., 2009) altogether suggests that the PPM research validity scale may simply be unreliable. Though further research should be conducted, as this is the first study investigating the applicability of the NEO-PI-R PPM research validity scale to the NEO-PI-3, it is doubtful that future studies will be able to find that the PPM research validity scale performs better than what has already been demonstrated. Therefore, future research should focus on re-derivation of the PPM research validity scale for the NEO-PI-3.

The second hypothesis stated that the PPM research validity scale will demonstrate concurrent validity using the MMPI-2 underreporting validity scales as a criterion. The third

hypothesis specified that the PPM research validity scale will demonstrate discriminant validity using the MMPI-2 overreporting validity scales as a criterion. Both hypotheses were supported. The PPM research validity scale was significantly related with MMPI-2 validity indicators in the expected directions. Additionally, as identified by the MMPI-2 L, K, and S scales, individuals who demonstrated positive response distortion scored significantly higher on the PPM research validity scale compared to valid responders. The effect size was also large.

When examining the factor analytic structure of positive response distortion, the fourth hypothesis, which stated that there will be two latent factor structures that underlie positive response distortion and that the PPM research validity scale will load strongly onto the factor related to impression management was also supported. The factors extracted were consistent with impression management and self-deception response styles and parallel prior research that indicates that K, S, and So load primarily on self-deception and that L, ODeep and Sd load primarily on impression management (Bagby & Marshall, 2004; Paulhus, 1984). The results not only support the robustness of these latent factors but also provide evidence for the MMPI-2 validity scales and their ability to measure such factors in a DPG sample. Furthermore, the findings suggest that PPM primarily loads on impression management, indicating that the validity scale is likely detecting individuals who are intentionally attempting to fake good. However, it is noted, that PPM had the lowest factor loading on both factors, and though primarily and moderately loaded on the impression management factor, it cross-loaded similarly on the self-deception factor. The findings may indicate that PPM is related to both stylistic and substantive response styles. As reported by Morey et al. (2002), who found a strong correlation between latent stylistic and substantive factors, the PPM (and NPM) research validity scale is “heavily intertwined with the respondent’s functional status...in light of the magnitude of this

estimated relationship between substantive and stylist factors, it would be a serious error in interpretation to consider elevations on scales such as NPM and PPM as de facto evidence of effortful response distortion” (p.595).

The strong correlation between the two latent factors also supports the relationship between impression management and self-deception, therefore, indicating the difficulty in dichotomously discriminating between individuals who are consciously engaging in positive response distortion and respondents who have a dispositional tendency to view themselves positively. On the other hand, the weak loadings for PPM may be attributed to the poor psychometric properties of the scale. While the results find support for the Impression Management and Self-Deception factor scales, more research is needed.

The fifth hypothesis which stated that the PPM research validity scale will accurately discriminate between participants who demonstrated positive response distortion and participants who responded in a valid manner was supported. The results indicated that the PPM research validity scale significantly distinguished respondents engaging in positive response distortion from valid responders. This indicates that the PPM research validity scale has potential utility in detecting positive response distortion when compared to the MMPI-2 validity scales. However, the PPM research validity scale only accounted for a small amount of variance between prediction and grouping.

The sixth hypothesis stated that there will be a moderate agreement between the PPM research validity and the established underreporting validity scales for the MMPI-2. This hypothesis was rejected. The overall percent agreement between the PPM and MMPI-2 L, K, and S scales in identifying individuals who are and who are not engaging in positive response distortion was 66.3%. Cohen’s kappa revealed only a fair agreement (Landis & Koch, 1977).

The Cohen's kappa for the current study is similar to the kappa obtained by Morasco, et al. (2007), who also investigated the utility of the PPM research validity scale, using the MMPI-2 as a criterion. Like Morasco et al. (2007), the PPM research validity scale yielded relatively high rates of false positive classification. For the current study, the PPM research validity scale identified 76 pilot applicant profiles as demonstrating positive response distortion in comparison to the MMPI-2 L, K, and S scales, which classified those participant profiles as valid. Therefore, the diagnostic accuracy of the PPM research validity scale is fair.

The cutoff points that best optimized sensitivity and specificity for the PPM research validity scale varied. In accordance with prior research (Caldwell-Andrews et al., 2000; Morasco et al., 2007; Sellbom & Bagby, 2008), a cutoff score of  $\geq 22$  for the current sample revealed optimal sensitivity. However, using a cutoff score of  $\geq 22$ , specificity for the PPM research validity scale was generally weak. Increasing the cutoff score to 23 on the other hand, improved PPM's specificity. Thus, for this current sample, using a cutoff score of 23 may be more optimal in accurately differentiating between profiles who engage in positive response distortion and those who respond in a valid manner. Nonetheless, post hoc analyses using a PPM cutoff score of 23 did not significantly improve diagnostic agreement between the PPM research validity scale and the MMPI-2 L, K, and S scales. Further, when using a cutoff score of 23 there is still a high risk for false positives.

Finally, the last hypothesis proposing that there is a significant difference in profiles between individuals who demonstrated positive response distortion on the PPM research validity scale and individuals who did not demonstrate positive response distortion was supported. It is noted that there were violations to the assumptions of conducting a MANOVA, and statistical procedures could also have been applied, such as randomly removing members from the group

of individuals who demonstrated positive response distortion on the PPM in order to create equal numbers per group, possibly controlling for equality of error variances. For this study, violations to the error variance assumption was accounted for by adjusting the alpha level to be more conservative, in order to control for Type 1 error. The findings of the analysis revealed that individuals who demonstrated positive response distortion obtained lower scores on the NEO-PI-3 Neuroticism Scale, and higher scores on the Extraversion, Agreeableness, and Conscientiousness scales. This is similar to findings from prior research that investigated profile differences on the NEO-PI-R between participants who demonstrated positive response distortion and participants who did not (Ballenger et al., 2001; Detrick et al., 2010; Juhel et al., 2012; Sellbom & Babgy, 2008). Individuals who demonstrated positive response distortion on the PPM research validity scale were observed to have significantly lower scores on all six of the Neuroticism facet scales, higher scores on five Extraversion facet scales, lower scores on one Openness facet scale and higher scores on two Openness facet scales, higher scores on four Agreeableness facet scales, and significantly higher scores on all six Conscientiousness facet scales.

When comparing the means between applicants who demonstrated positive response distortion and applicants who responded in a valid manner on the MMPI-2 Clinical and Content scales, the results of the analyses revealed that respondents who demonstrated positive response distortion scored significantly lower on MMPI-2 Clinical Scales 2 and 0 and significantly higher on Scale 3 in comparison to valid responders. The results obtained for the current study are similar to that obtained by Morasco et al. (2007), who found a significant difference on Scales 2, 7, and 0. When examining the face valid Content scales, applicants who demonstrated positive response distortion scored significantly lower on 13 out of the 15 content scales. Butcher,

Morfitt, Rouse, and Holden (1997), analyzed profile differences between respondents who were classified into valid and invalid groups based on the MMPI-2 L and K scale scores in a sample of 271 applicants for airline pilot positions. Butcher et al. (1997) found significant differences between the two groups on all of the Content scales. The implications of Butcher et al.'s (1997) and the current study's results suggest that the applicants who did not demonstrate positive response distortion were more likely to acknowledge psychological flaws. This is also consistent with the findings of Bagby and Marshall (2004) who found that motivation to underreport and the minimization of psychopathological symptoms are associated.

For the current study, individuals who demonstrated positive response distortion also scored significantly higher on Scale 3. Butcher (1994) investigated the MMPI-2 in a sample of 437 airline pilot applicants. In his study, Butcher found that airline pilots obtained lower mean scale scores on all of the MMPI-2 Clinical scales in comparison to the normative sample, except on Scale 3. The current sample's Scale 3 mean score for pilot applicants who demonstrated positive response distortion is similar to the Scale 3 mean score obtained by Butcher. However, no direct explanation was provided as to why pilot applicants scored higher on Scale 3 in comparison to the normative sample. Butcher conducted a principal-components analysis on the three validity scales (L, F, K) and the 10 Clinical scales of the MMPI-2 for the 437 pilot applicant and found that Scale 3, along with L and K positively loaded on what he summarized as the Defensiveness factor.

