

2022

The Reliability and Validity of the Open Enneagram of Personality Scales

Kayleigh Kastelein
kayleighkastelein@gmail.com

Follow this and additional works at: <https://digitalcommons.olivet.edu/elaia>



Part of the [Psychology Commons](#)

Recommended Citation

Kastelein, Kayleigh (2022) "The Reliability and Validity of the Open Enneagram of Personality Scales," *ELAIA*: Vol. 4, Article 4.

Available at: <https://digitalcommons.olivet.edu/elaia/vol4/iss1/4>

This Article is brought to you for free and open access by the Honors Program at Digital Commons @ Olivet. It has been accepted for inclusion in ELAIA by an authorized editor of Digital Commons @ Olivet. For more information, please contact digitalcommons@olivet.edu.

The Reliability and Validity of the Open Enneagram of Personality Scales

Cover Page Footnote

The completion of this research thesis would not have been possible without the amazing support of my supervisor, Dr. Kristian Veit. The inspiration for this project came during several classes taught by Dr. Veit, and through discussions with him, the idea for this project was formed. Dr. Veit was an immense help throughout my process, and his continual encouragement and direction was a driving factor in my success. Dr. Beth Schurman and Dr. Dan Sharda, my Honors faculty advisors, have also had a tremendous impact on this project with their guidance and teaching that kept me on track for finishing this project. I am also thankful to the Olivet Honors Program for the funding provided for this project that was crucial to my ability to obtain the high quantity of participants. The Olivet Honors Program has provided me with tremendous experiences, and I am so grateful for the opportunity to be in this phenomenal program.



**The Reliability and Validity
of the Open Enneagram of Personality Scales**

Kayleigh N. Kastelein

ACKNOWLEDGEMENTS

The completion of this research thesis would not have been possible without the amazing support of my supervisor, Dr. Kristian Veit. The inspiration for this project came during several classes taught by Dr. Veit, and through discussions with him, the idea for this project was formed. Dr. Veit was an immense help throughout my process, and his continual encouragement and direction was a driving factor in my success. Dr. Beth Schurman and Dr. Dan Sharda, my Honors faculty advisors, have also had a tremendous impact on this project with their guidance and teaching that kept me on track for finishing this project. I am also thankful to the Olivet Honors Program for the funding provided for this project that was crucial to my ability to obtain the high quantity of participants. The Olivet Honors Program has provided me with tremendous experiences, and I am so grateful for the opportunity to be in this phenomenal program.

ABSTRACT

The purpose of this study is to assess the reliability and validity of the thirty-six-item Open Enneagram of Personality Scales (OEPS). Our general hypothesis was that the OEPS would show adequate reliability evidence but not validity evidence. Participants were acquired through a small denominationally affiliated Midwest university, Amazon Mechanical Turk, and social media. Test-retest reliability was done with 249 participants, whereas internal consistency reliability, factor analysis, and correlations with the Big Five Inventory (BFI; John, Donahue, & Kentle, 1991) were done using 1,039 participants. An average Pearson’s correlation of .68 (range: 0.54 - 0.75) showed inadequate test-retest reliability for the OEPS factors. The average Cronbach’s Alpha was .46 (range: 0.27 - 0.56) for the internal consistency of the OEPS factors. Confirmatory factor analysis found insufficient evidence for the OEPS ($\chi^2 = 1255$, $p < .001$, CFI = 0.56, TLI = 0.50, and RMSEA = 0.08). This study used Pearson’s correlation coefficient to correlate OEPS factors with the BFI factors and found many correlations ($-0.30 > r > 0.30$) that support several of our predictions for convergent validity (See Table 2). There were also some relationships between the OEPS and BFI that were to be expected but were not supported in this study’s analysis, which is most likely due to the lack of strong psychometric support for the OEPS. Overall, this study showed OEPS did not show strong reliability or validity evidence.

Keywords: Enneagram, Open Enneagram of Personality Scales, Big Five Inventory, reliability, validity, confirmatory factor analysis, personality

INTRODUCTION

Personality is “the enduring configuration of characteristics and behavior that comprises an individual’s unique adjustment to life, including major traits, interests, drives, values, self-concepts, abilities, and emotional patterns.” (American Psychological Association, n.d.). Throughout the study of psychology, many people have created different theories about what makes up personality and how it affects a person’s thoughts, feelings, and behaviors (Kline, 2013; Friedman & Schustack, 2016). However, these theories are just speculation unless they can be tested using the scientific method (Kline, 2013). The scientific method requires making hypotheses that can be measured, but personality is a difficult thing to measure because it is not always possible to observe it directly. Psychologists create ways to measure personality through objective and projective measures (Reynolds & Livingston, 2012). Objective personality tests involve respondents using selected-response items to reflect their thoughts, feeling, or behaviors. Projective personality tests involve respondents being presented an ambiguous stimulus and a professional interpreting the respondents’ open-ended response to the stimulus (Reynolds & Livingston, 2012). This study will focus on the Enneagram, a popular yet not thoroughly tested theory of personality, and the psychometric study of one of its objective self-report measures.

The Enneagram theory

Not much is known about the early history of the Enneagram, but many believe that it is a personality theory based on Islamic mysticism that evolved after being adapted by Judeo-Christian and Greek philosophies (Bland, 2010; Matise, 2007; Wagner, 1980). George Gurdjieff introduced this idea to the West at a French conference in 1915 (Bland, 2010; Kam 2019). After spending time studying human nature in the Middle East, Gurdjieff learned of an idea of people having “chief features” or “passions” central to an individual’s personality (Kam, 2019). A Bolivian philosopher, Oscar Ichazo, also learned of nine personality types while studying in Asia and the Middle East. Ichazo connected the personality types with the symbol it is known for today and began teaching classes on the system in South America (Kam, 2019). Claudio Naranjo learned of the system from Ichazo while in Chili and brought it back to the U.S. (Matise, 2007, Wagner, 1980). Naranjo connected these Eastern spiritual practices with Western psychology and used it as a tool for helping people transcend their patterns and habits of behavior (Kam, 2019). Jesuit priests from Loyola University adopted Naranjo’s system and began to use the system in their counseling (Matise, 2007). Don Richard Riso learned of the Enneagram in his time studying to be a Jesuit priest and helped popularize the Enneagram in the 1980s through his research and writings (Matise, 2007; Bland, 2010). The Enneagram has gained much popularity in the last several years, gaining widespread usage in places such as Stanford University School of Business, the U.S. Postal Service, and the CIA (Bland, 2010).

