The Criterion-Related Validity of Personality Measures for Predicting GPA: A Meta-Analytic Validity Competition

Samuel T. McAbee and Frederick L. Oswald
Rice University

Interest in the role of personality traits in predicting academic performance outcomes has steadily increased over the last several decades, enough to produce a number of meta-analyses that summarize this research (e.g., Poropat, 2009; Richardson, Abraham, & Bond, 2012). These previous meta-analyses combine a variety of alternative personality measures under the assumption that they all reflect the same personality traits and thus predict outcomes similarly. The current meta-analysis tests this assumption by comparing different personality measures when predicting postsecondary grade point average (GPA). The operational validities ($r^*$) of 5 frequently used measures of the Big Five personality traits were compared: the NEO Personality Inventory—Revised (NEO-PI–R; Costa & McCrae, 1992), the NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992), the Big Five Inventory (BFI; e.g., Benet-Martínez & John, 1998), Goldberg’s (1992) unipolar Big Five Factor Markers (Markers), and the Big Five International Personality Item Pool (IPIP; Goldberg, 1999). A systematic review of the psychological literature from 1992 to 2012 was conducted, identifying 51 studies containing 274 correlations. Conscientiousness demonstrated the strongest criterion-related validity for predicting GPA ($r^* = .23$), consistent with previous meta-analyses; in addition, this overall validity was found to be robust across measures ($r^*_{BFI} = .24$, $r^*_{NEO-FFI} = .21$, $r^*_{Markers} = .15$, $r^*_{NEO-PI-R} = .24$, $r^*_{NEO-PI-R} = .26$). Although the criterion-related validities for Extraversion, Agreeableness, Neuroticism, and Openness to Experience (Intellect) differed by measure, they were generally low ($r^*$’s $< .10$). Practical implications of the findings and directions for future research are discussed.

Keywords: personality measures, Five Factor Model, grade point average, meta-analysis, post-secondary education

Supplemental materials: http://dx.doi.org/10.1037/a0031748.supp

The Five-Factor Model (or the Big Five) is a descriptive taxonomy of personality traits, providing a simplified framework for understanding situationally and temporally consistent patterns of thoughts, feelings, and behavior (John, Naumann, & Soto, 2008). The Big Five consists of five factors or traits, usually labeled as Extraversion, Agreeableness, Conscientiousness, Neuroticism (or its reverse, Emotional Stability), and Openness to Experience (Intellect; Digman, 1990). These traits have been studied in a variety of research contexts, resulting in a number of meta-analyses in the areas of job performance (e.g., Barrick & Mount, 1991; Tett, Jackson, & Rothstein, 1991) and academic achievement (e.g., O’Connor & Paunonen, 2007; Poropat, 2009; Richardson, Abraham, & Bond, 2012; Trappmann, Hell, Hrn, & Schuler, 2007). Past meta-analyses have typically combined the results of research findings using different Big Five personality scales, based on the critical but untested assumption that they measure Big Five personality traits in the same way (see Pace & Brannick, 2010; Steel, Schmidt, & Shultz, 2008). As a practical matter, some researchers have questioned whether such personality measures might in fact vary in their validity or effectiveness for predicting important life outcomes (e.g., Gruca & Goldberg, 2007; Hogan, 2005; Hogan, Hogan, & Roberts, 1996; Thalmaier, Saucier, & Eigenhuis, 2011). The current meta-analysis answers this question in the educational context by comparing the validity of five different Big Five personality measures for predicting college student grade point average (GPA).

Personality and Academic Performance

Since the early 20th century, psychological research has investigated relationships between personality traits and academic performance (e.g., Busato, Prins, Elshout, & Hamaker, 2000; Chamorro-Premuzic & Furnham, 2003a; Fleming, 1932; Spearman, 1927). Early research in this area, however, was challenged by a confusing swirl of personality constructs, measures, and empirical results, caused in part by the lack of an overarching theoretical framework for organizing diverse personality constructs (De Raad & Schouwenburg, 1996; Poropat, 2009) and in part due to sampling error variance in empirical findings being mistakenly attributed to differences in theories and measures (Schmidt & Hunter, 1977; Schmidt et al., 1993). The resurgence of applied personality research over the past two decades is largely due to the Five-Factor Model serving as an organizing framework.
There are substantive reasons to expect specific patterns of relationship between the Big Five personality traits and academic performance outcomes (see De Raad & Schouwenburg, 1996; Farsides & Woodfield, 2003; O’Connor & Paunonen, 2007; Poropat, 2009). Of the Big Five traits, Conscientiousness is the most clearly established dimension for predicting performance-related outcomes, as shown in previous meta-analyses of personality and performance in both academic and employment contexts (e.g., Barrick & Mount, 1991; Nofle & Robins, 2007; O’Connor & Paunonen, 2007; Poropat, 2009). Conscientiousness is the higher order trait that subsumes characteristics such as accomplishment (achievement-striving), self-efficacy, organization, orderliness, and self-discipline (Costa & McCrae, 1992; Goldberg, 1992). Although Conscientiousness generally shows high validity for predicting academic performance in meta-analysis (Corker, Oswald, & Donnellan, 2012; e.g., \( r = .23 \), Richardson et al., 2012; \( r = .27 \), Trapmann et al., 2007), researchers have reported a wide range of correlations between Conscientiousness and academic performance outcomes, from as low as \( .06 \) (e.g., Dwight, Cummings, & Glenar, 1998) to over \( .40 \) (e.g., Chamorro-Premuzic, 2006; Kappe & van der Flier, 2010). Openness to Experience (or Intellect) is another Big Five trait that would appear highly relevant to academic performance outcomes, as openness is associated with originality, imagination, interests, and intellectual engagement or curiosity (e.g., Ackerman, Chamorro-Premuzic, & Furnham, 2011; Chamorro-Premuzic & Furnham, 2008; De Raad & Schouwenburg, 1996; Goff & Ackerman, 1992; Lounsbury, Sundstrom, Loveland, & Gibson, 2003; von Stumm, Hell, & Chamorro-Premuzic, 2011). However, the literature reports correlations between Openness to Experience and academic performance that are more modest than might be expected (e.g., \( r = .10 \), Wolfe & Johnson, 1995; \( r = .05 \), Nofle & Robins, 2007; \( r = .12 \), Poropat, 2009), and they have even occasionally been negative (e.g., \( r = -.16 \), Furnham, Chamorro-Premuzic, & McDougall, 2002). Relationships between the other three Big Five dimensions of personality and academic performance are more tenuous (see Nofle & Robins, 2007; O’Connor & Paunonen, 2007; Richardson et al., 2012). Research has generally found a negative relationship between Neuroticism and academic performance outcomes (e.g., \( r = -.05 \), Chamorro-Premuzic & Furnham, 2008; \( r = -.25 \), Chamorro-Premuzic, Furnham, & Ackerman, 2006; \( r = -.03 \), O’Connor & Paunonen, 2007), in line with the hypothesis that students higher in Neuroticism tend to demonstrate higher levels of anxiety and stress that in turn result in lower performance on course examinations and in other high-stakes academic situations (De Raad & Schouwenburg, 1996; see also Ackerman et al., 2011; Chamorro-Premuzic & Furnham, 2005; O’Connor & Paunonen, 2007; Richardson et al., 2012). Although theoretical reasons do not lead one to expect a direct relationship between scores on Big Five Agreeableness scales and students’ overall academic achievement, Agreeableness is generally expected to influence academic performance through its effects on mediating processes, such as being more likely to attend class (Lounsbury et al., 2003; Richardson et al., 2012; Woodfield, Jessop, & McMilan, 2006), and facilitating performance on cooperative tasks (e.g., group course projects; De Raad & Schouwenburg, 1996). As such, individual studies often report inconsistent and sometimes practically nonsignificant find-
native models of personality (the Big Five, Eysenck’s three-factor model; e.g., Barrett, Petrides, Eysenck, & Eysenck, 1998) for predicting subjective well-being (see also Lucas & Fujita, 2000); and Pace and Brannick (2010) meta-analyzed estimates of reliability and convergent validities for a number of personality scales. However, the previous comparative-validity study of greatest relevance to the current meta-analysis is that of Thalmayer et al. (2011), who examined the criterion-related validities of a number of personality measures of various lengths (numbers of items). Thalmayer et al. found that personality traits (e.g., Conscientiousness and the HEXACO trait of honesty–humility; see Ashton et al., 2004) predicted GPA, student conduct criteria (e.g., student housing complaints), and behavioral observations (e.g., punctuality) similarly across scales. Although scales were generally comparable in their patterns of criterion-related validity, there were apparent practical differences between scales in breadth and efficiency for predicting outcomes (e.g., composite HEXACO scales showed slightly higher criterion-related validities).