Thumin (2002) also investigated the characteristics of the MMPI-2 in a sample of 82 applicants for an auditing position and found that Scale 3 correlated positively with L and K. Thumin (2002) suggested that together, L, K, and Scale 3 form a favorable or desirable cluster for well-educated job applicants, which is applicable to the current sample. Thumin (2002)

reported that the items on Scale 3 are two-fold, those reflecting somatic complaints and those that indicate the individual is well-socialized and well-adjusted. The latter is expected in motivated job applicants. In a previous study conducted by Thumin (1994), Scale 3 correlated significantly and positively with a measure of social desirability and with a scale that assessed for how kind and compassionate person is. Thus, as applied to the current sample, a correlation analyses revealed that Scale 3 is positively and significantly correlated with the L scale ( $r = .23$ ,  $p < .001$ ), K scale ( $r = .46$ ,  $p < .001$ ), and S scale ( $r = .41$ ,  $p < .001$ ). Therefore, that Scale 3 may be a reflection of socially desirable responding, which may account for the significantly higher mean score for pilot applicants who demonstrated positive response distortion, in comparison to pilot applicants who responded in a valid manner.

### **Clinical Implications**

The current study is consistent with previous research reports that personnel applicants engage in response distortion on self-reported measures administered as part of their selection process. In the current sample, more than half of the applicants demonstrated positive response distortion as identified by the L, K, and S scales of the MMPI-2 and the PPM research validity scale. The results of the study demonstrated that positive response distortion affects the clinical interpretability of the psychological measures of personality, specifically, the NEO-PI-3 and the MMPI-2. The findings revealed significant differences between profiles of individuals who demonstrated positive response distortion on the PPM research validity scale and profiles of individuals who did not demonstrate positive response distortion on the PPM research validity scale. Profiles of individuals who demonstrated positive response distortion on the PPM scored significantly lower on the Neuroticism scale and higher on the Extraversion, Agreeableness, and Conscientiousness scales of the NEO-PI-3 and significantly lower on Clinical Scales 2, and 0,

and 13 Content scales, and higher on Clinical Scale 3 for the MMPI-2. These significant profile differences can lead to inaccurate clinical interpretation, as these factors have found to be related to job performance. For instance, Piedmont and Weinstein (1994) found that neuroticism was negatively related to job performance, that conscientiousness and extraversion are significant predictors of success, and that agreeableness provided additional insight into interpersonal activities. Furthermore, other studies have found that conscientiousness is a central determining variable in job performance and can be a generalizable predictor of work motivation (Barrick & Mount, 2005; Schmidt & Hunter, 1998). Thus, significant differences in profile elevations, if not accounted for, can lead to inaccurate hiring decisions, which can then snowball into other serious consequences. Therefore, the clinical implications of these results indicate that clinicians need to be aware that positive response distortion occurs in personnel selection settings and that accurately assessing the validity of personality profiles prior to interpretation is vital.

In the current study, the PPM research validity scale demonstrated concurrent and discriminant validity, using the MMPI-2 validity scales as a criterion. Specifically, individuals who demonstrated positive response distortion, as identified by the L, K, and S scales of the MMPI-2, obtained significantly higher scores on the PPM. Furthermore, the PPM validity scale demonstrated moderate to strong positive correlations with the L, K, S, ODecp, Sd, and So scales (all measures of underreporting) and small to moderate negative correlations with the F and F<sub>B</sub> scales (all measures of overreporting). When added to the L, K, and S scales, the PPM research validity scale significantly contributed to the model and reliably distinguished valid and invalid profiles. Additionally, many of the individual items that make up the PPM research validity scale positively and moderately correlate with the L, K, S, ODecp, Sd, and So scales. The PPM



research scales demonstrates clinical utility in identifying participants who engage in positive response distortion.

Consistent with previous research reports, the study found support for the latent two factor structure of positive response distortion – impression management and self-deception. The findings indicated that K, S, and So load primarily on self-deception and that L, Sd, ODecp, and PPM load primarily on impression management. These findings can assist mental health professionals in making more informed interpretations regarding the applicant's motivation and response approach to the assessment. However, the strong relationship between the two factors highlight the overlap between impression management and self-deception and the difficulty in discriminating between effortful intent to deceive and a dispositional tendency to think positively of oneself. Although the results of the study can provide mental health professionals with additional information regarding the respondent's approach, without further research, existing measures of impression management and self-deception should continue to be utilized.

The current study was not able to substantiate the reliability of the PPM research validity scale as applied to the NEO-PI-3 in the current sample of airline pilot applicants. Though the scale demonstrated a comparable Cronbach alpha, the correlations among the items are weak, suggesting that the items are measuring different constructs. Agreement between the PPM research validity scale and the MMPI-2 L, K, and S scales in identifying profiles who did and did not demonstrate positive response distortion was low. Additionally, the cut-off score that best balances sensitivity and specificity for the PPM research validity, which will be discussed further, varied. Using the current recommended cut-off score in accordance with prior research, resulted in a high rate of false-positives. Therefore, before implementing the PPM research validity scale applied to the NEO-PI-3 in routine clinical practice, more research needs to be

done investigating the reliability of the scale in other samples, including airline pilot applicants. If further research continues to demonstrate the PPM's weak psychometric properties, than future research should focus on re-derivation of the PPM research validity scale.

### **Limitations**

There were several limitations apparent in the current study. First, there was minimal demographic information collected, limiting the generalizability of the findings. Furthermore, the study utilized archival data obtained from a sample of airline pilot applicants seeking employment at one airline company. Though the sample included a large number of applicants, the specific airline company utilized may attract a unique subset of airline pilot applicants, which again may limit the generalizability of the findings.

On the other hand, it is important to note that the findings for the current study may not be as unique to the current sample of pilot applicants. As discussed, though it appears that the current sample responded to the PPM research validity scale differently in comparison to the sample used to derive and normalize the PPM research validity scale, the applicants displayed similar characteristics to other published DPG and ARD samples. Thus, the PPM research validity scale appears to have generally weak reliability, which is consistent across previous research that have obtained internal consistency coefficients of .43 to .70 (Blanch et al., 2009). Notably, despite the poor reliability, the PPM research validity scale demonstrated robust convergent, discriminant, and criterion-related validity, demonstrated in the correlations between predictor and criterion variable and the ability of the scale to discriminate groups. But these findings may be limited to the design of the current study and may be attributed to common method variance (personal communication Dan Sass, March 26, 2018). Common method variance (CMV) may be a concern when self-report questionnaires are used to collect data at the

same time from the same participants. It is “variance that is attributable to the measurement method rather than to the construct the measures represent” (Podsakoff, Mackenzie, Lee, & Podsakoff, 2003, p.879). Common method bias describes measurement error that is compounded by the sociability of respondents who want to provide positive answers (Chang, van Witteloostuijn, & Eden, 2010). The most worrisome example of CMV occurs when “the data for both the predictor and the criterion variable are obtained from the same person in the same measurement context using the same item context and similar item characteristics” (Ling & Hiep, 2014, slide 3). Methods to test for CMV have been described in the literature, especially methods using variations of confirmatory factors analysis, to prevent or assess for CMV (Chang et al., 2010; Podsakoff et al., 2003; Podsakoff, Mackenzie, & Podasakoff, 2012). Future research should investigate the construct and discriminant validity of the PPM research validity scale, controlling for CMV.

Another limitation was that only one self-report measure (the MMPI-2) was used as the criterion to assess and form groups based on positive response distortion. Though there are an abundance of studies supporting the utility of the MMPI-2 validity scales, they are not without error (Sellbom & Bagby, 2008). Sellbom and Babgy (2008) reported that the MMPI-2 validity scales are associated with error in terms of classifying invalid and valid response styles, which impact our current study. However, in accordance with Sellbom and Bagby, the current study attempted to minimize such errors by using stringent criteria to dichotomize groups.

Finally, only the PPM research validity scale of the NEO-PI-R as applied to the NEO-PI-3 was examined in comparison to the MMPI-2 validity scales. The NPM and the INC research validity scales of the NEO-PI-R as applied to the NEO-PI-3 need to be investigated. Thus, in order to expand the generalizability of the findings, continuous examination of the concurrent

and discriminant validity of all of the research validity scales of the NEO-PI-R as applied to the NEO-PI-3 should be compared to established validity scales of multiple measures in other personnel assessments across various regions. The information gained from exploring this relationship will add to the literature regarding the use of validity scale and could again, be beneficial in aiding mental health professionals in understanding the effects of response distortion on a respondent's profile, aiding in interpretation.