The name Enneagram comes from the Greek words *ennea* (nine) and *gramma* (written) (Matise, 2007). The Enneagram labels individuals as one of nine personality types or “home styles” (Matise, 2007). Though nomenclature for each Enneagram type may vary, the overall descriptions are the same.

TABLE 1: DESCRIPTIONS OF ENNEAGRAM TYPES

These names and descriptions of types were acquired from The Enneagram Institute (2019).

| Enneagram Type | Description |
|----------------------------|--|
| Type 1 (The Reformer) | principled, purposeful, self-controlled, and perfectionistic |
| Type 2 (The Helper) | generous, demonstrative, people-pleasing, and possessive |
| Type 3 (The Achiever) | adaptable, excelling, driven, and image-conscious |
| Type 4 (The Individualist) | expressive, dramatic, self-absorbed, and temperamental |
| Type 5 (The Investigator) | perceptive, innovative, secretive, and isolated |
| Type 6 (The Loyalist) | engaging, responsible, anxious, and suspicious |
| Type 7 (The Enthusiast) | spontaneous, versatile, acquisitive, and scattered |
| Type 8 (The Challenger) | self-confident, decisive, willful, and confrontational |
| Type 9 (The Peacemaker) | receptive, reassuring, complacent, and resigned |

The Enneagram theory teaches that each personality type has a singular unconscious motivation that drives their behaviors (i.e., Type 1's unconscious motivation is perfection, whereas Type 6's is fear) (Sutton, Allinson, & Williams, 2013). According to the theory, there are many relationships between types, and each type can exhibit traits of other types in times of stress or growth. People can also exhibit traits of the types that are next to their central type (e.g., Type 5 may show traits of Type 6 or Type 4) (Sutton, Allinson, & Williams, 2013). Before being able to study all the relationship between types, there needs to be a way to measure the individual types by themselves. This study will focus on the classification of the nine types and not their relationships with other types.

There is still some debate about the validity of the Enneagram theory and the reliability and validity of various assessments used to measure the nine Enneagram types. There have been recent efforts to address these uncertainties by gaining psychometric evidence for the Enneagram, but there remains inconsistent and inadequate research on the reliability and validity of the Enneagram (Bland 2010; Matise, 2007). This study will attempt to address this lack of psychometric support of the Enneagram.

Reliability and validity

In the field of psychology, personality theories are assessed empirically. In research or practice, any assessments used should yield reliable scores, where reliability is defined as the ability of a test to produce consistent and stable results (Reynolds & Livingston, 2012). Ways to assess consistency of scores in a personality assessment include test-retest reliability, alternate forms reliability, internal consistency reliability, and inter-rater reliability. This study will focus mainly on test-retest reliability and internal consistency reliability because the nature of the assessment used in this study does not allow for alternate forms reliability or inter-rater reliability.

Internal consistency reliability measures the consistency of items measuring the same construct (Navarro & Foxcroft, 2019; Reynolds & Livingston, 2012). One way to assess internal consistency reliability would be to split the assessment into two equivalent halves and correlate the responses on both halves to see if they produce similar results. An example of this using the Enneagram would involve splitting the assessment into two equal parts, meaning same number of questions for each type on each half of the test. The two tests would be given to the same participant, and both halves would then be correlated to see if there is a strong relationship between the two equivalent tests. There are many ways that the items could be split in half, but Cronbach's alpha is a statistic that calculates the equivalency of all possible split halves (Navarro & Foxcroft, 2019). A strong assessment should be able to be split in any way and both halves produce similar results. An acceptable Cronbach's alpha would be above .70, which would mean that only 30% of the score of the measure is due to error variance.

Test-retest reliability is testing the assessment's ability to yield the same results over time (Reynolds & Livingston, 2012). Test-retest reliability correlates the responses of a participant from time 1 to time 2 of taking an assessment. In studying the Enneagram, participants will take a particular Enneagram assessment initially, and after a certain

designated amount of time, the participants would take the exact same Enneagram assessment. Their results would be correlated to see if they produce similar scores on each type across time. Though .70 is generally an adequate test-retest reliability coefficient for a personality measure, a coefficient over .80 represents a strong measure of reliability. A reliability coefficient of .70 means that 30% of the difference in responses is due to random error, whereas a coefficient of .80 means that only 20% of the differences in responses is due to random error (Reynolds & Livingston, 2012).

Validity is the ability of an assessment to measure what it is intended to measure (Friedman & Schustack, 2016). This study will collect two types of validity evidence. The first is convergent validity evidence. Convergent validation evidence exists if the construct from the observed assessment is related to a similar construct of another assessment (Friedman & Schustack, 2016). When assessing the Enneagram, participants will take the Enneagram and some other similar personality assessment. If the Enneagram theory was assessing some form of personality, there should be relationships between Enneagram types and other scientifically supported personality traits. A commonly used personality theory for validation and in the study of the Enneagram is the Big Five personality traits (Newgent et al., 2004; Yilmaz et al., 2016). This study will use the Big Five Inventory, created and validated by John, Donahue, and Kentle (1991), to find relationships with Enneagram types. A second method to assess the validity of the is by examining its internal structure using factor analysis. This study will focus on confirmatory factor analysis, which attempts to see how well the data fits the given model (Navarro & Foxcroft, 2019). When studying the Enneagram, factor analysis would look at the relationships between the questions in each type and assess whether the 9-type model is best fit with the data.

REVIEW OF LITERATURE

Wagner Enneagram Personality Style Scale

Very little scientific research was conducted on the Enneagram until Wagner created the Enneagram Personality Inventory (EPI) in 1981 (Wagner, 1980; Matise, 2007). The EPI is a 135-item measure but was adapted to a 200-item measure called the Wagner Enneagram of Personality Style Scale (WEPSS), which was published by Western Psychological Services (WPS) in 1999. WEPSS was normed using a sample of 1,429 individuals ranging from the ages 18 to 83 (Western Psychological Services, 2018).

