The NEO Personality Inventory-3:

A Step in the Right Direction

Pepperdine University
Among structured personality tests, one that has received significant research over the past decade has been the NEO Personality Inventory-Three. The NEO Personality Inventory-Three (named for the traits of “neuroticism”, “extroversion”, and “openness”, which it measures) was authored by Paul Costa, Jr. and Robert McCrae, and published in 2005 by Psychological Assessment Resources (Irving & Weiner, 2011). The original NEO-PI-3 is itself an updated version of the NEO Personality Inventory-Revised, which was published in 1992 (Irving & Weiner, 2011). Thirty-seven items on the NEO-PI-R that were deemed difficult for adolescents to understand were replaced, resulting in the NEO-PI-3 (Irving & Weiner, 2011). Comprehensive exams kits (for either adult or adolescent populations) are available for $340, which includes the test manual, ten self-report form booklets, ten observer form booklets, 25 answer sheets, and 25 profile sheets for each test form (NEO Personality Inventory-3). Scoring can be done by hand or by mailing in scannable answer sheets (10 for $25 or 25 for $300), which are then returned with customized score reports (NEO-PI-3 and NEO-PI-R Professional Report Service). Clients can also take a computerized version of the test by using Psychological Assessment Resource’s NEO Software System (NEO Software System).

The NEO-PI-3 is designed to give detailed insight into adult and adolescent personality traits (NEO Personality Inventory-3). It “attempts to provide a multipurpose inventory for predicting interests, health and illness behavior, psychological well-being, and characteristic coping styles” (Kaplan & Saccuzzo, 2013), taking a multidimensional approach to personality by defining it in terms of five traits: neuroticism, extraversion, openness, agreeableness, and conscientiousness (Irving & Weiner, 2011). Theoretically, the test is based on the five-factor model of personality, an approach that uses factor analysis to determine the minimum number of coherent traits that can explain personality variance (Kaplan & Saccuzzo, 2013). The test authors
chose test items that loaded onto the test’s five factors, balancing the amount of positively and negatively worded items (Benson, 2014). The test is designed for patients as young as 12 years of age (NEO Personality Inventory-3), with an overall grade reading level of 5.3 (McCrae, Costa, & Martin, 2005a). Structurally, it consists of 240 items, each of which applies to one of 30 personality facets (with each of the test’s five dimensions consisting of six of these facets) (Benson, 2014). Items are presented in Likert format, with a given statement and five possible responses ranging from “strongly disagree” to “strongly agree” (Benson). The test contains no validity or social desirability subscales (Kluck, 2014). The NEO-PI-3 can be administered to groups or individuals, using either a paper-and-pencil or a computerized version (Benson, 2014). Both self-report and observer forms are available, and test administration usually lasts from 35 to 45 minutes (NEO Personality Inventory-3). Test administrators are allowed to read the test items to the patient in cases of limited reading ability (Irving & Weiner, 2011). According to the test’s publisher (NEO Personality Inventory-3), administrators must possess either a license to work in a health care occupation or a bachelor’s degree in psychology (or a related field) that included coursework in assessment interpretation. For tests taken on a computer, interpretive score reports are automatically generated (NEO Software System). Yet, the answer sheets provided by the exam kit are hand-scorable (and are also available for individual purchase) (NEO Personality Inventory-3). Scores for each item range from 0 to 4, depending on the item’s direction (“strongly agree” is a four point response on some questions and a zero point response on others). Since each item corresponds to only one facet, a facet’s scores (ranging from zero to 32) is simply the sum of the scores for the eight items item pertaining to it. Domain scores range from 0 to 192 (though test norms are truncated at 25 and 172) and consist of the sum of the six facet scores pertaining to that domain (Irving & Weiner, 2011). Profile plots (included in the
exam kit) allow administrators to plot clients’ facet scores for each domain side-by-side, connecting them with a line to provide a graphical representation for how the client responded to each domains (Irving & Weiner, 2011). As mentioned above, scannable answer sheets are available for additional purchase, and can be mailed to the publisher to obtain customized score reports (NEO-PI-3 and NEO-PI-R Professional Report Service).