### **Recommendations for Future Research**

It is important that research examining the reliability of the PPM research validity scale applied to the NEO-PI-3 continue in order to investigate its psychometric properties in other samples. Previous research has applied the PPM research validity scale to personnel selection samples. However, to the researcher's knowledge, this is the first study conducted investigating the PPM research validity scale as applied to the NEO-PI-3 in a personnel selection sample. Furthermore, though the current study demonstrated that the characteristics of the current sample as assessed by comparison of means between the NEO-PI-3 factor and facet scales and MMPI-2 Validity and Clinical scales obtained by the current sample, standardization samples, and published ARD and DPG samples, were similar to that of other published DPG and ARD samples, some of the differences analyzed, were computed between NEO-PI-3 and NEO-PI-R factor and facet scales. Though McCrae and Costa (2010) acknowledge the similarities between the two measures, it would be important that other studies be conducted comparing these differences on NEO-PI-3 assessments. This would assist in providing further evidence that the weak reliability of the PPM research validity scale is not significantly influenced by the current sample.

However, due to the evidence accumulated in the current research indicating the scales poor reliability which is not accounted for by differences in the current sample, and prior research which found lower than expected reliability for the PPM, it is unlikely that future research will find that the PPM research validity scale performs any better than what has been demonstrated. It is therefore suggested, that future research should focus on re-derivation of the PPM research validity scale.

It may also be beneficial that future research investigates the construct validity of the NEO-PI-3 factor structure when positive response distortion is present. For instance, Marshall et al. (2005) examined the factorial validity of the factor structure of the FFM, which is a form of construct validity. Marshall et al. found that the factor structure of the FFM is stable across empirically derived socially desirable responding groups and in DPG groups. Thus, though the current study demonstrated that positive response distortion affects scale elevations, which in turn affects the clinical interpretability of the profile, future research should examine whether positive response distortion affects all aspects of profile validity.

Moreover, the balance between specificity and sensitivity using the PPM research validity scale deserves important clinical consideration. The rate of false positives identified by the PPM research validity scale is troublesome as this could lead to negative consequences. For instance, as applied to the current sample, if the examiner using the PPM research validity scale cutoff score was led to believe that the pilot applicant invalidated his psychological evaluation, this could lead to the loss of the applicant's job opportunity. On the other hand, current standard practice for psychological assessments for the current sample includes retesting the pilot applicants who demonstrate invalid or atypical profiles. Retesting taxes a significant amount of resources, including monetary expenses on behalf of the airline company who must pay for the

re-evaluation, and time and extended resources on behalf of the psychologist and the applicant. Therefore, the PPM research validity scale's high rate of false positives warrants continued research investigating the PPM research validity scale and its cutoff scores before implementing the validity scale in routine clinical practice.

However, when considering sensitivity in regards to the PPM research validity scale, that is, the scales ability to correctly identify individuals who are engaging in positive response distortion, may have more importance when it comes to the current sample. That is, the consequences of false negatives appear dire in comparison to false positives. Though, as discussed false positives can pose great financial and time impositions, false negatives can lead to the incorrect hiring of pilots based on their clinical profiles. As demonstrated, profiles of individuals who engage in positive response distortion significantly differs from individuals who respond in a valid manner. These differences, if not accounted for by positive response distortion, could lead to inaccurate conclusions regarding an individual's personality and behavior. As such, a psychologist may inaccurately conclude that an applicant profile is congruent with the qualifications of a pilot. If that applicant then gets hired as a pilot, the potential consequences are imaginably frightful. Therefore, though as mentioned, continued research is necessary before using the PPM research validity scale in clinical practice, and consideration of the PPM's sensitivity in regards to the sample and the outcome is vital.

Finally, to improve the diagnostic accuracy of the PPM research validity scale, future research should focus on examining and cross-validating the cutoff scores used to identify individuals engaging in positive response distortion from valid responders. Additionally, though the current research study focused on maximizing both sensitivity and specificity when

determining an optimal cutoff score, future researchers should consider whether or not one is more important than the other, depending on their outcome goal.

### **Conclusion**

One of the primary functions as a psychologist is to use psychological tests to gather information pertinent to the provision of clinical services (Ben-Porath & Waller, 1992). Psychologists are tasked to conduct psychological assessments to evaluate the suitability of a job applicant for positions within an organization, including pilot selection (Rothstein & Goffin, 2006). Research indicates that airline pilot applicants, along with others who take personality tests as part of their personnel selection process, attempt to minimize adjustment problems, engage in positive response distortion, and highlight attributes that align with demand characteristics (Butcher, 1994; Detrick, Chibnall, & Call, 2010). Further, the validity of personality measures has found to be severely reduced by response bias, specifically by positive response distortion in personnel selection (Christiansen, Burns, & Montgomery, 2005; Mueller-Hanson, Heggstad, & Thornton, 2003; Reid-Seiser & Fritzsche, 2001; Rothstein & Goffin, 2006).

Regarding airline pilot applicants, research has indicated that they are a unique group of individuals as they tend to be an unusually well-functioning group psychologically (Butcher, 1994). Furthermore, airline pilot applicants are commonly of above-average intelligence, making it possible therefore, to readily distinguish appropriate from inappropriate responses on measures used for aviator selection (North & Griffin, 1977). Therefore, the ability to detect positive response distortion and its effect on profile interpretation is vital. The purpose of the current research was to investigate the applicability of the NEO-PI-R PPM research validity scale, a

measure of positive response distortion, to the NEO-PI-3, a popular psychological test used in pilot personnel selection.

Though there continues to be an ongoing debate regarding the use of validity scales in personality testing, the current study demonstrated that a significant portion of the pilot applicant sample engaged in positive response distortion, as identified by established underreporting validity scales on the MMPI-2 (L, K, and S scales) and by the previously studied PPM research validity scale. Furthermore, the current study indicated that there were significant differences in profile elevations on the NEO-PI-3 and the MMPI-2 between individuals who demonstrated positive response distortion on the PPM research validity scale and individuals who demonstrated valid responding. Thus, response bias significantly affects aspects of profile validity and clinical interpretability. The results support previous research indicating that response bias has a significant effect on profiles and warrants continued use of response bias indicators in personnel selection.

When assessing the reliability and validity of the NEO-PI-R PPM research validity scale as applied to the NEO-PI-3, the current study found support for the scale's concurrent, criterion-related, and discriminant validity using the MMPI-2 validity scales as a criterion. The PPM research validity demonstrated clinical utility in identifying participants who engage in positive response distortion. However, the reliability of the PPM research validity scale was found to be weak in accordance with traditional views regarding the appropriate psychometric properties of a scale. In summary, the PPM research validity scale as applied to the NEO-PI-3 demonstrated low internal-consistency but was significantly and positively correlated with the MMPI-2 validity scales assessing for underreporting. The psychometric properties of the PPM research validity scale as applied to the NEO-PI-3 is weak, but not dissimilar to previous research reports.



Furthermore, using the MMPI-2 validity scales as a criterion, resulted in high rates of false positive classification using the recommended cutoff score of greater than or equal to 22 on the PPM. Therefore, continued research is needed to assess the reliability and validity of the PPM research validity scale and appropriate cutoff scores as applied to the NEO-PI-3 in other samples.

Overall, the results of the current study indicate that the NEO-PI-R PPM research validity scale has potential validity and utility in identifying individuals who demonstrate positive response distortion when applied to the NEO-PI-3. However, the reliability of the scale when applied to the NEO-PI-3 using standard indices of reliability is weak. Given the overall weak performance, in light of the previous research, the scale deserves work. Continued research investigating the use of the PPM research validity scale as applied to the NEO-PI-3 and developing a more psychometrically sound scale is important as the use of validity scales assists mental health professionals in accurately obtaining information to make informed clinical decisions. The ability to correctly interpret psychological measures is especially important in regards to pilot applicants, as this leads to hiring decisions. It is important for members of our society to be able to trust the airline company to hire candidates who will keep them safe. Validity scales help psychologists to ensure accurate interpretation of an individual's clinical profile, leading to a more competent workforce and increased safety for the members of society.