Brown (2003) and Bernt (2003) critiqued the WEPSS in the *Mental Measurements Yearbook*, which to date is the only Enneagram assessment to be assessed in the *Mental Measurements Yearbooks*. Their first critique of the studies on WEPSS was of the small sample size. Bernt (2003) claims that the population group consists of mostly college educated participants. No other demographics were discussed in the test manual other than age and gender. Despite this critique, the results from the WEPSS study were indicative of strong internal consistency reliability. The range of Cronbach's alpha coefficient values for each of the Enneagram types were between .73 and .88, which all indicate a fairly strong internal consistency reliability. All the test-retest reliability coefficients for each of the Enneagram types were between .75 to .81, indicating strong

stability coefficients. However, more studies may need to be done with larger and more representative samples to support reliability (Bernt, 2003). Both Bernt (2003) and Brown (2003) suggest future studies perform factor analysis and correlation of the WEPSS to Big Five personality scales. Factor analysis for Enneagram types should load into nine factors that match each type's description (Reynolds & Livingston, 2012). Sharp (1994) performed factor analysis on the WEPSS and found a five-factor solution best fit for the assessment, which suggests that the WEPSS does not show strong construct validity. With small sample sizes, weak support from factor analysis, and low quantity of studies, there is not enough support for reliability and validity of the WEPSS for it to be considered a strong assessment of the Enneagram.

Riso-Hudson Enneagram Type Indicator (Version 2.5)

Another scale in the scientific study of the Enneagram is the 144-term Riso-Hudson Enneagram Type Indicator Version 2.5 (RHETI). It claims to be “the most popular Enneagram-based test” and a “scientifically validated test” (The Enneagram Institute, 2019), yet this is not supported in the scientific research.

One study based on forty-four participants correlated the RHETI with the Revised Neo Personality Inventory (NEO PI-R) (Newgent, Gueulette, Newman, & Parr, 2000). This pilot study revealed significant relationships between the Enneagram and the NEO PI-R factors, but further studies are needed with a larger sample size. Another similar study was conducted examining correlations between the RHETI and the NEO PI-R using a convenience sample of 287 people (Newgent, Parr, Newman, & Higgins, 2004). The researchers used Cronbach's coefficient alpha and found that internal consistency reliability coefficients for each type ranged from .56 to .82. Six of the nine types had a reliability coefficient greater than the acceptable standard of .70, so this study did not support adequate reliability for all types. In terms of validity evidence, the study found some moderately strong relationships between NEO PI-R big five personality traits and RHETI Enneagram types.

One limitation to the RHETI is that it uses an ipsative scale that forces participants to choose between two statements instead of responding to a single statement, which may affect psychometric estimates (Newgent, Parr, Newman, & Higgins, 2004). The small amount of research conducted shows that the RHETI is not strong enough to be used as an assessment for the Enneagram. It is thus unclear why the RHETI is described as “a scientifically validated test” when there is not enough research to support that claim (The Enneagram Institute, 2019).

Nine Types Temperament Model

One study created the Nine Types Temperament Model (NTTM), which is an assessment based on the Enneagram. NTTM uses temperament types versus personality types because its authors believe temperament has biological and genetic underpinning that make up personality. Their intent was to create a temperament scale to test biological underpinnings of types. They created the NTTM and assessed its reliability and validity. The NTTM is a ninety-one-item scale using a three-point Likert-type scale. The first study had a sample of 990 students (Yilmaz, Gencer, Aydemir, Yilmaz, Kesebir, Unal,

Orek, & Bilici, 2014). The study looked at validity using confirmatory and exploratory factor analysis. Exploratory factor analysis found that all types factored except for Type 4. Type 7, 3, and 9 all factored partially, whereas Type 8, 5, 2, 6, and 1 all fully factored. Confirmatory factor analysis showed that the whole scale and all types were significant with Type 4 being the least acceptable factor. One possible reason for this could be that Type 4 has the fewest people in statistical assessment. Also, Type 4 personalities are characterized as preferring to be unique and different from others, which could mean their personality skews their responses in this assessment. Type 3 and 7 also have lower CFI scores. This could be explained because Type 4, 3, and 7 are all narcissistic personalities and may not respond to self-report accurately when items refer to negative qualities. This study also found adequate internal consistency for all types except for Type 3 (Yilmaz, Gencer, Aydemir, Yilmaz, Kesebir, Unal, Orek, Bilici, 2014). A second study compared the Five Factor Model of Personality (FFM) and the NTTM and found multiple moderate relationships with Pearson's correlations greater than 0.30 using a cluster sampling of 247 participants (See **Table 2**) (Yilmaz et al., 2016).

TABLE 2: RELATIONSHIP BETWEEN BIG FIVE PERSONALITY TRAITS AND ENNEAGRAM TRAITS

All correlations coefficients are significant at the $p < .01$

| Big Five Personality | RHETI | NTTM | OEPS |
|----------------------|-------|-------|-------|
| Extraversion | | | |
| Type 2 | 0.43 | 0.35 | 0.32 |
| Type 3 | | 0.44 | |
| Type 4 | -0.31 | | |
| Type 5 | -0.39 | -0.67 | |
| Type 6 | | -0.67 | |
| Type 7 | 0.45 | 0.57 | 0.54 |
| Type 8 | | 0.42 | 0.37 |
| Conscientiousness | | | |
| Type 1 | 0.46 | 0.58 | 0.35 |
| Type 2 | | 0.35 | |
| Type 4 | -0.36 | -0.39 | -0.36 |
| Type 7 | -0.30 | -0.58 | |
| Openness | | | |
| Type 2 | 0.30 | | |
| Type 6 | -0.38 | | |
| Type 7 | 0.33 | 0.33 | 0.31 |
| Neuroticism | | | |
| Type 2 | | 0.32 | |
| Type 4 | 0.49 | 0.43 | 0.37 |
| Type 6 | | 0.64 | |
| Agreeableness | | | |
| Type 2 | | 0.34 | 0.37 |
| Type 8 | | -0.33 | |
| Type 9 | 0.46 | 0.51 | |

Open Enneagram of Personality Scales

The Open Enneagram of Personality Scales (OEPS) is a thirty-six-item assessment derived from an initial scale containing seventy-two items that were developed from reading descriptions of types from a variety of sources. This initial survey was given to 7,898 participants who were confident in their self-type after spending several hours studying the Enneagram. Each item was assigned to a type based on which type it is correlated strongest with. The top four questions for each type were used to make up the OEPS. However, it is not yet known if the OEPS is a reliable and valid assessment.