Most of these previous findings are based on individual studies, whereas the current comparative-validity meta-analysis averages validities within Big Five measures across studies using specialized methods, and then compares these averaged validities across the different measures (cf. Steel et al., 2008). Comparative validity is a practical question when deciding which measure to use in applied or research settings; it also incidentally informs the theoretical assumption about the appropriateness of collapsing across measures in meta-analyses of personality scales, as is typical in practice.

Specific Measures of the Big Five Personality Traits

The current meta-analysis examined the criterion-related validity of several alternative Big Five personality measures that have been widely used to predict GPA, as described below.

The NEO Personality Inventory—Revised. The NEO Personality Inventory—Revised (NEO-PI–R; Costa & McCrae, 1992) is a 240-item measure of the Big Five personality traits (48 items per trait). The NEO-PI–R includes six lower order facet scales per trait (eight items each) that can be combined to obtain scale scores for each of the Big Five traits. Pace and Brannick (2010) provided meta-analytic reliability estimates (α) for the NEO-PI–R scales of Extraversion (.86), Agreeableness (.86), Conscientiousness (.91), Neuroticism (.90), and Openness to Experience (.85).

The NEO Five-Factor Inventory (NEO-FFI; Costa & McCrae, 1992) is a shortened 60-item version of the NEO-PI–R, with 12 items per Big Five trait. Unlike its parent scale, the NEO-FFI was designed to measure the higher order factors of the Big Five without assessing its underlying facets. The NEO-FFI only includes items that were most strongly associated (demonstrated the highest factor loadings) with the Big Five traits rather than representatively sampling across individual facets (McCrae & Costa, 1989; Thalmayer et al., 2011). Pace and Brannick (2010) provided reliability estimates for the NEO-FFI scales of Extraversion (.79), Agreeableness (.74), Conscientiousness (.81), Neuroticism (.80), and Openness to Experience (.75).

The Big Five Inventory. The Big Five Inventory (BFI; Benet-Martinez & John, 1998; John, Donahue, & Kentle, 1991; John & Srivastava, 1999) was developed as a brief, noncommercial measure of the Big Five (John et al., 2008; Thalmayer et al., 2011). The BFI contains 44 items (ranging from eight to 10 items per scale) and was developed to reflect the dimensions of the Big Five identified by Costa and McCrae (1992). John et al. (2008) provided internal consistency reliability estimates for the BFI scales of Extraversion (.86), Agreeableness (.79), Conscientiousness (.82), Neuroticism (.87), and Openness to Experience (.83).

Goldberg’s (1992) unipolar Markers. Goldberg’s unipolar Big Five Factor Markers are a set of descriptive adjectives that reflect the Big Five personality traits, with 20 adjectives per trait, or 100 adjectives total (see Saucier, 1994). Note that the fifth dimension of Goldberg’s Markers is labeled as Intellect, rather than Openness to Experience, incorporating items that are primarily associated with intellectual orientation (e.g., imagination, curiosity, creativity) and excluding other facets of the broader openness construct (e.g., artistic interests, liberalism; see also Thalmayer et al., 2011). Several short-form alternatives to Goldberg’s Markers have been developed for use (Saucier, 1994; Thompson, 2008). Although these scales differ with respect to the number of items included, they are comparable in their item content and formatting, and are hereafter subsumed under the label Markers.

The International Personality Item Pool (IPIP). The IPIP is the product of an ongoing online, public collaborative effort designed to further the progress of personality research in a multinational context (see Goldberg et al., 2006; www.ipip.ori.org). The IPIP scales include 100-item and 50-item measures of the Big Five developed from the IPIP items (Goldberg, 1999). Unlike the various scales from which the IPIP is derived, adjectives for the IPIP scales were embedded in sentences to improve interpretation of the items (see Thalmayer et al., 2011); however, like Goldberg’s Markers, items for the fifth factor of the IPIP scales reflect intellect, rather than openness. Thalmayer et al. further noted that the Agreeableness scale of the IPIP differs from other similarly labeled scales (i.e., the NEO-FFI, BFI) in its “emphasis on empathy and interest in others, and lack of items referring to quarrelsomeness” (p. 997). Pace and Brannick (2010) provided reliability estimates for the (combined) IPIP and Markers scales of Extraversion (.82), Agreeableness (.79), Conscientiousness (.82), Neuroticism (.80), and Intellect (.79).