## References

- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Association.
- Bagby, R. M., & Marshall, M. B. (2003). Positive impression management and its influence on the revised NEO Personality Inventory: A comparison of analog and differential prevalence group designs. *Psychological Assessment, 15*(3), 333-339. doi:10.1037/1040-3590.15.3.333
- Bagby, R. M., & Marshall, M. B. (2004). Assessing underreporting response bias on the MMPI-2. *Assessment, 11*(2), 115-126. doi:10.1177/1073191104265918
- Bagby, R. M., Rogers, R., Nicholson, R. A., Buis, T., Seeman, M. V., & Rector, N. A. (1997). Effectiveness of the MMPI-2 validity indicators in the detection of defensive responding in clinical and non-clinical samples. *Journal of Personality Assessment, 9*(4), 406-413.
- Ballenger, J. F., Caldwell-Andrews, A., & Baer, R. A. (2001). Effects of positive impression management on the NEO Personality Inventory-Revised in a clinical population. *Psychological Assessment, 13*(2), 254-260. doi:10.1037//1040-3590.13.2.245
- Barrick, M. R., & Mount, M. K. (2005). Yes, personality matters: Moving on to more important matters. *Human Performance, 18*(4), 359-372. doi:10.1027/s15327043hup1804\_3
- Bartlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistical Society, 16*, 296-298.
- Ben-Porath, Y.S., & Waller, N.G. (1992). "Normal" personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment, 4*, 14-19.
- Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A., (2006). A meta-analytic investigation of job applicant faking on personality measures. *International*

- Journal of Selection and Assessment*, 14(4), 317-335.
- Blanch, A., Aluja, A., Gallart, S., & Dolcet, J. M. (2009). A review on the use of NEO-PI-R validity scales in normative, job selection, and clinical samples. *The European Journal of Psychiatry*, 23(2), 121-129.
- Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences*, 12(3), 291-294. doi:10.1016/0191-8869(91)90115-R
- Butcher, J. N. (1994). Psychological assessment of airline pilot applicants with the MMPI-2. *Journal of Personality Assessment*, 62(1), 31-44.
- Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring*. Minneapolis: MN: University of Minnesota Press.
- Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom W. G., & Kaemmer, B. (2001). *Minnesota Multiphasic Personality Inventory – 2 (MMPI-2)*. Minneapolis: MN: University of Minnesota Press.
- Butcher, J. N., Gucker, D. K., & Hellervik, L. W. (2012). Clinical personality assessment in the employment context. In J. N. Butcher (Ed.), *Oxford handbooks online* (pp. 1-35). Twin Cities, MN: Oxford University Press.
- Butcher, J. N., Morfitt, R. C., Rouse, S. V., & Holden, R. R. (1997). Reducing MMPI-2 defensiveness: The effect of specialized instructions on retest validity in a job applicant sample. *Journal of Personality Assessment*, 68(2), 385-401.
- Caldwell-Andrews, A., Baer, R. A., & Berry, D. T. R. (2000). Effects of response sets on NEO-PI-R scores and their relations to external criteria. *Journal of Personality Assessment*.

74(3), 472-488.

- Callister, K., King, R., Retzlaff, P., & Marsh, R. (1999). Revised NEO Personality Inventory of male and female U.S. Air Force pilots. *Military Medicine: An International Journal*, 164(12), 885-890.
- Can, S. (2011). Effects of stress caused by the public personnel selection examination on the performance of physical education and other teacher trainees in turkey. *Social Behavior and Personality*, 31(10), 1367-1378. doi:10.2224/sbp.2011.39.10.1367
- Carducci, B. K. (2009). *The psychology of personality: Viewpoints, research, and applications* (2<sup>nd</sup> ed.). United Kingdom: Wiley-Blackwell Publishing.
- Cascio, W. (1995). Whither industrial and organizational psychology in a changing world of work. *American Psychologist*, 50, 928-939.
- Catell, R. B. (1966). The scree test for number of factors. *Multivariate Behavioral Research*, 1, 245-276.
- Chang, S.J., van Witteloostuijn, A., Eden, L (2010). From the editors: Common method variance in international business research. *Journal of International Business Research*, 41(2), 178-184.
- Chapelle, W. L., Novy, P. L., Sowin, T. W., & Thompson, W. T. (2010). NEO PI-R Normative personality data that distinguish U.S. Air Force female pilots. *Military Psychology*, 22, 158-175. doi:10.1080/08995600903417308
- Christiansen, N. D., Burns, G. N., & Montgomery, G. E. (2005). Reconsidering forced-choice item formats for applicant personality assessment. *Human Performance*, 18(3), 267-307. doi:10.1207/s15327043hup1803\_4
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale

- development. *Psychological Assessment*, 7(3), 309-319.
- Costa, P. T., Jr., & McCrae, R. R. (1985). *The NEO Personality Inventory manual*. Odessa, FL: Psychological Assessment Resources.
- Costa, P. T., & McCrae, R. R. (1989). *NEO-PI/FFI: Manual supplement for use with the NEO Personality Inventory and the NEO Five Factor Inventory*. Odessa, FL: Psychological Assessment Resources.
- Costa, P. T., Jr., & McCrae, R.R. (1992). Normal personality in clinical practice: The NEO Personality Inventory. *Psychological Assessment*, 4, 5-13.
- Costa, P. T., Jr., & McCrae, R. R. (1997). Stability and change in personality assessment: The Revised NEO Personality Inventory in the year 2000. *Journal of Personality Assessment*, 68(1), 86-94.
- Crowne, D.P., & Marlow, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24, 349-354. doi:10.1037/h0047358
- Detrick, P., & Chibnall, J. T. (2013). Revised NEO Personality Inventory normative data for police officer selection. *Psychological Services*, 10(4), 372-377. doi:10.1037/a0031800
- Detrick, P., Chibnall, J. T., & Call, C. (2010). Demand effects on positive response distortion by police officer applicants on the Revised NEO Personality Inventory. *Journal of Personality Assessment*, 92(5), 410-415.
- Detrick, P., Chibnall, J. T., & Rosso, M. (2001). Minnesota Multiphasic Personality Inventory—2 in police officer selection: Normative data and relation to the Inwald Personality Inventory. *Professional Psychology: Research and Practice*, 32(5), 484-490. doi:10.1037/0735-7028.32.5.484
- Donovan, J. J., Dwight, S. A., & Hurtz, G. M. (2003). An assessment of the prevalence, severity, and verifiability of entry-level applicant faking using the randomized response technique.

*Human Performance*, 16, 81–106.

- Edwards, A. L. (1957). *The social desirability variable in personality assessment research*. New York: Dryden.
- Fitzgibbons, A., Davis, D., & Schutte, P. C., (2004) *Pilot personality profile using the NEO PI-R* (Report No. NASA/TM-2004-213237). Hampton, VA: National Aeronautics and Space Administration Langley Research Center.
- Furnham, A. F. (1997) Knowing and faking one's five-factor personality score. *Journal of Personality Assessment*, 69(1), 229-243. doi:10.1207/s15327752jpa6901\_14
- Gallen, R. T., & Berry, D. T. R. (1996). Detection of random responding in MMPI-2 protocols. *Assessment*, 3(2), 171-178.
- Gaskin, C. J., Lamber, S. D., Bowe, S. J., & Orellana, L. (2017). Why sample selection matters in exploratory factor analysis: implications for the 12-item World Health Organization Disability Assessment Schedule 2.0. *BioMed Central Medical Research Methodology*, 17(40), 1-9. doi: 10.1186/s12874-017-0309-5
- Geller, A. (2004). Now, tell the computer why you want this job: PCs take lead role in screening hourly workers. *Calgary Herald*, F.3.
- Goffin, R. D., & Christiansen, N. D. (2003). Correcting personality tests for faking: Review of popular personality tests and initial survey of researchers. *International Journal of Selection and Assessment*, 11(4), 340-344.
- Graham, J. R. (2012). *MMPI-2 assessing personality and psychopathology* (5<sup>th</sup> ed.). New York: Oxford University Press, Inc.
- Graham, J. R., Watts, D., & Timbrook, R. E. (1991). Detecting fake-good and fake-bad MMPI-2 profiles. *Journal of Personality Assessment*, 57(2), 264-277.
- Greene, R. L. (2000). *The MMPI-2: An interpretive manual* (2<sup>nd</sup> ed.). Needham Heights, MA: Allyn and Bacon.