Though there are psychometric studies done on other Enneagram assessments, such as the WEPSS, RHETI, and NTTM, many of these tests are still lacking in their psychometric support. This study will analyze the OEPS, which is a free source located at Open-Source Psychometric Project (“Development of the OSPP Enneagram of Personality Scale,” accessed 2020). Test-retest reliability, internal consistency, correlations with Big Five personality traits, and factor analysis will be performed as in previous studies on the WEPSS, RHETI, and NTTM. These analyses will determine if the OEPS would be a reliable and valid alternative to the other assessments of the Enneagram. Compared to previous studies, a larger sample of participants will be used in this study. Also, OEPS is not ipsative, which allows for more sophisticated analysis.

Hypotheses

Though there are mixed results on reliability evidence in previous assessments, I hypothesize that the OEPS is a reliable measure to assess personality. In this study, I will assess the test-retest reliability of the OEPS and hypothesize that there will be a correlation greater than 0.70 between time 1 and time 2 on OEPS types. I will also assess internal consistency of the OEPS. I hypothesize that there will be sufficient evidence to support internal consistency with a Cronbach’s alpha coefficient of greater than .70.

In contrast, I hypothesize that the OEPS is not a valid measure to assess personality. Previous research on more formal Enneagram assessments has failed to find sufficient evidence to support the validity of the Enneagram, so I predict that the OEPS will also fall short of validity standards. I will perform a correlational study to find relationships between OEPS Enneagram types and Big Five personality traits. Based on previous studies, relationships may exist between OEPS and Big Five personality scales. Based on results from previous research done on the RHETI and the NTTM (Newgent et al., 2004; Yilmaz et al., 2016), conscientiousness should positively correlate with Type 1 and negatively correlate with Type 7. Extraversion should positively correlate with Type 2, Type 3, Type 7, and Type 8 and negatively correlate with Type 5 and Type 6. Neuroticism should positively correlate with Type 4 and Type 6. Agreeableness will positively correlate with Type 9 and negatively correlate with Type 5 and Type 8. Openness to Experience will positively correlate with Type 7. To assess the internal structure of the OEPS, this study will perform confirmatory factor analysis on OEPS responses. Because the Enneagram theory contains nine types, the OEPS should yield a nine-factor model through factor analysis if it is a valid assessment. However, since the OEPS was created in a less strategic and scientific method as earlier tests, I hypothesize that the OEPS will not support a 9-factor model. Overall, I predict that the OEPS will yield reliable, but not valid results.

METHOD

Participants

I sought to enroll thirty participants per Enneagram type, or 270 participants total, for this study. Accounting for 50% attrition rate between initial assessment and second assessment, this study required at least 540 participants for the initial survey. After IRB approval and informed consent, participants were acquired through Amazon Mechanical Turk, an email sent to students at a small, denominationally affiliated Christian university in the Midwestern United States, and through social media. A total of 1,286 participants responded to the initial survey. The survey consisted of two discrimination questions that tested the participant attentiveness to the survey. Each question told the participant to select a specific response. Participants were removed if they incorrectly responded to at least one of the two discrimination question. Of the 1,286 participants, 247 participants were excluded from this study because they failed at least one of the two discrimination questions or were under the age of eighteen. A total of 1039 participants were used in the time 1 analysis. A second survey was sent to participants six months after the initial survey. The sample from the second participant group was acquired through email from those participants who agreed in initial survey to be sampled again. There were 259 participants who responded to the second survey, but seventeen of those were excluded because they failed at least one of the two discrimination questions. A total of 242 participants were used in analysis for time 2.

Demographics

Several demographics were collected from participants such as age, race/ethnicity, gender, knowledge of the Enneagram, and Enneagram number. Gender was assessed by choosing male, female, or prefer not to answer. There were 1,039 participants used in the initial analysis (599 women, 441 men, $M_{age} = 30.1$, $SD = 13.1$ years). Based on suggestions by Hughes, Camden, and Yangchen (2016), participants chose one of seven races or ethnicities, other, or prefer not to answer.

TABLE 3: FREQUENCIES OF RACE FOR TIME 1

| Levels | Counts | % of Total | Cumulative % |
|---|--------|------------|--------------|
| White | 726 | 69.9 % | 69.9 % |
| Black/African American | 75 | 7.2 % | 77.1 % |
| Hispanic, Latino, or Spanish origin | 59 | 5.7 % | 82.8 % |
| Asian | 153 | 14.7 % | 97.5 % |
| American Indian or Alaska Native | 5 | 0.5 % | 98.0 % |
| Middle Eastern or North African | 6 | 0.6 % | 98.6 % |
| Mixed | 8 | 0.8 % | 99.3 % |
| I prefer not to answer | 5 | 0.5 % | 99.8 % |
| Other | 1 | 0.1 % | 99.9 % |
| Native Hawaiian or other Pacific Islander | 1 | 0.1 % | 100.0 % |

The second participant group consisted of 242 participants (174 women, 68 males, $M_{age} = 26.2$, $SD = 11.9$ years).

TABLE 4: FREQUENCIES OF RACE FOR TIME 2

| Levels | Counts | % of Total | Cumulative % |
|-------------------------------------|--------|------------|--------------|
| White | 200 | 82.6 % | 82.6 % |
| Black/African American | 9 | 3.7 % | 86.4 % |
| Hispanic, Latino, or Spanish origin | 12 | 6.2 % | 91.3 % |
| Asian | 16 | 6.6 % | 97.9 % |
| Mixed | 1 | 0.4 % | 98.3 % |
| I prefer not to answer | 1 | 0.4 % | 98.8 % |

Participant responses from time 1 and time 2 were connected using their emails, but eighteen participants' emails did not match any emails from original responses. Their data was kept and used in analysis of the second responses but not used in analyzing the relationship between responses. Two hundred and twenty-four participants answered both the first and second survey and were used when analyzing test-retest reliability.