Potential sources of differences in personality measures and validities. Broadly speaking, personality tests may differ in their definitions of trait domains and the content of the items that measure those domains. These differences, in turn, may result in validity differences for predicting outcomes such as GPA. As an example, the IPIP and Markers Conscientiousness scales are more heavily saturated with items reflecting orderliness and organization facets, with fewer items reflecting achievement-striving (Noffle & Robins, 2007), the latter being a facet of Conscientiousness that has been shown to be highly related to GPA (e.g., Chamorro-Premuzic & Furnham, 2003a; Gray & Watson, 2002). As such, one might expect these scales to demonstrate lower correlations with GPA than those found for Conscientiousness scales that include a greater proportion of achievement-striving-related items.

In addition to differences in the definitions of personality traits on which measures are based, measures also differ in terms of their item content and response format. The contextualization of the items is an important difference between personality measures (Pace & Brannick, 2010), where contextualization refers to the
degree of additional information provided to foster respondent interpretation of an item (see Goldberg & Kilkowski, 1985; John et al., 2008). Some measures like the Markers (e.g., Goldberg, 1992; Saucier, 1994) only provide adjectives (e.g., “talkative”), whereas some measures add context to each personality item (e.g., the IPIP item “talk to a lot of different people at parties”; Goldberg, 1999). Contextualization therefore may reflect a researcher’s definition of how personality is measured, not only the fact that each test must reflect a finite sample of items that can be constructed from a virtually infinite pool of potential items. Even subtle differences in item wording unrelated to the trait can affect its interpretation, hence the need for multiple contextualized items in a scale, in order for construct-relevant covariance between items to dominate item-specific variance.

**Aim of the Current Meta-Analysis**

The current meta-analysis serves to compare the criterion-related validity of several commonly used personality measures for predicting college GPA using meta-analytic techniques. Even though the different personality measures we meta-analyze all claim to measure the same Big Five traits, any differences might translate into differences in predicting important personality-relevant outcomes such as academic performance (GPA). Alternative measures of a construct may be said to be commensurate to the extent that they demonstrate comparable patterns of criterion-related validity (see McDonald, 1999; Sharpe, 1997); however, researchers have expressed some doubt in previous meta-analytic findings that have combined results from diverse sets of personality measures with systematic differences in their construction and underlying definitions of the Big Five personality traits (e.g., Hogan, 2005; Hogan et al., 1996; Pace & Bramnick, 2010; Steel et al., 2008). As such, the current meta-analysis was undertaken to address the practical research question: Do alternative measures of the Big Five personality traits (i.e., the BFI, IPIP, Markers, NEO-FFI, and NEO-PI–R) demonstrate commensurate patterns of criterion-related validity across dimensions for predicting GPA?

**Method**

**Literature Search**

A systematic search of the literature on the relationship between the Big Five personality traits and GPA in postsecondary student populations was conducted to identify relevant studies for the current meta-analysis. Articles published between January 1992 (coinciding with the publication of the NEO-FFI and NEO-PI–R; Costa & McCrae, 1992; and Goldberg’s, 1992, unipolar Markers) and February 2012 were considered for inclusion. Studies were located using several online databases, including PsycINFO, ERIC, and Google Scholar. The following keywords were entered into the respective databases to locate articles in various combinations: personality, Big Five, Five Factor Model, academic performance, academic achievement, GPA, and grade. Following database searches, the following online journal records were manually searched, starting from 2005 to online-first publications as of February 2012: Contemporary Educational Psychology, Human Performance, Intelligence, Learning and Individual Differences, Journal of Applied Psychology, Journal of Educational Psychology, Journal of Research in Personality, Personality and Individual Differences. Finally, the references of previous meta-analyses and large reviews on personality and academic performance were searched to identify additional studies (Nofzige & Robins, 2007; O’Connor & Paunonen, 2007; Poropat, 2009; Richardson et al., 2012; Trapmann et al., 2007). Article abstracts, and Method sections as necessary, were examined to determine whether the study included measures of personality related to academic performance outcomes (e.g., GPA, course grades).

**Inclusion Criteria**

Studies from the literature search were included in the current meta-analysis if we could identify or calculate at least one correlation between the Big Five and GPA in postsecondary student populations. Academic performance reflected official college GPA, self-reported college GPA, or individual course grades, but official GPA was the preferred criterion of interest when it was available. Table 1 lists the information extracted from the individual articles included in the meta-analysis. All studies included in the current meta-analysis were published in English.

Effects were excluded from the current meta-analysis for several reasons. First, we limited our analyses to the specific Big Five scales of interest (i.e., the BFI, IPIP, Markers, NEO-FFI, and NEO-PI–R). We therefore excluded effects that were identified as relevant but used other personality scales (e.g., Busato et al., 2000; Lounsbury et al., 2003; Paunonen & Ashton, 2001b), and effects based on scales that reflect alternative models of personality such as the Myers-Briggs Trait Indicator (Myers, McCaulley, Quenk, & Hammer, 1998; e.g., Kahn, Nauta, Gailbreath, Tipps, & Chartrand, 2002), the HEXACO model of personality (Ashton et al., 2004; e.g., de Vries, de Vries, & Born, 2011), and Eysenck’s three-factor model of personality (Barrett et al., 1998; e.g., Petrides, Chamorro-Premuzic, Fredrickson, & Furnham, 2005). Second, we excluded effects on the basis of noncollege samples (e.g., high school, military), effects apparently based on the same data and reported across multiple studies (e.g., Komarraju et al., 2009; Komarraju, Karau, Schmeck, & Avdic, 2011; Sisco & Reilly, 2007a, 2007b), and effects that did not appear to include a relevant academic outcome (e.g., Natalianis, 2010; Paunonen & Ashton, 2001a). Finally, we excluded a single study (Colquitt & Simmering, 1998) in which GPA was intentionally range restricted to 3.75 or lower. As a result of these procedures, 51 individual studies (57 independent samples) identified in the literature search were included in the current meta-analysis.

The current meta-analysis included several studies with unique aspects that deserve mention. First, correlations for five studies were tied to experimental manipulations (Hirsh & Peterson, 2008; Huws, Reddy, & Talcott, 2009; Lievens, De Corte, & Schollaert, 2008; Peeters & Lievens, 2005; Schmit, Ryan, Stierwalt, & Powell, 1995), where we extracted correlations between personality scales and GPA only when participants did not receive an experimental manipulation at the time they responded to personality measures (e.g., control groups or initial screening). Second, three of the studies only reported personality—GPA correlations by gender (Richardson & Abraham, 2009; Wintre & Sugar, 2000; Woodfield et al., 2006). Here we extracted correlations separately for males and females and entered them as independent samples.
Third, for those studies with tables that only indicated a range of sample sizes for correlations between the Big Five and GPA (e.g., Dwight et al., 1998; Thalmayer et al., 2011; Woodfield et al., 2006), we took the average of each range as the study sample size that went into meta-analytic calculations. Finally, one longitudinal study (Lievens, Ones, & Dilchert, 2009) reported correlations between personality scales and GPA over a 7-year period. For this study, only the correlation between participants’ personality scale scores and first-year GPA were extracted in order to retain the largest available sample and be comparable with other effects in the analysis.