- Heller (2005). Court ruling that employer's integrity test violated ADA could open door to litigation. *Workforce Management*, 84(9), 74-77.
- Holden, R. R., & Jackson, D. N. (1981). Subtlety, information, and faking effects in personality assessment. *Journal of Clinical Psychology*, 37(2), 379-386.
- Hsu, C. (2004). The testing of America. *U.S. News and World Report*, 137(9), 68-69.
- Jackson, D. N. (1970). A sequential system for personality scale development. In C.D. Spielberger (Ed.), *Current topics in clinical and community psychology* (pp. 62-97). New York: Academic Press
- Juhel, J., Brunot, S., & Zapata, G. (2012). Response distortion on the NEO-PI-R among candidates taking the entrance examination to the national school of civil aviation (ENAC-France). *Scientific Research*, 3(5), 393-398. doi:10.4236/psych.2012.35055
- Kaiser, H. (1970). A second generation Little Jiffy. *Psychometrika*, 35, 401-415.
- King, R. E., Schroeder, D. J., Manning, C. A., Retzlaff, P. D., & Williams, C. A. (2008). Screening air traffic control specialists for psychopathology using the Minnesota Multiphasic Personality Inventory-2 (DOT Publication No. DOT/ FAA/AM-08/13). Washington, DC: FAA Office of Aerospace Medicine.
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.
- Ling, S. L., Hiep, P. H. (2014). *The issue of common method variance or common method bias* [PowerPoint slides] Retrieved from <https://www.slideshare.net/>
- Lochner, K., & Nienhaus, N. (2016). *The predictive power of assessment for pilot selection*. Retrieved from <https://www.cut-e.com>
- Loevinger, J. (1959). Theory and techniques of assessment. *Annual Review of Psychology*, 10(1), 287-316.
- Lowmaster, S. A., & Morey, L. C. (2012). Predicting law enforcement officer job performance

- with the Personality Assessment Inventory. *Journal of Personality Assessment*, 94(3), 254-261. doi: 10.1080/00223891.2011.648295
- Marshall, M. B., De Fruyt, F., Rolland, J. P., & Bagby, R. M. (2005). Socially desirable responding and the factorial stability of the NEO-PI-R. *Psychological Assessment*, 17(3), 379-384. doi: 10.1037/1040-3590.17.3.379
- McCrae, R. R., & Costa, P. T., Jr. (2010). *NEO Inventories for the NEO Personality Inventory-3 (NEO-PI-3), NEO Five-Factor Inventory-3 (NEO-FFI-3), NEO Personality-Inventory-Revised (NEO PI-R) Professional Manual*. Lutz: FL: PAR.
- McCrae, R. R., & Costa, P. T., Jr. (1983). Social desirability scales: more substance than style. *Journal of Consulting and Clinical Psychology*, 51(6), 882-888.
- McCrae, R. R., & Costa, P. T. (1994). The stability of personality: Observations and evaluations. *Current Directions in Psychological Science*, 3, 173-175.
- McCrae, R. R., Costa, P. T., Jr., Dahlstrom, W. G., Barefoot, J. C., Siegler, I. C., & Williams, R. B., Jr. (1989). A caution of the use of the MMPI K-correction in research on psychosomatic medicine. *Psychosomatic Medicine*, 51, 58-65.
- McCrae, R. R., Stone, S. V., Fagan, P. J., & Costa, P. T., Jr. (1998). Identifying causes of disagreement between self-reports and spouse ratings of personality. *Journal of Personality*, 66(3), 285-313.
- McGrath, R. E., Mitchell, M., Kim, B. H., & Hough, L. (2010). Evidence for response bias as a source of error variance in applied assessment. *Psychological Bulletin*, 136(3), 450-470. doi:10.1037/a0019216
- Minter, A. (2015, March 29). Low-cost airlines are high stress for pilots. *Bloomberg*. Retrieved from <https://www.bloomberg.com>
- Morasco, B. J., Gfeller, J. D., & Elder, K. A. (2007). The utility of the NEO-PI-R validity scales to detect response distortion: A comparison with the MMPI—2. *Journal of*



- Personality Assessment*, 88(3), 276-283. doi:10.1080/00223890701293924
- Morey, L. C. (1991). *Personality Assessment Inventory: Professional Manual*. Odessa, FL: Psychological Assessment Resources.
- Morey, L. C. (2012). Detection of response bias in applied assessment: Comment on McGrath et al. (2010). *Psychology Injury and Law*, 5, 153-161. doi:10.1007/s12207-012-9131-x
- Morey, L. C., & Lanier, V. W. (1998). Operating characteristics of six response distortion indicators for the Personality Assessment Inventory. *Assessment*, 5(3), 203-214.
- Morey, L. C., Quigley, B. D., Sanislow, C. A., Skodol, A. E., McGlashan, T. H., Shea, M.T., ... Gunderson, J. G. (2002). Substance or Style? An investigation of the NEO-PI-R Validity Scales. *Journal of Personality Assessment*, 79, 583-599.
- Mueller-Hanson, R., Heggstad, E. D., & Thornton III, G. C. (2003). Faking and selection: Considering the use of personality from select-in and select-out perspectives. *Journal of Applied Psychology*, 88(2), 348-355. doi:10.1037/0021-9010.88.2.346
- North, R. A., & Griffin, G. R. (1977). *Aviation Selection 1919-1977 (NAMRL-SR-77-2)*. Pensacola, FL: Naval Aerospace Medical Research Laboratory.
- Nichols, D. S. (1991). *Development of a global measure for positive mental health*. Unpublished manuscript.
- Nichols, D. S., & Greene, R. L. (1991, March). *New measures for dissimulation on the MMPI/MMPI-2*. Paper presented at the 26<sup>th</sup> Annual Symposium on Recent Developments in the Use of the MMPI (MMPI-2/MMPI-A), St. Petersburg Beach, FL.
- Pallant, J. (2013). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (5<sup>th</sup> ed.). New York: The McGraw-Hill Companies.
- Paulhus, D. L. (1984) Two-component models of socially desirable responding. *Journal of*

- Personality and Social Psychology*, 46(3), 598-609. doi:10.1037/0022-3514.46.3.598
- Paulhus, D. L. (1989). *Manual for the Balanced Inventory of Desirable Responding: Version 6*. Unpublished manual, University of British Columbia.
- Paulhus, D. L. (1998). *Manual for the Balanced Inventory of Desirable Responding (BIDR-7)*. Toronto, Ontario, Canada: Multi-Health Systems.
- Paulhus, D. L. (2002). Social desirable responding: The evolution of a construct. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ: Lawrence Erlbaum.
- Paulhus, D. L., Bruce, M. N., & Trapnell, P. D. (1995). Effects of self-presentation strategies on personality profiles and their structure. *Personality and Social Psychology Bulletin*, 21(2), 100-108.
- Piedmont, R. L., & Ciarrochi, J. W. (1999). The utility of the Revised NEO Personality Inventory in an outpatient, drug rehabilitation context. *Psychology of Addictive Behaviors*, 13, 213-226.
- Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology*, 78(3), 582-593. doi:10.1037//0022-3514.78.3.582
- Piedmont, R. L., & Weinstein, H. P. (1994). Predicting supervisor ratings of job performance using the NEO Personality Inventory. *The Journal of Psychology*, 128(3), 255-265.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879-903. doi:10.1037/0021-

9010.88.5.879

- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, 63(1), 539-569. doi:10.1146/annurev-psych-120710-100452
- Quirk, S. W., Christiansen, N. D., Wagner, S. H., & McNulty, J. L. (2003). On the usefulness of measures of normal personality for clinical assessment: Evidence of the incremental validity of the Revised NEO Personality Inventory. *Psychological Assessment*, 15(3), 311-325. doi:10.1037/1040-3590.15.3.311
- Raubenheimer, J. (2004). An item selection procedure to maximise scale reliability and validity. *SA Journal of Industrial Psychology*, 30(4), 59-64.
- Reid-Seiser, H. L., & Fritzsche, B. A., (2001). The usefulness of the NEO PI-R Positive Presentation Management Scale for detecting response distortion in employment contexts. *Personality and Individual Differences*, 31, 639-650.
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, 29(5), 615-620. doi:10.1007/s10979-005-6832-7
- Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74(2), 166-169.
- Ross, S. R., Bailey, S. E., & Millis, S. R. (1997). Positive self-presentation effects and the detection of defensiveness on the NEO-PI-R. *Assessment*, 4, 395-408.
- Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, 16, 155-180. doi:10.1016/j.hrmr.2006.03.004

- Salkind, N. J. (2007). *Statistics for people who (think they) hate statistics: The Excel edition*. Thousand Oaks, CA: Sage Publishing, Inc.
- Scandell, D. J. (2000). Development and initial validation of validity scales for the NEO-Five Factor Inventory. *Personality and Individual Differences, 29*, 1153-1162.
- Schinka, J. A., Kinder, B. N., & Kremer, T. (1997). Research validity scales for the NEO-PI-R: Development and initial validation. *Journal of Personality Assessment, 68*, 127-138.
- Sellbom, M., & Bagby, R. M. (2008). The validity and utility of the Positive Presentation Management and Negative Presentation Management scales for the Revised NEO Personality Inventory. *Assessment, 15*(2), 165-176. doi:10.1177/1073191107310301
- Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*(1), 1–26.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4<sup>th</sup> ed.). Needham Heights, MA: Allyn & Bacon.
- Tellegen, A. (1982). *Brief manual of the Multidimensional Personality Questionnaire*. Unpublished manuscript, University of Minnesota.
- Thumin F. J. (1994). Correlations for a new personality test with age, education, intelligence, and the MMPI-2. *Perceptual and Motor Skills, 79*, 459-466.
- Thumin, F. J. (2002). Comparison of the MMPI and MMPI-2 among job applicants. *Journal of Business and Psychology, 17*(1), 73-86.
- Wagner, W. F. (2000). All skill, no finesse. *Workforce, 79*(6), 108-116.
- Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks: CA: Sage Publications, Inc.
- Widiger, T. A., & Costa, P. T., Jr. (2002). Five-factor model personality disorder research. In P. Costa & T. Widiger (Eds.), *Personality disorders and the five-factor model of personality* (2nd ed., pp. 59 – 87). Washington, DC: American Psychological Association.