Materials

Open Enneagram of Personality Scale

The nine Enneagram types were assessed using the Open Enneagram of Personality Scale (OEPS) taken from openpsychometrics.org. The OEPS consists of thirty-six statements (four items per Enneagram type) and uses a 5-point Likert scale (1 = disagree to 5 = agree). Total scores for each type were calculated by taking a sum of all the responses for the type. There is no published research to date on the OEPS, so this study will attempt to assess the reliability and validity of this assessment.

Big Five Inventory

The Big Five Inventory (BFI) was used to obtain the convergent validity of the OEPS. It was created by John, Donahue, and Kentle (1991) as a shorter alternative to a Big Five personality assessment. It is a 44-item inventory that has been tested and shown to be just as strong as other larger Big Five personality assessments (John, Donahue, & Kentle, 1991). Total scores were calculated by taking a sum of the responses for that factor, accounting for the reverse coded items. The BFI has an internal consistency of .83. When correlated to other major Big Five personality tests, the BFI correlated strongly to both the NEO-FFI (mean $r = .73$) and the Trait Descriptive Adjectives (TDA; mean $r = .73$), which indicates strong convergent validity evidence. The BFI also showed strong evidence from confirmatory factor analysis with all items correlating over .90 to the factors (John & Srivastava, 1999).

Procedure

After IRB approval, a survey link was sent via email to all undergraduate students at a small denominationally affiliated Midwest university. Participants were given two weeks to complete the survey. The same survey was also posted on Amazon Mechanical

Turk and social media. After giving their consent, participants completed both the OEPS and the BFI along with some demographic questions. Amazon Mechanical Turk participants were given compensation of \$0.30 for taking the survey. All other participants were entered into a drawing for a \$25 Visa gift card. At the end of the survey, participants were asked if they would be willing to participate in a follow-up survey. Those who responded yes were sent the same survey to the emails they provided six months after initial intake. The second survey also remained up for two weeks to give participants time to respond. Responses from both times were connected using participants' email addresses. Emails were discarded after the analysis of their data to maintain privacy.

RESULTS

Data screening

Again, a number of participants were excluded from the data set because they failed at least one of the two discrimination questions at either time 1 or time 2. Generalizability and normality of both the OEPS and the BFI data were assessed using skewness and kurtosis statistics. The OEPS showed moderate to low skewness (between -1 and 0) and low kurtosis (between -0.323 and 0.308) for each type. The BFI showed low skewness (between -0.44 And 0.07) and moderate to low kurtosis (between -0.54 to .42) for each trait. There was no missing data in the data because all questions were required to be answered.

Analysis

Test-retest reliability

Test-retest reliability was assessed using Pearson's correlation coefficient. There was a significant positive relationship between OEPS responses from March and OEPS responses from September, $r(222) = .68$, $p < .001$. (For individual type results see table 5.) Moreover, there is a significant positive relationship between BFI responses from March and BFI responses from September, $r(222) = .85$, $p < .001$. With an acceptable correlation of .70, this study demonstrates that the OEPS does not show test-retest reliability, but the BFI does show strong test-retest reliability.

TABLE 5: TEST-RETEST RELIABILITY FOR OEPS TYPES AND BFI TYPES

All p-values significant at the .001 level.

| Trait | Pearson correlation coefficient | Trait | Pearson correlation coefficient |
|--------|---------------------------------|-------------------|---------------------------------|
| Type 1 | 0.70 | OEPS total | 0.68 |
| Type 2 | 0.75 | Conscientiousness | 0.91 |
| Type 3 | 0.61 | Extraversion | 0.86 |
| Type 4 | 0.74 | Openness | 0.85 |
| Type 5 | 0.70 | Agreeableness | 0.82 |
| Type 6 | 0.54 | Neuroticism | 0.82 |
| Type 7 | 0.75 | BFI Total | 0.85 |
| Type 8 | 0.64 | | |
| Type 9 | 0.73 | | |

Internal consistency reliability

Internal consistency of the OEPS scales and the BFI scales was analyzed using Cronbach's alpha. The internal consistency from the time 1 survey was analyzed using the responses from 1,039 participants. The results from the first survey suggests inadequate internal consistency for the OEPS. The average Cronbach's alpha of all the OEPS type scales was .46 with a range of .27 (Type 6) and .56 (Type 9). The internal consistency for BFI was supported with an average Cronbach's alpha of .80 and a range of .76 (Openness) to .90 (Extraversion). The internal consistency from the time 2 survey was analyzed from the responses of 242 participants. The average internal consistency for time 2 for the OEPS was .40 with a range of .17 (Type 3) to .67 (Type 9), which shows inadequate internal consistency. The internal consistency for the BFI for time 2 had a Cronbach's alpha coefficient average of .83 with a range of .77 (Openness) and .89 (Extraversion). With a .70 Cronbach's Alpha to show adequate internal consistency, this study shows that the OEPS does not show internal consistency, whereas the BFI does show internal consistency.

TABLE 6: INTERNAL CONSISTENCY OF OEPS AND BFI
FOR TIME 1 AND TIME 2

| Scale | Cronbach's alpha time 1 | Cronbach's alpha time 2 |
|-------------------|-------------------------|-------------------------|
| Type 1 | 0.51 | 0.49 |
| Type 2 | 0.50 | 0.53 |
| Type 3 | 0.46 | 0.17 |
| Type 4 | 0.47 | 0.37 |
| Type 5 | 0.48 | 0.50 |
| Type 6 | 0.27 | 0.19 |
| Type 7 | 0.55 | 0.51 |
| Type 8 | 0.36 | 0.19 |
| Type 9 | 0.56 | 0.67 |
| OEPS total | 0.46 | 0.40 |
| Extraversion | 0.90 | 0.89 |
| Agreeableness | 0.81 | 0.81 |
| Conscientiousness | 0.78 | 0.81 |
| Neuroticism | 0.84 | 0.84 |
| Openness | 0.76 | 0.77 |
| BFI Total | 0.80 | 0.83 |

Factor analysis

Confirmatory factor analysis (CFA) was performed on all thirty-six OEPS items. Items were put into factors based on the question that was intended for that factor. The analysis was performed on time 1 responses with 1,039 participants. The results showed that a 9-factor model was not a strong fit for the model (CFI = .56, TLI = .50, and RMSEA = .08). Though we do have a large chi-square value (1,255, $p < .001$), this is probably due to large sample size.