### Meta-Analytic Procedures

Meta-analyses of relationships between the Big Five personality traits and GPA were conducted at overall (all measures included) and scale-specific (BFI, IPIP, Markers, NEO-FFI, and NEO-PI–R) levels. Three meta-analytic procedures were used. First, sample size-weighted validities were averaged to yield conservative meta-analytic estimates in the sense that they correct for sampling error variance alone (known as “bare-bones” meta-analysis; Hunter & Schmidt, 2004). Next, random effects meta-analyses were conducted to correct for both sampling error variance and unreliability in the GPA criteria. These analyses are of general interest to researchers and practitioners who are concerned with the use of personality scales as they operate in real-world conditions (i.e., measurement unreliability should not be corrected in these predictors), yet they seek to predict a criterion construct of interest and not the criterion measure itself (i.e., measurement unreliability in GPA should be corrected for as an estimate of academic performance). Finally, random effects meta-analyses were conducted that correct for sampling error variance and measurement error variance (unreliability) in both the predictor (Big Five personality) and criterion (GPA) measures. This analysis is of interest to researchers and practitioners concerned with the underlying (latent) relationship in both the Big Five and GPA. Random effects models are generally appropriate for meta-analysis (see Borenstein, Hedges, Higgins, & Rothstein, 2010; Hedges & Vevea, 1998; National Research Council, 1992; Schmidt, Oh, & Hayes, 2009), as they are designed to estimate meaningful heterogeneity when it exists, as it might between personality scales. Furthermore, previous meta-analyses examining the relationship between personality traits and academic performance (e.g., Richardson et al., 2012) have generally used random-effects models.

### Reliability estimation

The vast majority of studies examining the relationship between personality and academic performance fail to provide estimates of internal consistency reliability for academic performance criteria, complicating the meta-analytic correction of observed correlations for measurement error in the criterion. Previous meta-analyses in this area have used a variety of approaches to perform such corrections (see Poropat, 2009; Richardson et al., 2012; Trapmann et al., 2007). One way to conceptualize the reliability of GPA is to treat official undergraduate GPA in one of two ways: either as the criterion that is to be predicted, including any minor errors in how it is reported, or a proxy for the true score of latent academic performance (see Kuncel, Credé, & Thomas, 2005; Richardson et al., 2012). In either case, official GPA is assumed to be perfectly reliable (i.e., reliability of official GPA = 1.0), with students’ self-reported GPA and individual course grades as less reliable indicators of official GPA. Under this framework, an estimate of reliability for self-reported undergraduate GPA and individual course grades is obtained from their respective squared correlations with official GPA (see McDonald, 1999). Reliability estimates for official GPA were obtained for five

---

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Information Collected From Individual Studies</strong></td>
</tr>
<tr>
<td><strong>Item</strong></td>
</tr>
<tr>
<td><strong>Study descriptive information</strong></td>
</tr>
<tr>
<td>Full APA reference</td>
</tr>
<tr>
<td>Type of publication</td>
</tr>
<tr>
<td>2 = Book/manual</td>
</tr>
<tr>
<td>3 = Dissertation</td>
</tr>
<tr>
<td>4 = Conference paper</td>
</tr>
<tr>
<td>5 = Unpublished source</td>
</tr>
<tr>
<td>Year of publication</td>
</tr>
<tr>
<td>Type of sample</td>
</tr>
<tr>
<td>2 = Undergraduate</td>
</tr>
<tr>
<td>3 = Other</td>
</tr>
<tr>
<td>Total sample size</td>
</tr>
<tr>
<td>Location of study</td>
</tr>
<tr>
<td>2 = Great Britain</td>
</tr>
<tr>
<td>3 = Europe</td>
</tr>
<tr>
<td>4 = Other</td>
</tr>
<tr>
<td>Research design</td>
</tr>
<tr>
<td>2 = Longitudinal</td>
</tr>
<tr>
<td>3 = Experimental/within subjects</td>
</tr>
<tr>
<td>4 = Experimental/within subjects</td>
</tr>
<tr>
<td>Sample size used in analysis</td>
</tr>
<tr>
<td>Personality scale information</td>
</tr>
<tr>
<td>Personality scale</td>
</tr>
<tr>
<td>2 = NEO-FFI (Costa &amp; McCrae, 1992)</td>
</tr>
<tr>
<td>3 = IP (e.g., Goldberg, 1999)</td>
</tr>
<tr>
<td>4 = Other</td>
</tr>
<tr>
<td>5 = NEO-PI–R (Costa &amp; McCrae, 1992)</td>
</tr>
<tr>
<td>6 = Markers (e.g., Goldberg, 1992; Saucier, 1994)</td>
</tr>
<tr>
<td>Number of scale items</td>
</tr>
<tr>
<td>Scale mean</td>
</tr>
<tr>
<td>Scale SD</td>
</tr>
<tr>
<td>Reliability (α) of scale</td>
</tr>
<tr>
<td>Neuroticism or emotional stability</td>
</tr>
<tr>
<td>−1 = Emotional Stability</td>
</tr>
<tr>
<td>Outcome information</td>
</tr>
<tr>
<td>Outcome variable</td>
</tr>
<tr>
<td>2 = Course grades</td>
</tr>
<tr>
<td>3 = GPA (Marks)</td>
</tr>
<tr>
<td>4 = Course exam</td>
</tr>
<tr>
<td>5 = Other</td>
</tr>
<tr>
<td>Outcome mean (e.g., GPA)</td>
</tr>
<tr>
<td>Outcome SD (e.g., GPA)</td>
</tr>
<tr>
<td>Reliability (α) for outcome measure</td>
</tr>
<tr>
<td>Validity coefficient between scale and outcome</td>
</tr>
</tbody>
</table>

*Note. APA = American Psychological Association; BFI = Big Five Inventory; NEO-FFI = NEO Five-Factor Inventory; IP = International Personality Item Pool; NEO-PI–R = NEO Personality Inventory—Revised; GPA = grade point average. *Values were coded separately for each dimension of the Big Five.
studies in the meta-analysis (Chamorro-Premuzic, 2006; Chamorro-Premuzic et al., 2006; Higgins, Peterson, Pihl, & Lee, 2007; Kappe & van der Flier, 2010; Lievens & Coetsier, 2002). For all other studies, we relied on the Kuncel et al. (2005) meta-analysis that reported a correlation of \( r = .90 \) between self-reported GPA and official GPA (i.e., reliability of self-reported GPA = \( r^2 = .81 \)), and Richardson et al. (2012), who reported a meta-analytic correlation of \( r = .60 \) between individual course grades and official GPA (i.e., reliability of individual course grades = \( r^2 = .36 \)).