- Widiger, T. A., Costa, P. T., Jr., Gore, W. L., & Crego, C. (2012). Five factor model personality disorder research. In T. Widiger & P. Costa (Eds.), *Personality disorders and the five-factor model of personality* (3<sup>rd</sup> ed.). Washington, DC: American Psychological Association.
- Wiggins, J. S. (1959) Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. *Journal of Consulting Psychology*, 23 419-427.
- Wiggins, J. S., & Trapnell, P. D. (1997). Personality structure: The return of the big five. In R. Hogan, J. A. Johnson, & S. Briggs (Eds.), *Handbook of Personality Psychology* (pp. 737-765). San Diego, CA: Academic Press.
- Young, M. S., & Schinka, J. A. (2001). Research validity scales for the NEO-PI-R: Additional evidence for reliability and validity. *Journal of Personality Assessment*, 76(3), 412-420.
- Youngstrom, E. A. (2014). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: We are ready to ROC. *Journal of Pediatric Psychology*, 39(2), 204-221. doi:10.1093/jpepsy/jst062

## Appendices

## Appendix A

*IRB Letter of Certification*

June 23, 2017

Jolene Young  
7160 Hawaii Kai Dr. #255  
Honolulu, HI 96825

jolenestacy@gmail.com

Dear Ms. Young

Your application, "An Investigation into the Applicability of the NEO-PI-R PPM Research Validity Scale to the NEO-PI-3 and the Effects of Positive Response Distortion," is fully certified by the Institutional Review Board as of 6-23-2017.

Please note that research must be conducted according to this application that was certified by the IRB. Your proposal should have been revised to be consistent with your application. Please note that you also need to abide by any requirements specified in your letter of permission. Any changes you make to your study need to be reported to and certified by the IRB.

Any adverse events or reactions need to be reported to the IRB immediately.

Your full application is certified for one year from 6-23-2017. Please be aware that if your study is not likely to be completed one year from 6-23-2017, you will need to file a **Continuing Review for IRB or Continuing Certification of Compliance** form with the IRB at least two months before that date to obtain recertification. If your proposal is not recertified within the year specified (365 days), your IRB certification expires and you must immediately cease data collection.

When you have completed your research, you will also need to inform the IRB of this in writing and complete the required forms. You may use the **Project Completion Report** form for this purpose. Records must be retained for at least three years.

Good luck with your research!

Please be careful not to use this letter.

If you have questions please feel free to contact me

Sincerely,

A handwritten signature in cursive script, appearing to read "Robert M. Anderson Jr.", followed by a small circular mark.

Robert M. Anderson Jr., Ph.D. Co-Chair  
Institutional Review Board

cc: Dr. Lianne Philhower



## Appendix B

Table 1

*Reliability of the PPM Research Validity Scale (N=303)*

<i>Scale</i>	<i>Alpha</i>	<i>Scale M</i>	<i>Scale SD</i>	<i>Inter-Item Correlation</i>	<i>Mean Inter-Item Correlation</i>
PPM	.56	23.49	4.10	-.17 to .42	.12

*Note:* PPM = positive presentation management; scale score range is 0 to 40.

Table 2

*Correlations Among PPM Research Validity Scale Items and  
MMPI-2 Validity Scales*

Variable	L	K	S	ODeccp	Sd	So	1	2	3	4	5	6	7	8	9	10
MMPI-2																
L	1.00															
K	.50***	1.00														
S	.57***	.85***	1.00													
ODeccp	.67***	.51***	.62***	1.00												
Sd	.59***	.23***	.32***	.83***	1.00											
So	.40***	.72***	.71***	.52***	.29***	1.00										
PPM Items																
1	.29***	-.26***	.32***	.26***	.12*	.30***	1.00									
2	.30***	.26***	.32***	.34***	.24***	.37***	.21***	1.00								
3	.07	.14*	.14*	.22***	.18**	.16**	.23***	.13*	1.00							
4	.11	.19**	.27***	.14*	.09	.25***	.08	.12*	.15*	1.00						
5	.38***	.28***	.30***	.48***	.37***	.29***	.27***	.23***	.25***	.08	1.00					
6	.20***	.24***	.24***	.29***	.20**	.32***	.12*	.31***	.30***	.15*	.28***	1.00				
7	.13*	.16**	.22***	.22***	.16**	.15*	.08	.15*	.00	.01	.04	.06	1.00			
8	.09	-.05	.06	.12*	.11	.05	.00	.08	-.01	-.02	.04	.09	.42***	1.00		
9	.26***	.18**	.20**	.19**	.22***	.13*	.11*	.15**	.08	.13*	.23***	.17**	-.17**	-.13*	1.00	
10	.32***	.25***	.31***	.31***	.20**	.16**	.16**	.18**	-.02	.01	.27***	.08	.03	-.01	.13*	1.00

*Note:*  $N=303$  for all scales, except for ODeccp ( $N=298$ ), Sd ( $N=295$ ), and So ( $N=299$ ) as scales were not able to be calculated for all participants due to missing items. MMPI-2 = Minnesota Multiphasic Personality Inventory-2; PPM = Positive Presentation Management scale; L = Lie scale; K = K scale; S = Superlative Self-Presentation Scale; ODeccp = Other Deception Scale; Sd = Wiggins' Social Desirability Scale; So = Edwards' Social Desirability Scale. \* $p<.05$ , two-tailed. \*\* $p<.01$ , two-tailed. \*\*\* $p<.001$ , two-tailed.

Table 3

*Factor Loadings for Principal Axis Factoring with Varimax Rotation  
of Two Factor Solution of PPM Research Validity Scale Items*

<i>PPM Item</i>	<i>Rotated Factor Loading</i>		<i>Initial Communality</i>	<i>Extraction Communality</i>
	<i>Factor 1</i>	<i>Factor 2</i>		
1	.59	.01	.23	.35
2	.52	.07	.21	.27
3	.48	.15	.17	.25
4	.41	.04	.14	.17
5	.40	-.01	.17	.16
6	.36	-.25	.13	.20
7	.29	-.00	.11	.08
8	.23	-.03	.05	.06
9	.07	.73	.21	.54
10	.03	.56	.19	.32

*Note:*  $N=303$ . Eigenvalue for Factor 1 following rotation = 1.45 (percentage of variance = 14.5); eigenvalue for Factor 2 following rotation = .95 (percentage of variance 9.5). PPM = Positive Presentation Management scale

Table 4

*Percentage of Participants Who Obtained Elevated PRD Scores on the L, K, or S, or PPM Scales (N=303)*

<i>Scale</i>	<i>N</i>	<i>Percentage</i>
L	44	14.5
K	112	37.0
S	147	48.5
PPM	209	69.0

*Note:* Profiles were classified as invalid or valid using a cut-off score  $T \geq 65$  on either the L, K, or S scale and a raw score of  $\geq 22$  on the PPM research validity scale. L = Lie scale; K = K scale; S = Superlative Self-Presentation Scale; PPM = Positive Presentation Management research validity scale.