CFA was performed on all forty-four items of the BFI with items assigned to factors based on criteria from the BFI. Analysis was performed on responses from time 1 with 1,039 participants. The results for a 5-factor model showed inadequate evidence to support the model (CFI = .60, TLI = .58, and RMSEA = .09). Though we do have a large chi-square value (8,248, $p < .001$), again, this is probably due to a large sample size.

Convergent validity

Correlations were performed between BFI traits and OEPS types using Pearson's correlation coefficient. All correlation between Enneagram types and Big Five personality types found in this study are relationships that were found in previous research. This study considered any correlation of .30 or greater to be a notable correlation. Based on results from previous research, conscientiousness should positively correlate with Type 1 and negatively correlate with Type 7. This study partially supported this hypothesis by finding a correlation with conscientiousness and Type 1 but no correlation with Type 7. Extraversion should positively correlate with Type 2, Type 3, Type 7, and Type 8 and negatively correlate with Type 5 and Type 6. This study partially supported this hypothesis and found that Extraversion only correlated with Type 2, Type 7, and Type 8, but no correlation with Type 3, Type 5, or Type 6. Neuroticism should positively correlate with Type 4 and Type 6, yet this study partially supported this theory by only finding a correlation with Type 4. Agreeableness should positively correlate with Type 9 and negatively correlate with Type 5 and Type 8, yet this study did not support this hypothesis and only found Type 2 to correlate with Agreeableness. Openness to Experience should positively correlate with Type 7, and this study did support that hypothesis. (See Table 2 for all results.)

DISCUSSION

Overall, this study found inadequate evidence to support the use of the OEPS as an assessment for the Enneagram personality types. The test-retest reliability for the OEPS types had an average correlation coefficient of .68 between time 1 and time 2, which is less than the desired .70 correlation. Only five of the nine OEPS types had a test-retest reliability coefficient that was greater than .70. In comparison, the BFI types had an average stability coefficient of .85 with all test-retest reliability correlations above .82. This demonstrates that the OEPS does not have adequate test-retest reliability to assess personality.

Internal consistency reliability was also inadequate for OEPS's assessment of personality. The average Cronbach's alpha was .46, with no type reaching above a .56. These Cronbach's alphas are well below the accepted minimum of .70 for internal consistency to be supported. In comparison, the BFI had an average Cronbach's alpha of .80 with a range of .76 to .90. All Big Five types were higher than the desired .70 internal consistency coefficient. Against what was hypothesized, the OEPS does not show sufficient reliability results. Because the BFI still shows strong reliability evidence, it can be concluded that the OEPS's lack of reliability evidence is due to poor quality assessment rather than poor participant effort.

Factor analytic evidence was insufficient for both the OEPS and the BFI. Both the OEPS and BFI had high chi-square values, but this is most likely due to large sample size. Looking at the fit indices, the hypothesis supported that the OEPS did not show adequate fit indices for a nine-factor model ($CFI = .56$, $TLI = .50$, and $RMSEA = .08$). Though the BFI showed higher fit indices ($CFI = .60$, $TLI = .58$, and $RMSEA = .09$), it did not reach the desired level of fit for a five-factor model. It is unclear why the BFI did not show adequate construct validity in this study, but it could potentially be due to underlying correlations between Big Five traits. The major take-away from this evidence is that the BFI still shows stronger evidence than the OEPS in terms of validity.

Lastly, convergent validity of the OEPS was analyzed through correlating the BFI and the OEPS. All the relationships found in this study were supported in previous research supporting the strength of our experimental approach. (See **Table 2**.) There were several relationships that were predicted in the hypotheses but were not found in the data. Part of this could be due to the poor psychometric properties of the OEPS. There were some results found that were not hypothesized but were still supported in previous research. Even though these were found in previous research, it was not hypothesized that these results existed because the relationship were moderate. For example, this study found Type 4 negatively correlated with conscientiousness and Type 2 positively correlated with agreeableness. Previous research showed similar relationships, but the relationships were moderate and not thought to be strong enough to replicate in this study. Despite all of this, it was interesting to see that the OEPS still supported many of the hypotheses.

In order to have a greater understanding of the underlying factors of the OEPS, this study also performed an exploratory factor analysis (EFA). The Kaiser-Meyer-Olkin (KMO) analysis measures were used to find the sampling adequacy for the analysis to see if the sample distribution was adequate for using factor analysis. The overall KMO was .84, which is “meritorious” according to Hucheson & Sofroniou, (1999) (as cited in Field, 2013), and all KMO values for individual items were greater than .68, which is above the acceptable limit of .5 (Fields, 2013). The chi-square test was calculated as 8,112 ($p < .001$) in the Bartlett’s test, which means that the factors are unrelated (Fields, 2013).

The EFA was conducted on the time 1 data. A principal axis factor analysis was conducted on the thirty-six items with oblique promax rotation. An initial analysis was run to obtain eigenvalues for each factor in the data. Three factors had eigenvalues over Kaiser’s criterion of one and in combination explained 24.9% of the variance.