The NEO-PI-3 is normed for both the self-report form (Form S) and the observer form (Form R) (Benson, 2014). Form S is normed on a sample of 635 adults (ages 21 to 91) and 500 adolescents (ages 14 to 20), while Form R is normed on a sample of 649 adults and 465 adolescents (Benson, 2014). The adult sample was 92.6% White, 1.6% Asian/Pacific Islander, 1.3% Hispanic, and 1.1% Black (McCrae, Martin, & Costa, 2005b). The adolescent sample was also predominantly White (84.6%) (McCrae et al., 2005b). According to McCrae, Martin, and Costa (2005b), the adult standardization sample mostly consisted of acquaintances of the research assistants, who contacted them and explained to them the nature of the study. The NEO-PI-3 was mailed to willing participants, who were given as much time as needed to complete and return the assessment. Most participants were recruited in pairs, who completed the self-report form for themselves and the observer form for the person they were paired with. Unpaired participants completed the observer form for someone they knew well. Similar procedures were used in obtaining the adolescent standardization sample (McCrae et al., 2005b). The demographic representativeness of the norm groups is suspect. In addition to being predominantly White, 63% of the adult sample resided in Pennsylvania, and 75.6% had more than a high school education (McCrae et al., 2005b). The adult sample was also fairly high-earning, with almost twenty percent earning over $100,000 a year (McCrae et al., 2005). McCrae et al. (2005b) also admits that the adolescent sample over-selected high academic achievers.
Overall, the demographic, geographic, and socioeconomic diversity of the norming samples is a limitation of the NEO-PI-3 (Benson, 2014; Kluck, 2014). The NEO-PI-3 does provide adequate subnorms, giving separate normative information for males and females, as well as supplementary norms for younger adults (ages 21 to 30) and older adults (ages 30 and up) (Benson, 2014). Psychometrically, the NEO-PI-3 demonstrates high internal consistency. Internal consistency coefficients for each of the five domains range from .87 to .93 for adolescents and .88 to .95 for adults (Kluck, 2014). However, there are no published values for test-retest reliability (Kluck, 2014). Retest reliability values for NEO-PI-R can be used as estimates, given that “psychometrically, the NEO-PI-3 is similar enough to the NEO-PI-R that previous research findings and clinical familiarity and cautions should generalize to the new version” (Kluck, 2014). A study by Terracciano, Costa, and McCrae used path analytic formulas to estimate domain retest reliabilities, and found coefficients ranging from .83 to .92 (McCrae, Kurtz, Yamagata, & Terracciano, 2011). Kurtz and Parrish administered the NEO-PI-R twice to the same sample, with a retest interval of one to two weeks, and found domain retest coefficients ranging from .91 to .93 (McCrae et al., 2011). Yet there are no officially published short-term retest reliability studies for the NEO-PI-R in American non-clinical samples (McCrae et al., 2011); data presented here should be regarded as reliability estimates only. While internal consistency of the test seems psychometrically sound, a lack of retest reliability information for the NEO-PI-3 is a severe limitation, especially given that one of the key features of personality tests is to measure traits that are (presumably) stable over time. If the test cannot be shown to provide consistent results in multiple administrations in a well-defined study, then variations among individual test scores (and thus, variations in the personality traits it aims to measure) are drawn into question. In terms of validity, the domains on the NEO-PI-3 correlate highly (.98 to
.99) with the domains of its predecessor, the NEO-PI-R (Kluck, 2014). All 30 facets loaded onto their assigned domains when results from the adult standardization sample were factor analyzed, supporting the test’s construct validity (Kluck, 2014). Also supporting its construct validity are replications of the test’s five-factor personality structure in independent samples. A study by Costa, McCrae, and Martin (2008) found that factor analysis in test results for a sample of 12 and 13 year olds supported the test’s underlying five domains. Similarly, another study analyzed test results of adolescents (ages 12 to 17) from 24 cultures, and found that factor analysis supported the five domains (De Fruyt, De Bolle, McCrae, Terracciano, & Costa, 2009). McCrae et al. (2011) analyzed three very specific aspects of the test’s validity. They correlated the self-report and observe report forms for a fixed sample and found that domain correlations ranged from .45 to .59, demonstrating modest agreement between observer reports and self-reports. They also analyzed data from twin studies in three countries to quantify the amount of variance in facet and domain scores attributable to genetics (heritability), and found that domains are fairly heritable (heritability ranging from .41 to .52). They also analyzed long term stability and found five to ten year retest correlations of .76 to .83 for each of the five domains. Perhaps one of the test’s greatest limitations is its lack of internal validity scales (Kluck, 2014). Other researchers have created an inconsistency scale that can identify patterns of random responding, but the scale has not been standardized (Irving & Weiner, 2011). The lack of established validity scales make the NEO-PI-3’s utility with pathological clients suspect, as patients suffering from certain psychological conditions may be inclined to “fake good” or to randomize their responses out of apathy.