Table 5

*Means and Standard Deviations for MMPI-2 Validity Scales and the PPM Research Validity Scale*

<i>Scale</i>	<i>M</i>	<i>SD</i>
PPM	56.9	8.8
L	51.0	11.1
K	60.8	8.4
S	63.2	9.7
ODecp	56.9	10.4
Sd	52.5	9.2
So	59.1	6.1
F	44.2	7.1
F <sub>B</sub>	43.5	4.0
F <sub>P</sub>	45.3	6.9
TRIN	53.2	4.6
VRIN	40.7	8.6

*Note:*  $N=303$  for all scales, except for ODecp ( $N=298$ ), Sd ( $N=295$ ), and So ( $N=299$ ) as scales were not able to be calculated for all participants due to missing items. PPM = Positive Presentation Management research validity scale; L = Lie scale; K = K scale; S = Superlative Self-Presentation Scale; ODecp = Other Deception Scale; Sd = Wiggins' Social Desirability Scale; So = Edwards' Social Desirability Scale; F = F scale; F<sub>B</sub> = Back F Scale; F<sub>P</sub> = Infrequency Psychopathology Scale; TRIN = True Response Inconsistency Scale; VRIN = Variable Response Inconsistency Scale.

Table 6

*Correlations Between the PPM Research Validity Scale and MMPI-2 Validity Scales*

Variable	PPM	L	K	S	ODecp	Sd	So	F	F <sub>B</sub>	F <sub>P</sub>	TRIN	VRIN
NEO-PI-3												
1. PPM	1.00											
MMPI-2												
2. L	.48***	1.00										
3. K	.42***	.50***	1.00									
4. S	.52***	.57***	.85***	1.00								
5. ODecp	.57***	.67***	.51***	.62***	1.00							
6. Sd	.43***	.59***	.23***	.32***	.83***	1.00						
7. So	.47***	.40***	.72***	.71***	.52***	.29***	1.00					
8. F	-.34***	-.19***	-.44***	-.47***	-.28***	-.12*	-.50***	1.00				
9 F <sub>B</sub>	-.24***	-.14**	-.43***	-.44***	-.17**	.03	-.50***	.63***	1.00			
10. F <sub>P</sub>	.08	.34***	-.08	-.05	.29***	.39***	-.07	.43***	.56***	1.00		
11. TRIN	-.22***	-.17***	-.34***	-.33***	-.32***	-.26***	-.47***	.32***	.18***	.01	1.00	
12. VRIN	-.45***	-.39***	-.64***	-.65***	-.49***	-.30***	-.70***	.33***	.33***	.06	.37***	1.00

*Note:*  $N=303$  for all scales, except for ODecp ( $N=298$ ), Sd ( $N=295$ ), and So ( $N=299$ ) as scales were not able to be calculated for all participants due to missing items. NEO-PI-3 = NEO Personality Inventory-3; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; PPM = Positive Presentation Management research validity scale; L = Lie scale; K = K scale; S = Superlative Self-Presentation Scale; ODecp = Other Deception Scale; Sd = Wiggins' Social Desirability Scale; So = Edwards' Social Desirability Scale; F = F scale; F<sub>B</sub> = Back F Scale; F<sub>P</sub> = Infrequency Psychopathology Scale; TRIN = True Response Inconsistency Scale; VRIN = Variable Response Inconsistency Scale. \* $p<.05$ , one-tailed. \*\* $p<.01$ , one-tailed. \*\*\* $p<.001$ , one-tailed.

Table 7

*Correlations Among the PPM Research Validity Scale, NEO-PI-3 Factor Scales, and MMPI-2 Clinical Scales (N=303)*

Variable	PPM	N	E	O	A	C	Hs	D	Hy	Pd	Mf	Pa	Pt	Sc	Ma	Si
NEO-PI-3																
1. PPM	1.00															
2. N	-.61***	1.00														
3. E	.23***	-.08	1.00													
4. O	.13*	-.03	.07	1.00												
5. A	.28***	-.28***	.06	.26***	1.00											
6. C	.48***	-.26***	.09	.13*	.12*	1.00										
MMPI-2																
7. Hs	.02	-.05	-.01	-.03	.16**	-.04	1.00									
8. D	-.25***	.29**	-.29***	-.12*	-.01	-.11*	.25***	1.00								
9. Hy	.22***	-.16**	.23***	.12*	.23***	.07	.57***	-.07	1.00							
10. Pd	-.13*	.15**	.11*	-.02	-.08	-.12*	.35***	.07	.29***	1.00						
11. Mf	-.13*	.13*	-.20***	.22***	-.14**	-.08	.01	.16**	.04	.07	1.00					
12. Pa	.01	.20***	.08	.04	-.01	-.01	.20***	.11*	.27***	.34***	.12*	1.00				
13. Pt	-.20***	.32***	.04	.11*	.10*	-.21***	.43***	.34***	.24***	.41***	.12*	.42***	1.00			
14. Sc	-.22***	.23***	.02	.01	-.01	-.19**	.43***	.15**	.26***	.51***	.13*	.36***	.64***	1.00		
15. Ma	-.08	.25***	.28***	.07	-.21***	-.04	-.01	-.08	-.04	.25***	.02	.18**	.14**	.34***	1.00	
16. Si	-.46***	.38***	-.52***	-.22***	-.10*	-.20***	-.02***	.58***	-.32***	-.06	.11*	.01	.21***	.11*	-.14**	1.00

*Note:* NEO-PI-3 = NEO Personality Inventory-3; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; PPM = Positive Presentation Management research validity scale; N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness; Hs = Hypochondriasis; D = Depression; Hy = Hysteria; Pd = Psychopathic Deviate; Mf = Masculinity-Femininity; Pa = Paranoia; Pt = Psychasthenia; Sc = Schizophrenia; Ma = Hypomania; Si = Social Introversion. \* $p < .05$ , one-tailed. \*\* $p < .01$ , one-tailed. \*\*\* $p < .001$ , one-tailed.

Table 8

*Patterns and Structure Matrix for Principal Axis Factoring with Direct Oblimin Rotation of Two Factor Solution of Validity Scales*

<i>Validity Scales</i>	<i>Pattern coefficients</i>		<i>Structure Coefficients</i>		<i>Initial Communality</i>	<i>Extraction Communality</i>
	<i>Factor 1</i>	<i>Factor 2</i>	<i>Factor 1</i>	<i>Factor 2</i>		
K	.97	-.09	.92	.39	.76	.85
S	.90	.06	.93	.52	.80	.87
So	.75	.05	.77	.43	.58	.60
Sd	-.19	1.003	.31	.91	.77	.85
ODecp	.20	.839	.62	.94	.84	.91
L	.30	.54	.57	.69	.54	.54
PPM	.35	.38	.54	.55	.40	.40

*Note:* Eigenvalue for Factor 1 = 4.24 (percentage of variance = 60.6); eigenvalue for Factor 2 = 1.22 (percentage of variance = 17.4). K = K scale; S = Superlative Self-Presentation Scale; So = Edward's Social Desirability Scale; Sd = Wiggins' Social Desirability Scale; ODecp = Other Deception Scale; L = Lie scale; PPM = Positive Presentation Management scale



Table 9

*Agreement Between the PPM Research Validity Scale and the MMPI-2 L, K, and S Validity Scales (N=303)*

<i>Scale</i>	<i>Validity</i>	<i>MMPI-2 L, K, and S Scales</i>		<i>Overall Agreement</i>	<i>Cohen's Kappa</i>
		<i>Valid (T score ≤ 64)</i>	<i>Invalid (T score ≥ 65)</i>		
PPM	Valid (Raw ≤ 21)	68	26	66.3%	.31
	Invalid (Raw ≥ 22)	76	133		

*Note:* PPM = Positive Presentation Management research validity scale; L = Lie scale; K = K scale; S = Superlative Self-Presentation Scale.