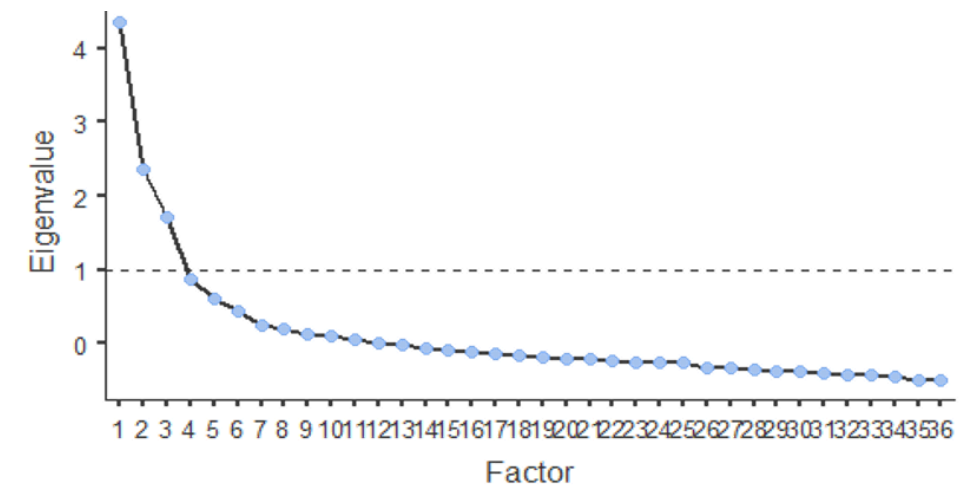


Figure 1: Scree plot for exploratory factor analysis This figure demonstrates the scree plot for the exploratory factor analysis of the OEPS. Dots represent the eigenvalue for each given factor. All eigenvalues over 1 (represented by the dotted line) are considered as a significant factor. In this plot, it shows three factors above the eigenvalue of 1.

The scree plot showed an inflection that would justify retaining three factors. Table 7 shows the factor loadings after rotation. The items that cluster on the same factor suggests that factor 1 represents assertiveness, factor 2 represents people pleasing, and factor 3 represents passiveness, which show similarities to Karen Horney’s neurotic needs of Moving Toward, Moving Away, and Moving Against. This finding may support the idea that there is a relationship between the Enneagram and Horney’s neurotic needs as some researchers have explored (Wagner, 2001; Nettmann & van Deventer, 2013).

TABLE 7: FACTOR LOADING OF OEPS ITEMS

'Principal axis factoring' extraction method was used in combination with a 'promax' rotation.

| | Factor | | | Uniqueness |
|------------------|--------|-------|-------|------------|
| | 1 | 2 | 3 | |
| OEPS 1 (time 1) | | | | 0.833 |
| OEPS 2 (time 1) | | 0.362 | | 0.821 |
| OEPS 3 (time 1) | | | | 0.816 |
| OEPS 4 (time 1) | | | 0.305 | 0.803 |
| OEPS 5 (time 1) | | | 0.437 | 0.718 |
| OEPS 6 (time 1) | | | 0.377 | 0.781 |
| OEPS 7 (time 1) | 0.581 | | | 0.616 |
| OEPS 8 (time 1) | 0.551 | | | 0.636 |
| OEPS 9 (time 1) | | | 0.544 | 0.731 |
| OEPS 10 (time 1) | | 0.395 | | 0.848 |
| OEPS 11 (time 1) | | | 0.425 | 0.784 |
| OEPS 12 (time 1) | 0.632 | | | 0.598 |

| | Factor | | | |
|------------------|--------|-------|-------|------------|
| | 1 | 2 | 3 | Uniqueness |
| OEPS 13 (time 1) | 0.510 | | | 0.598 |
| OEPS 14 (time 1) | | | | 0.900 |
| OEPS 15 (time 1) | | 0.381 | | 0.813 |
| OEPS 16 (time 1) | 0.511 | | | 0.654 |
| OEPS 17 (time 1) | 0.658 | | | 0.602 |
| OEPS 18 (time 1) | | | 0.732 | 0.532 |
| OEPS 19 (time 1) | | | | 0.804 |
| OEPS 20 (time 1) | | 0.641 | | 0.606 |
| OEPS 21 (time 1) | | 0.372 | | 0.806 |
| OEPS 22 (time 1) | | | | 0.937 |
| OEPS 23 (time 1) | | | | 0.803 |
| OEPS 24 (time 1) | | | 0.550 | 0.680 |
| OEPS 25 (time 1) | 0.462 | | | 0.793 |
| OEPS 26 (time 1) | | 0.341 | | 0.861 |
| OEPS 27 (time 1) | | 0.478 | | 0.763 |
| OEPS 28 (time 1) | | 0.359 | | 0.871 |
| OEPS 29 (time 1) | | 0.563 | | 0.709 |
| OEPS 30 (time 1) | | | | 0.839 |
| OEPS 31 (time 1) | 0.484 | | | 0.751 |
| OEPS 32 (time 1) | | 0.322 | | 0.838 |
| OEPS 33 (time 1) | | 0.619 | | 0.647 |
| OEPS 34 (time 1) | | | | 0.836 |
| OEPS 35 (time 1) | | 0.462 | | 0.798 |
| OEPS 36 (time 1) | -0.377 | | 0.654 | 0.604 |

Limitations

The first limitation was that this study was based on self-report data. That means that responses are dependent on the ability and willingness of participants to respond honestly and accurately. Another limitation was the lack of representativeness of the sample. Most participants were younger and Caucasian. Because the sample is not representative of the population, results cannot be generalized to the whole population. A limitation for the test-retest reliability and convergent validity results is that each of the OEPS scales had low internal consistency. Test-retest reliability and convergent validity are reliant on having an internally consistent assessment. Because of this, we can assume that some of the lack of support in test-retest reliability and the convergent validity is due to the OEPS's lack of internal consistency. Another limitation is that this study had a high attrition rate. The study initially had 1,039 participants but only 242 responded to the follow-up survey. This is a 77% attrition rate. This high attrition rate affects the representativeness of the sample because the attrition is not controlled. A certain personality trait may be less likely to respond to the second survey, which would skew time 2 results.