The NEO-PI-3 has a grade reading level of 5.3, with colloquially and simply worded items, making it appropriate for its use with adolescents (McCrae et al., 2005a). Its predecessor,
the NEO-PI-R, had a grade reading level of 5.7, and was analyzed for items that adolescents found difficult to read or understand. Most of these items were replaced with easier to understand ones on the NEO-PI-3 (McCrae et al., 2005a). The test manual also includes a dictionary with definitions for 28 uncommon words used on the test (Kluck, 2014). In addition, the NEO-PI-3 test manual allows an administrator to read the items to the participant if they are unable to read them (Irving & Weiner, 2011). Administering the NEO-PI-3 takes approximately 35 to 45 minutes (NEO Personality Inventory-3), and “the test manual and item booklet directions are clear” (Kluck, 2014). The greatest challenge with administering is assuring that the client is motivated to answer honestly, since the lack of internal validity scales makes identifying social desirability or random responding difficult (Irving & Weiner, 2011). Scoring the test is simple, as it can be done by hand without the need for any additional scoring keys or templates (NEO Personality Inventory-3). Test interpretation is a matter of computing raw scores for each facet and comparing them to the norms provided in the test manual (Irving & Weiner, 2011).

The NEO-PI-3 is, overall, a strong and psychometrically sound measure of personality. Perhaps its greatest strength its basis in the five-factor model of personality, which enjoys a great deal of empirical support and has been observed in cross-cultural and adolescent samples (Costa et al., 2008; Kaplan & Saccuzzo, 2013; De Fruyt et al., 2009). The test is also supported by many different empirical measures of validity. Two important assumptions in personality theory are that traits remain stable and that traits manifest themselves through behavior. Both of these assumptions are supported by the test’s high long term stability and its modest self-observer correlations (McCrae et al., 2011). The test’s other major strength is its accessibility. A low grade reading level and an option for items to be read aloud make literacy almost a non-issue for test administration (Irving & Weiner, 2011; McCrae et al., 2005a), and the entire test can be
administered and scored using only pencil and paper (NEO Personality Inventory-3). Still, certain weaknesses need to be addressed. A major limitation of the test is its lack of retest reliability data (Benson, 2014), which places an upper limit on the test’s ability to accurately report personality stability or change over time. Another major limitation is the lack of internal validity checks, such as social desirability scales (Kluck, 2014). The test’s other limitation is the lack of diversity in the standardization samples, which was predominantly white and overscored high academic achievers (McCrae et al., 2005b). Interesting enough, the NEO-PI-3 was developed to allow for a lower reading level than its predecessor (McCrae et al., 2005a), yet in overselecting high academic achievers, the test developers may have excluded from their standardization sample the very demographic the test was developed for (those with poor reading ability).

The empirical support for the NEO’s five factor model (Costa et al., 2008; Kaplan & Saccuzzo, 2013; De Fruyt et al., 2009), the correlations between its domains and facets (Kluck, 2014), and its long term stability (McCrae et al., 2011) suggest that it is, overall, a valid measure of personality. It would be best used to measure personality in normal individuals-those without serious pathologies or psychological disorders. It is logical to assume that random responding, dishonest responding, and “faking good” or “faking bad” would be more common among pathological individuals, and without built-in validity checks, the NEO-PI-3 is unfit as a measure of personality in patients who may tend towards these test-taking behaviors. The NEO-PI-3 should not be used to measure change or stability in personality traits over time. Its lack of published retest reliability data places an upper limit on its long term validity; changes or stability in a client’s score over time cannot be assumed to represent trait change or stability if retest error is unknown. Thus, further research is still needed for the NEO-PI-3. While the test
has been studied cross-culturally (De Fruyt et al., 2009), norms should be updated to include more ethnically and racially diverse standardization samples. In addition, normative data should be obtained for individuals with low reading ability. The test itself was designed to be very accessible to those with poor literacy, given its low grade reading level (McCrae et al., 2005a) and option for items to be read aloud (Irving & Weiner, 2011), yet its standardization samples are mostly comprised of educated individuals (McCrae et al., 2005b). The test would benefit from including individuals with little to no formal education in future standardization samples, to further validate its use with the population for which it was designed. Studies should also establish formal retest reliability coefficients for the test, to validate its longitudinal use. Future revisions of the test should also include internal validity scales, such as those included in the MMPI-2, to check for random responding and social desirability. This will allow test administrators to identify response styles that may need therapeutic attention. Still, the generally satisfactory psychometric properties of the NEO-PI-3 and its ease of administration make it a promising and useful personality measure; further research will only add to its use as a powerful clinical tool.
References