Table 10

*Classification Estimates for Selected Cutoff Scores Based on ROC Analyses on PPM in Detecting Positive Response Distortion*

<i>Cutoff Score</i>	<i>SN</i>	<i>SP</i>
PPM (AUC = .75)		
≥20	.94	.30
≥21	.87	.39
≥22	.84	.47
≥23	.77	.60
≥24	.69	.69
≥25	.59	.78

*Note:* ROC = receiving operating characteristics; AUC = Area under the ROC curve; PPM = Positive Presentation Management research validity scale; SN = sensitivity; SP = specificity

Table 11

*Factor and Clinical Scores of Participants with Invalid (Raw  $\geq 22$ ) Versus Valid (Raw  $\leq 21$ ) Scores on the PPM Scale*

Scale	Invalid PPM <sup>a</sup>		Valid PPM <sup>b</sup>		Statistic (df)	Partial eta squared
	M	SD	M	SD		
NEO-PI-3						
N	38.1	7.9	46.8	7.1	$F(1, 301) = 84.52^*$	.220
E	58.5	7.7	54.7	7.9	$F(1, 301) = 15.56^*$	.049
O	55.6	9.0	54.5	11.3	$F(1, 301) = .89$	.003
A	60.0	7.8	56.6	8.3	$F(1, 301) = 11.85^*$	.038
C	54.7	7.5	49.0	8.8	$F(1, 301) = 33.43^*$	.100
MMPI-2						
Hs	49.0	5.4	48.3	7.2	$F(1, 301) = 1.02$	.003
D	45.0	6.3	48.4	7.4	$F(1, 301) = 16.60^*$	.052
Hy	52.1	5.9	49.5	7.4	$F(1, 301) = 10.70^*$	.034
Pd	52.7	6.5	53.5	7.7	$F(1, 301) = 1.01$	.003
Mf	43.7	10.1	46.0	9.2	$F(1, 301) = 3.59$	.012
Pa	51.1	7.8	49.5	8.5	$F(1, 301) = 2.49$	.008
Pt	50.4	5.1	52.2	7.5	$F(1, 301) = 6.15$	.020
Sc	49.1	5.7	51.6	7.9	$F(1, 301) = 9.78$	.032
Ma	50.3	8.3	50.6	8.2	$F(1, 301) = .12$	.000
Si	39.1	6.2	46.6	9.6	$F(1, 301) = 66.69^*$	.182

*Note:* PPM = Positive Presentation Management research validity scale; NEO-PI-3 = NEO Personality Inventory-3; MMPI-2 = Minnesota Multiphasic Personality Inventory-2; N = Neuroticism; E = Extraversion; O = Openness; A = Agreeableness; C = Conscientiousness; Hs = Hypochondriasis; D = Depression; Hy = Hysteria; Pd = Psychopathic Deviate; Mf = Masculinity-Femininity; Pa = Paranoia; Pt = Psychasthenia; Sc = Schizophrenia; Ma = Hypomania; Si = Social Introversion. After controlling for multiple comparisons, the threshold for statistical significance was  $p < .0014$  on the NEO-PI-3 and  $p < .002$  on the MMPI-2 for variables that did not violate any assumptions and  $p < .001$  for variables that violated assumptions, controlling for Type I error. <sup>a</sup>  $n = 209$ . <sup>b</sup>  $n = 93$ . \* Denotes a statistically significant difference between the two groups.

Table 12

*Means and Standard Deviations for MMPI-2 Validity Scales:  
Published DPG and ARD Studies*

<i>MMPI-2 Validity Scales</i>	<i>Current Sample</i>		<i>Butcher Manual</i>		<i>Butcher Pilots</i>		<i>Detrick et al</i>		<i>King et al ATC</i>		<i>Bagby &amp; Marshall</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
L	50.99	11.09	50.00	9.94	57.41	11.67	61.10	11.50	57.50	11.70	64.60	16.17
K	60.79	8.39	50.01	9.90	65.67	6.41	63.80	7.20	57.20	9.30	58.16	8.47
S	63.18	9.69	--	--	--	--	--	--	--	--	61.54	10.11
So	59.10	6.10	--	--	--	--	--	--	--	--	55.37	10.55
ODecp	56.90	10.40	--	--	--	--	--	--	--	--	67.44	15.42
Sd	52.50	9.20	--	--	--	--	--	--	--	--	67.17	15.79

*Note:* MMPI-2 = Minnesota Multiphasic Personality Inventory -2; DPG = Differential Prevalence Group; ARD = Analog Research Design; L = Lie Scale; K = Correction Scale; S = Superlative Scale; So = Edwards' Social Desirability Scale; ODecp = Other Deception Scale; Sd = Wiggins' Social Desirability Scale. Adapted from Butcher (1994); Butcher Pilots (N=437). Butcher et al. (1989); Butcher Manual (Norms N=1,138). Detrick, Chibnall, & Rosso (2001); Caucasian Males (N=395). King, Schroeder, Manning, Retzlaff, & Williams (2008); (N=1,014). Bagby & Marshall (2004); Fake Good Instruction (N=70).

Table 13

*Means and Standard Deviations for MMPI-2 Clinical Scales:  
Published DPG and ARD Studies*

<i>MMPI-2 Clinical Scales</i>	<i>Current Sample</i>		<i>Butcher Manual</i>		<i>Butcher Pilots</i>		<i>Detrick et al</i>		<i>King et al ATC</i>		<i>Bagby &amp; Marshall</i>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Hs	48.84	6.01	50.01	10.04	48.25	4.49	49.30	6.20	50.30	7.50	44.59	9.59
D	46.05	6.82	49.99	9.94	42.89	4.59	45.80	5.70	47.10	7.60	45.96	9.29
Hy	51.30	6.50	50.16	10.10	52.27	5.53	50.90	5.90	48.90	7.50	50.07	9.37
Pd	52.97	6.94	50.03	9.98	49.26	5.99	51.50	6.80	50.40	7.90	45.89	8.85
Mf	44.39	9.87	50.08	10.05	38.90	6.15	--	--	--	--	--	--
Pa	50.71	8.13	50.15	10.06	47.83	5.73	47.80	6.50	48.30	8.70	49.29	9.19
Pt	51.03	6.04	50.09	9.97	48.43	4.71	48.20	5.40	47.90	8.00	44.24	9.00
Sc	50.03	6.89	50.06	10.02	47.84	4.57	47.50	5.00	49.20	8.40	46.41	10.65
Ma	50.55	8.74	49.98	10.02	45.95	5.52	44.90	6.30	52.60	9.50	51.40	8.64
Si	41.43	8.20	49.94	9.98	39.89	5.76	--	--	--	--	--	--

*Note:* MMPI-2 = Minnesota Multiphasic Personality Inventory – 2; DPG = Differential Prevalence Group; ARD = Analogue Research Design; Hs = Hypochondriasis; D = Depression; Hy = Hysteria; Pd = Psychopathic Deviate; Mf = Masculinity-Femininity; Pa = Paranoia; Pt = Psychasthenia; Sc = Schizophrenia; Ma = Hypomania; Si = Social Introversion. Adapted from Butcher (1994); Butcher Pilots (N=437). Butcher et al. (1989); Butcher Manual (Norms N=1,138). Detrick, Chibnall, & Rosso (2001); Caucasian Males (N=395). King, Schroeder, Manning, Retzlaff, & Williams (2008); (N=1,014). Bagby & Marshall (2004); Fake Good Instruction (N=70).

Table 15

*Factor and PPM T-Score Means and Standard Deviations for NEO- PI-3 and NEO-PI-R, Current Sample, and Published DPG and ARD Studies*

Factors	McCrae & Costa (2010)		Current Sample <sup>a</sup>		Detrick & Chibnall (2013) <sup>a</sup>		Sellbom & Bagby (2008) <sup>a</sup>		Ballenger et al. (2001) <sup>b</sup>	
	M	SD	M	SD	M	SD	M	SD	M	SD
Neuroticism	50.00	10.00	40.90	8.82	37.00	6.80	44.21	10.51	43.70	12.71
Extraversion	50.00	10.00	57.38	7.98	55.30	6.20	47.37	10.26	54.47	7.87
Openness	50.00	10.00	55.18	9.77	47.90	7.90	46.06	9.56	52.93	7.31
Agreeableness	50.00	10.00	58.80	8.29	50.50	7.90	61.10	10.57	58.93	8.76
Conscientiousness	50.00	10.00	53.02	8.39	57.30	8.30	53.29	6.83	55.57	10.18
PPM (raw)	---		23.49	4.11	27.2	3.90	24.50	4.02	20.33	2.72
PPM (T score)	---		56.90	8.84	---		---		---	

*Note:* PPM = Positive Presentation Management Research Validity Scale. McCrae & Costa (2010), NEO PI-3 standardization sample (N = 635). Current Sample (N = 303). Detrick & Chibnall (2013) police candidates (N = 288). Sellbom & Bagby (2008) MMPI-2 L, K, or S  $\geq$  65 (N = 20). Ballenger et al (2001). Fake good (N = 30).<sup>a</sup> DPG = Differential Prevalence design, <sup>b</sup> ARD = Analog Research Design

## Appendix C

## Figure

Figure 1

*NEO-PI-3 and NEO-PI-R Mean T-Score Profiles by Group*