Internal consistency reliability estimates for the respective personality scales were obtained from primary studies to correct for measurement error variance. When a study did not report observed reliabilities (\( \alpha \)), reliability estimates provided by Pace and Branick (2010; IPIP, Markers, NEO-FFI, NEO-PI–R) and John et al. (2008; BFI) were applied. Consistent with previous meta-analyses (e.g., Poropat, 2009; Richardson et al., 2012), we did not correct for range restriction effects, as studies in this area do not typically report adequate information to perform such corrections, and these effects are thought to be small.

**Dependent correlations.** A small number of studies identified in the literature search included multiple correlations between personality measures and GPA within the same sample. In these cases, only the relationship between Big Five scales and the most stable estimate of undergraduate GPA (official or self-report) was coded when available. However, three studies included criterion-related validities within samples of equal relevance for the purposes of the current meta-analyses: Two studies provided dependent correlations between multiple personality scales (e.g., BFI, NEO-FFI, and IPIP) and GPA for one sample (Biderman & Reddon, 2012; Thalmayer et al., 2011), and one study presented correlations for the same personality scale with multiple relevant GPA outcomes (independent semester GPAs; McKenzie & Gow, 2004). Given only a handful of dependent effects, sample size-weighted average correlations and reliabilities were computed for each of these studies for the overall and individual scale meta-analyses as appropriate.

**Outlier identification.** For the current research, studies reporting sample sizes greater than twice the size of the next largest sample within a specific meta-analysis were considered to be outliers (cf. Pace & Branick, 2010). One sample (\( n = 10,497 \), \( \bar{z}_{overall} = 7.41 \); Nofle & Robins, 2007) was substantially larger than any of the remaining samples included in the overall (next largest \( n = 1,193 \); McKenzie & Gow, 2004) and BFI-specific meta-analyses (next largest \( n = 437 \); Richardson & Abraham, 2009). In addition, one sample for the NEO-FFI specific meta-analyses (\( n = 1,193 \), \( z_{NEO-FFI} = 3.88 \); McKenzie & Gow, 2004) was more than twice the size of the next largest sample (\( n = 475 \); Nofle & Robins, 2007). Results are reported and tabulated with and without these outlier (large \( N \)) studies included.

**Results**

In total, 274 correlations were obtained from the 51 studies (57 independent samples) identified through the literature search. Of these, 199 correlations were obtained for the Big Five dimensions of personality with official college GPA, 63 with self-reported estimates of GPA, and 12 with individual course grades. Sample sizes for the individual studies ranged from 70 (Chamorro-Premuzic & Furnham, 2003b) to 10,497 (Nofle & Robins, 2007). Nine studies (\( k = 10 \)) used the BFI (e.g., Benet-Martínez & John, 1998); 10 studies (\( k = 10 \)) used some variation of the IPIP scales (e.g., Goldberg, 1999); five studies (\( k = 5 \)) used Goldberg’s (1992) unipolar Markers or its short-form alternatives (two studies: Saucier, 1994; one study: Thompson, 2008); 19 studies (\( k = 21 \)) used the NEO-FFI; and 14 studies (\( k = 15 \)) used the NEO-PI–R (Costa & McCrae, 1992).

Tables 2–6 present the meta-analytic findings for the relationship between the Big Five personality traits and GPA at the overall and scale-specific levels. Results are presented (a) for the sample-weighted (bare-bones) analyses correcting for sampling error variance only, (b) for the random-effects analysis correcting for sampling error variance and measurement unreliability in the criterion measure (i.e., operational validities), and (c) for the meta-analyses correcting for sampling error variance and measurement unreliability in both the predictor and the criterion. The 95% confidence intervals for bare-bones estimates are statistically significant when they do not include zero. Operational validities (\( r^* \)) for the individual scales with outlier (large \( N \)) samples removed are discussed, as these estimates are likely to be of greatest interest to researchers and practitioners when deciding between various measures of personality.

Meta-analytic validities for Conscientiousness were consistent, and they were statistically and practically significant for predicting college GPA, both across measures (\( r^*_{overall} = .23 \)) and for specific measures (\( r_{BFI} = .24, r_{IPIP} = .21, r_{Markers} = .15, r_{NEO-FFI} = .24, r_{NEO-PI–R} = .26; \) Table 2). Also, within each of the measures, Conscientiousness always had the highest operational validity relative to those for the other four personality traits. The 95% credibility intervals for the individual Conscientiousness scales were generally wide but did not include zero, suggesting that the population validities for the relationship between Conscientiousness and GPA are heterogeneous but positive for each Conscientiousness scale across all measures. Differences in the operational validities were calculated in a pairwise manner between the specific Conscientiousness scales. In general, these differences were not statistically significant (\( p > .05 \)); although the criterion-related validity for the Markers Conscientiousness scale was statistically significantly lower for predicting GPA than the validity for similar scales on the BFI (\( r_{BFI} – r_{Markers} = .09, 95% CI [.02, .16] \), NEO-FFI (\( r_{NEO-FFI} – r_{Markers} = .09, 95% CI [.01, .17] \)), and NEO-PI–R (\( r_{NEO-PI–R} – r_{Markers} = .10, 95% CI [.05, .17] \)), but the practical magnitudes of these differences were small to weak (Cohen, 1988).

The overall meta-analytic correlation between Extraversion and GPA was nonsignificant and weak, at \( r^* = -.03 \) (see Table 3). The operational validities for all of the specific Extraversion scales were also weak and nonsignificant (e.g., \( r_{NEO-FFI} = -.03 \)). Going beyond the meta-analytic operational validities for each measure, the differences in these operational validities were calculated between each pair of specific Extraversion scales; no such differences were statistically significant (all \( p > .05 \)).

The overall meta-analytic correlation between Agreeableness and GPA was statistically significant but weak, with \( r^* = .08 \) (see Table 4). The NEO-FFI Agreeableness scale produced the largest meta-analytic correlation with GPA for the measure-specific meta-analyses (\( r_{NEO-FFI} = .11 \)). The criterion-related validities for the majority of the scales were statistically significant; however, that
for the Markers scale was nonsignificant due to the relatively few samples identified for this scale ($k_{\text{Markers}} = 4$ vs. $k_{\text{NEO-FFI}} = 18$). The 95% credibility intervals for the Markers and NEO-FFI Agreeableness scales crossed zero, suggesting that the operational validities for these two scales may not generalize to future settings. Differences in the operational validities were calculated for each pair of specific Agreeableness scales; no such differences were statistically significant (all $p > .05$).