Future research

I believe future research should be done on many Enneagram assessments before their use in any decision-making processes, whether that be in counseling, job hiring, or even decisions on one's own personality. Assessments and theories for the Enneagram need to be supported scientifically before being used for any of these purposes. This could be done by improving the current OEPS to see if the assessment can present better psychometric properties. More research also needs to be done on the major Enneagram test such as the RHETI (version 2.5) and the IEQ9. Both of these tests claim adequate reliability and validity evidence, but currently there simply is not any peer reviewed and published evidence to support these popular tests. Lastly, if additional support can be found for Enneagram assessments, subsequent research should be done on the connection between Enneagram types and Karen Horney's theory of neurotic needs.

REFERENCES

- American Psychological Association. (n.d.). *Personality*. In *APA Dictionary of Psychology*. Retrieved February 4, 2021, from <https://dictionary.apa.org/personality>.
- Bernt, F. M. (2003). [Test review of Wagner Enneagram Personality Style Scale]. B. S. Plake, J. C. Impara, & R. A. Spies (Eds.), *The Fifteenth Mental Measurements Yearbook*. Buros Center for Testing.
- Bland, A. M. (2010). The Enneagram: A review of the empirical and transformational literature. *The Journal of Humanistic Counseling, Education and Development*, 49(1), 16-31.
- Brown, J. B. (2003). [Test review of Wagner Enneagram Personality Style Scale]. B. S. Plake, J. C. Impara, & R. A. Spies (Eds.), *The Fifteenth Mental Measurements Yearbook*. Buros Center for Testing.
- Confirmation bias. (2009). In A. S. Reber, R. Allen, & E. S. Reber, *The Penguin Dictionary of Psychology* (4th ed.). Penguin.
- Development of the OSPP Enneagram of Personality Scales*. <https://openpsychometrics.org/tests/OEPS/development/>.
- Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics* (Fourth ed.). London: SAGE.

Friedman, H. S., & Schustack, M. W. (1999). *Personality: Classic theories and modern research* (p. 576). Allyn and Bacon.

Hughes, J. L., Camden, A. A., & Yangchen, T. (2016). Rethinking and updating demographic questions: Guidance to improve descriptions of research samples. *Psi Chi Journal of Psychological Research*, 21(3), 138-151. doi:10.24839/b21.3.138

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). Big five inventory. *Journal of Personality and Social Psychology*.

John, O. P., & Srivastava, S. (1999). *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives* (Vol. 2, pp. 102-138). Berkeley: University of California.

Kam, C. (2019). Enneagram. In D. A. Leeming (Ed.), *Encyclopedia of Psychology and Religion*. (Vol. 3, pp. 1-5). Springer International Publishing.

Kline, P. (2013). *Personality: The psychometric view*. Routledge.

Matise, M. (2007). The enneagram: An innovative approach. *Journal of Professional Counseling: Practice, Theory & Research*, 35(1), 38-58.

Navarro, D. J., & Foxcroft, D.R. (2019). *Learning Statistics with Jamovi: A Tutorial for Psychology Students and Other Beginners*. (Version 0.70). DOI: 10.24384/hgc3-7p15

Nettmann, R. W., & van Deventer, V. (2013). The relationship between Enneagram type and Karen Horney's interpersonal trends measured as compliance, aggression and detachment. *The Enneagram Journal*, 6(1), 41.

Newgent, R., Gueulette, C., Newman, I., & Parr, P. (2000). An investigation of the Riso-Hudson Enneagram Type Indicator constructs of personality as a unique estimate of personality when considering the Revised NEO Personality Inventory and the five-factor model of personality. *Manuscript submitted for publication*.

Newgent, R. A., Parr, P. E., & Newman, I. (2002). *The enneagram: Trends in validation*. Retrieved from ERIC database. (ED468827)

Newgent, R. A., Parr, P. E., Newman, I., & Higgins, K. K. (2004). The Riso-Hudson Enneagram Type Indicator: Estimates of reliability and validity. *Measurement & Evaluation in Counseling & Development*, 36(4), 226-237. https://doi.org/10.1080/07481756.2004.11909744

Reynolds, C. R., & Livingston, R. B. (2012). *Mastering Modern Psychological Testing: Theory and Methods*. Pearson Education.

Sharp, P. M. (1994). A factor analytic study of three Enneagram personality inventories and the Vocational Preferences Inventory [Disertation, Texas Tech University]. https://ttu-ir.tdl.org/bitstream/handle/2346/61024/3129500846149.pdf?sequence=1

Sutton, A., Allinson, C., & Williams, H. (2013) Personality type and work-related outcomes: An exploratory application of the Enneagram model. *European Management Journal*, 31(3), 234-249. doi: 10.1016/j.emj.2012.12.004

The Enneagram Institute. (2019). *The Riso-Hudson Enneagram Type Indicator* (RHETI® version 2.5). https://www.enneagraminstitute.com/rheti

Wagner Enneagram Personality Style Scale. (2020). *About the WEPSS*. https://www.wepss.com/aboutWEPSS.asp

Wagner, J. (2001). Karen Horney meets the Enneagram. *Enneagram Monthly*, April, 1-13.

Wagner, J.P.(1980).A descriptive, reliability, and validity study of the Enneagram personality typology [Dissertation, Loyola University]. https://ecommons.luc.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com/&httpsredir=1&article=3108&context=luc_diss

Wagner, J. P., & Walker, R. E. (1983). Reliability and validity study of a Sufi personality typology: The Enneagram. *Journal of Clinical Psychology*, 39(5), 712-717. https://doi.org/10.1002/1097-4679(198309)39:5<712::AID-JCLP2270390511>3.0.CO;2-3

Western Psychological Services. (2018). (WEPSS™) *Wagner Enneagram Personality Style Scales™*. Accessed 14 Sept. 2020. https://www.wpspublish.com/wepss-wagner-enneagram-personality-style-scales.

Yilmaz, E. D., Gençer, A. G., Ünal, Ö., & Aydemir, Ö. (2014). From Enneagram to Nine Types Temperament Model: A proposal. *Egitim ve Bilim*, 39(173).

Yılmaz, E. D., Ünal, Ö., Palancı, M., Gençer, A. G., Örek, A., Tatar, A., ... & Aydemir, Ö. (2016). The relation between the Nine Types Temperament Model and the Five Factor Personality model in a Turkish sample group. *Journal of Advances in Medicine and Medical Research*, 1-11.