The overall meta-analytic correlation between Neuroticism and GPA was statistically significant but weak, with $r^* = -.03$ (see Table 5). For the measure-specific meta-analyses, the operational validities for the IPIP and Markers Neuroticism scales were also statistically significant but weak (e.g., $r_{\text{IPIP}} = -.06$); however, the operational validities for the NEO-FFI and NEO-PI–R Neuroticism scales were nonsignificant. Unexpectedly, the operational validity for the BFI Neuroticism scale was weakly and positively related to college GPA ($r_{\text{BFI}} = .06$), whereas the other specific Neuroticism scales all demonstrated weak and negative relationships with GPA. The 95% credibility intervals for the BFI, IPIP, and Markers scales did not include zero, suggesting that the validities for these scales, albeit weak, may generalize to future settings. Differences in the operational validities were calculated between each pair of specific Neuroticism scales. In general, these differences were not statistically significant ($p > .05$); however, the operational validity for the BFI Neuroticism scale was positive, as mentioned, and significantly differed from all other scales whose validities were negative in sign.

The overall meta-analytic correlation between Openness to Experience and GPA was statistically significant but weak, with $r^* = .08$ (see Table 6). For measure-specific meta-analyses, the NEO-FFI Openness to Experience scale produced the highest meta-analytic correlation with GPA ($r^* = .12$), but also exhibited wide associated heterogeneity across studies, including negative values. The Markers Intellect scale showed similar validity ($r^* = .11$), but was not statistically significant due to the relatively few samples identified ($k_{\text{Markers}} = 4$ vs. $k_{\text{NEO-FFI}} = 19$). The meta-analytic validities for the NEO-FFI Openness to Experience and IPIP Intellect scales were statistically significant, but those for the BFI and NEO-PI–R Openness to Experience scales were not. The 95% credibility intervals for the BFI Openness to Experience and

| Measure | $N$ | $k$ | $r$ | $SD_r$ | 95% CI | $r^*$ | $SD_{r^*}$ | 95% CR$_r$ | $\rho$ | $SD_\rho$ | 95% CR$_\rho$
|---------|----|----|-----|-------|-------|------|---------|-----------|------|-------|---------|
| Overall | 24,740 | 50 | -.02 | 0.05 | [-0.04, 0.00] | -.02 | 0.05 | [-0.13, 0.08] | -.03 | 0.06 | [-0.14, 0.09]
| Overall (– outlier) | 14,243 | 49 | -.02 | 0.07 | [-0.05, 0.00] | -.03 | 0.07 | [-0.16, 0.11] | -.03 | 0.08 | [-0.17, 0.12]
| BFI | 12,152 | 8 | -.02 | 0.00 | [-0.04, 0.01] | -.03 | 0.00 | [-0.03, 0.03] | -.03 | 0.00 | [-0.03, 0.03]
| BFI (– outlier) | 1,655 | 7 | -.04 | 0.00 | [-0.08, 0.00] | -.04 | 0.00 | [-0.04, 0.04] | -.04 | 0.00 | [-0.04, 0.04]
| IPIP | 2,271 | 8 | -.02 | 0.07 | [-0.09, 0.04] | -.02 | 0.07 | [-0.17, 0.13] | -.02 | 0.08 | [-0.18, 0.14]
| Markers | 1,165 | 4 | -.01 | 0.00 | [-0.02, 0.01] | -.01 | 0.00 | [-0.01, 0.01] | -.01 | 0.00 | [-0.01, 0.01]
| NEO-FFI | 5,684 | 20 | -.03 | 0.04 | [-0.06, 0.00] | -.03 | 0.04 | [-0.10, 0.04] | -.03 | 0.04 | [-0.11, 0.04]
| NEO-FFI (– outlier) | 4,491 | 19 | -.02 | 0.04 | [-0.06, 0.01] | -.03 | 0.04 | [-0.11, 0.06] | -.03 | 0.05 | [-0.12, 0.06]
| NEO-PI–R | 3,884 | 12 | -.01 | 0.11 | [-0.09, 0.06] | -.01 | 0.11 | [-0.24, 0.21] | -.02 | 0.12 | [-0.26, 0.22]

Note. GPA = grade point average; $k$ = number of independent samples; $r$ = sample-weighted mean correlation; CI = confidence interval; $r^*$ = mean correlation corrected for measurement error variance in the criterion only (operational validity); CR = credibility interval; $\rho$ = mean correlation corrected for measurement error variance in both predictor and criterion; Overall = all measures included; – outlier = outlier effect removed and meta-analysis recalculated; BFI = Big Five Inventory; IPIP = International Personality Item Pool; Markers = Goldberg’s Big Five Factor Markers, Saucier’s Mini-Markers, Thompson’s Mini-Markers; NEO-FFI = NEO Five Factor Inventory; NEO-PI–R = NEO Personality Inventory—Revised. 95% confidence intervals that do not include zero are statistically significant.
IPIP Intelllect scales did not include zero, suggesting that the weak operational validities for these two scales may generalize to future settings. Differences in the operational validities were tested between each pair of specific Openness to Experience/Intelllect scales. In general, these differences were not statistically significant ($p > .05$); however, the operational validity for the NEO-FFI Openness to Experience scale was significantly higher for predicting GPA than that for the BFI ($r_{NEO-FFI} = .10, 95\% \text{CI} [.04, .16]$) and the NEO-PI–R Openness to Experience scales ($r_{NEO-FFI} = .07, 95\% \text{CI} [.00, .14]$), but the practical magnitudes of these differences were small to weak (Cohen, 1988).

Discussion

A general but untested assumption in meta-analysis is that alternative measures of presumably the same construct are sufficiently comparable to allow for their aggregation for predicting important outcomes (cf. Oswald & McCloy, 2003; Pace & Branrick, 2010; Steel et al., 2008). The current meta-analysis examined this assumption explicitly in the context of comparing the criterion-related validities of several widely used Big Five personality scales (the BFI; IPIP; Goldberg’s, 1992, Markers; NEO-FFI; NEO-PI–R) for predicting college GPA.

Findings for the meta-analytic validities of the Big Five personality dimensions for each measure were generally similar to those found for previous meta-analyses of the relationship between the Big Five and GPA that aggregate across measures (e.g., O’Connor & Paunonen, 2007; Poropat, 2009; Richardson et al., 2012). Consistent with previous research, Conscientiousness demonstrated the strongest overall criterion-related validity for predicting college GPA ($r^* = .23$), and this finding was robust across the specific personality measures ($r_{BFI} = .24$, $r_{IPIP} = .21$, $r_{Markers} = .15$, $r_{NEO-FFI} = .24$, $r_{NEO-PI–R} = .26$). Openness to Experience (Intelllect) demonstrated a weak, but positive overall relationship with GPA, and operational validities were generally weak to small in magnitude across measures (e.g., $r_{BFI} = .02$ vs. $r_{NEO-FFI} = .12$). Agreeableness also demonstrated a weak but positive overall relationship with GPA, and this finding was generally consistent across scales. The NEO-FFI measure demonstrated the highest criterion-related validities for scales marked * statistically significantly differ from scales marked $^b$, $p < .05$. 

### Table 5

<table>
<thead>
<tr>
<th>Measure</th>
<th>$N$</th>
<th>$k$</th>
<th>$r$</th>
<th>$SD_r$</th>
<th>95% CI</th>
<th>$r^*$</th>
<th>95% CI</th>
<th>$\rho$</th>
<th>SD$_{\rho}$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>24,968</td>
<td>51</td>
<td>.00</td>
<td>.06</td>
<td>[.02, .02]</td>
<td>.00</td>
<td>.06</td>
<td>[.13, .12]</td>
<td>.00</td>
<td>.07</td>
</tr>
<tr>
<td>Overall (outlier)</td>
<td>14,471</td>
<td>50</td>
<td>.03</td>
<td>.06</td>
<td>[.05, .01]</td>
<td>.03</td>
<td>.06</td>
<td>[.16, .09]</td>
<td>.03</td>
<td>.07</td>
</tr>
<tr>
<td>BFI</td>
<td>12,152</td>
<td>8</td>
<td>.04</td>
<td>.00</td>
<td>[.03, .05]</td>
<td>.05</td>
<td>.00</td>
<td>[.05, .05]</td>
<td>.05</td>
<td>.00</td>
</tr>
<tr>
<td>BFI (outlier)</td>
<td>1,655</td>
<td>7</td>
<td>.06</td>
<td>.00</td>
<td>[.03, .08]</td>
<td>.06</td>
<td>.00</td>
<td>[.06, .06]</td>
<td>.06</td>
<td>.00</td>
</tr>
<tr>
<td>IPIP$^b$</td>
<td>2,499</td>
<td>9</td>
<td>.05</td>
<td>.00</td>
<td>[.09, .02]</td>
<td>.06</td>
<td>.00</td>
<td>[.06, .06]</td>
<td>.06</td>
<td>.00</td>
</tr>
<tr>
<td>Markers$^b$</td>
<td>1,165</td>
<td>4</td>
<td>.07</td>
<td>.00</td>
<td>[.11, .02]</td>
<td>.08</td>
<td>.00</td>
<td>[.08, .08]</td>
<td>.08</td>
<td>.00</td>
</tr>
<tr>
<td>NEO-FFI</td>
<td>5,684</td>
<td>20</td>
<td>.03</td>
<td>.07</td>
<td>[.07, .01]</td>
<td>.03</td>
<td>.07</td>
<td>[.17, .10]</td>
<td>.04</td>
<td>.08</td>
</tr>
<tr>
<td>NEO-FFI (outlier)$^b$</td>
<td>4,491</td>
<td>19</td>
<td>.04</td>
<td>.08</td>
<td>[.08, .01]</td>
<td>.04</td>
<td>.08</td>
<td>[.19, .12]</td>
<td>.04</td>
<td>.09</td>
</tr>
<tr>
<td>NEO-PI–R$^b$</td>
<td>3,884</td>
<td>12</td>
<td>.04</td>
<td>.08</td>
<td>[.10, .02]</td>
<td>.04</td>
<td>.08</td>
<td>[.20, .12]</td>
<td>.05</td>
<td>.09</td>
</tr>
</tbody>
</table>

Note. GPA = grade point average; $k$ = number of independent samples; $r$ = sample-weighted mean correlation; CI = confidence interval; $r^*$ = mean correlation corrected for measurement error variance in both predictor and criterion; Overall = all measures included; + outlier = outlier effect removed and meta-analysis recalculated; BFI = Big Five Inventory; IPIP = International Personality Item Pool; Markers = Goldberg’s Big Five Factor Markers, Saucier’s Mini-Markers, Thompson’s Mini-Markers; NEO-FFI = NEO Five Factor Inventory; NEO-PI–R = NEO Personality Inventory—Revised. 95% confidence intervals that do not include zero are statistically significant. Operational validities for scales marked * statistically significantly differ from scales marked $^b$, $p < .05$. 

### Table 4

<table>
<thead>
<tr>
<th>Measure</th>
<th>$N$</th>
<th>$k$</th>
<th>$r$</th>
<th>$SD_r$</th>
<th>95% CI</th>
<th>$r^*$</th>
<th>95% CI</th>
<th>$\rho$</th>
<th>SD$_{\rho}$</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>24,968</td>
<td>51</td>
<td>.00</td>
<td>.06</td>
<td>[.02, .02]</td>
<td>.00</td>
<td>.06</td>
<td>[.13, .12]</td>
<td>.00</td>
<td>.07</td>
</tr>
<tr>
<td>Overall (outlier)</td>
<td>14,471</td>
<td>50</td>
<td>.03</td>
<td>.06</td>
<td>[.05, .01]</td>
<td>.03</td>
<td>.06</td>
<td>[.16, .09]</td>
<td>.03</td>
<td>.07</td>
</tr>
<tr>
<td>BFI</td>
<td>12,152</td>
<td>8</td>
<td>.04</td>
<td>.00</td>
<td>[.03, .05]</td>
<td>.05</td>
<td>.00</td>
<td>[.05, .05]</td>
<td>.05</td>
<td>.00</td>
</tr>
<tr>
<td>BFI (outlier)</td>
<td>1,655</td>
<td>7</td>
<td>.06</td>
<td>.00</td>
<td>[.03, .08]</td>
<td>.06</td>
<td>.00</td>
<td>[.06, .06]</td>
<td>.06</td>
<td>.00</td>
</tr>
<tr>
<td>IPIP$^b$</td>
<td>2,499</td>
<td>9</td>
<td>.05</td>
<td>.00</td>
<td>[.09, .02]</td>
<td>.06</td>
<td>.00</td>
<td>[.06, .06]</td>
<td>.06</td>
<td>.00</td>
</tr>
<tr>
<td>Markers$^b$</td>
<td>1,165</td>
<td>4</td>
<td>.07</td>
<td>.00</td>
<td>[.11, .02]</td>
<td>.08</td>
<td>.00</td>
<td>[.08, .08]</td>
<td>.08</td>
<td>.00</td>
</tr>
<tr>
<td>NEO-FFI</td>
<td>5,684</td>
<td>20</td>
<td>.03</td>
<td>.07</td>
<td>[.07, .01]</td>
<td>.03</td>
<td>.07</td>
<td>[.17, .10]</td>
<td>.04</td>
<td>.08</td>
</tr>
<tr>
<td>NEO-FFI (outlier)$^b$</td>
<td>4,491</td>
<td>19</td>
<td>.04</td>
<td>.08</td>
<td>[.08, .01]</td>
<td>.04</td>
<td>.08</td>
<td>[.19, .12]</td>
<td>.04</td>
<td>.09</td>
</tr>
<tr>
<td>NEO-PI–R$^b$</td>
<td>3,884</td>
<td>12</td>
<td>.04</td>
<td>.08</td>
<td>[.10, .02]</td>
<td>.04</td>
<td>.08</td>
<td>[.20, .12]</td>
<td>.05</td>
<td>.09</td>
</tr>
</tbody>
</table>
Table 6

Meta-Analytic Correlations Between Openness to Experience/Intellect Scales and Undergraduate GPA

<table>
<thead>
<tr>
<th>Measure</th>
<th>N</th>
<th>k</th>
<th>r</th>
<th>SD</th>
<th>95% CI</th>
<th>r²</th>
<th>SD²</th>
<th>95% CR²</th>
<th>ρ</th>
<th>SD</th>
<th>95% CR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>24,996</td>
<td>51</td>
<td>0.07</td>
<td>0.06</td>
<td>[0.05, 0.09]</td>
<td>0.07</td>
<td>0.06</td>
<td>[-0.05, 0.19]</td>
<td>0.08</td>
<td>0.07</td>
<td>[-0.06, 0.22]</td>
</tr>
<tr>
<td>Overall (− outlier)</td>
<td>14,499</td>
<td>50</td>
<td>0.07</td>
<td>0.07</td>
<td>[0.05, 0.10]</td>
<td>0.08</td>
<td>0.08</td>
<td>[-0.08, 0.23]</td>
<td>0.08</td>
<td>0.09</td>
<td>[-0.09, 0.26]</td>
</tr>
<tr>
<td>BFI</td>
<td>12,152</td>
<td>8</td>
<td>0.05</td>
<td>0.01</td>
<td>[0.04, 0.07]</td>
<td>0.06</td>
<td>0.02</td>
<td>[0.03, 0.09]</td>
<td>0.07</td>
<td>0.02</td>
<td>[0.03, 0.10]</td>
</tr>
<tr>
<td>BFI (− outlier)</td>
<td>1,655</td>
<td>7</td>
<td>0.02</td>
<td>0.00</td>
<td>[0.03, 0.06]</td>
<td>0.02</td>
<td>0.00</td>
<td>[0.02, 0.02]</td>
<td>0.02</td>
<td>0.00</td>
<td>[0.02, 0.02]</td>
</tr>
<tr>
<td>IPPI</td>
<td>2,271</td>
<td>8</td>
<td>0.07</td>
<td>0.00</td>
<td>[0.04, 0.11]</td>
<td>0.07</td>
<td>0.00</td>
<td>[0.07, 0.07]</td>
<td>0.08</td>
<td>0.00</td>
<td>[0.08, 0.08]</td>
</tr>
<tr>
<td>Markers</td>
<td>1,165</td>
<td>4</td>
<td>0.09</td>
<td>0.12</td>
<td>[-0.04, 0.22]</td>
<td>0.11</td>
<td>0.14</td>
<td>[-0.16, 0.38]</td>
<td>0.12</td>
<td>0.15</td>
<td>[-0.18, 0.42]</td>
</tr>
<tr>
<td>NEO-FFI</td>
<td>5,684</td>
<td>20</td>
<td>.10</td>
<td>.06</td>
<td>[0.06, 0.14]</td>
<td>.11</td>
<td>0.07</td>
<td>[-0.03, 0.25]</td>
<td>.12</td>
<td>0.08</td>
<td>[-0.04, 0.29]</td>
</tr>
<tr>
<td>NEO-FFI (− outlier)</td>
<td>4,491</td>
<td>19</td>
<td>.11</td>
<td>.07</td>
<td>[0.07, 0.15]</td>
<td>.12</td>
<td>0.08</td>
<td>[-0.04, 0.27]</td>
<td>.13</td>
<td>0.09</td>
<td>[-0.05, 0.32]</td>
</tr>
<tr>
<td>NEO-PI–R³</td>
<td>4,140</td>
<td>13</td>
<td>.05</td>
<td>.08</td>
<td>[-0.00, 0.10]</td>
<td>.05</td>
<td>0.08</td>
<td>[-0.11, 0.21]</td>
<td>.05</td>
<td>0.09</td>
<td>[-0.12, 0.23]</td>
</tr>
</tbody>
</table>

Note. GPA = grade point average; k = number of independent samples; r = sample-weighted mean correlation; CI = confidence interval; r² = mean correlation corrected for measurement error variance in the criterion only (operational validity); CR = credibility interval; ρ = mean correlation corrected for measurement error variance in both predictor and criterion; Overall = all measures included; − outlier = outlier effect removed and meta-analysis recalculated; BFI = Big Five Inventory; IPPI = International Personality Item Pool; Markers = Goldberg’s Big Five Factor Markers, Saucier’s Mini-Markers, Thompson’s Mini-Markers; NEO-FFI = NEO Five Factor Inventory; NEO-PI–R = NEO Personality Inventory—Revised. 95% confidence intervals that do not include zero are statistically significant. Operational validities for scales marked * statistically significantly differ from scales marked **, p < .05.

Future Research Considerations

One future research consideration concerns the fact that the current meta-analysis had relative few studies for three of the five measures: Goldberg’s (1992) unipolar Markers and related scales (i.e., Saucier, 1994; Thompson, 2008; five studies), the BFI (nine studies; e.g., Benet-Martínez & John, 1998), and the IPPI (10 studies; Goldberg, 1999). We were able to analyze more studies for the NEO-PI–R (19 studies) and the NEO-FFI (14 studies; Costa & McCrae, 1992). Note that the noncommercial Big Five scales (e.g., the IPPI and Markers) are becoming more popular in use. Therefore, it may be useful to reexamine the issue of criterion-related validities across Big Five measures as future research accumulates for noncommercial and web-based personality scales.

A second direction for future research in light of the current meta-analysis is to examine personality traits that extend beyond the Big Five. Although the Big Five is the predominant organizing structure of personality traits in the research literature, research and practice have begun to investigate the six-factor HEXACO model more thoroughly, which includes the Big Five plus a sixth honesty–humility factor (Ashton, Lee, & Goldberg, 2007; Ashton et al., 2004). In addition, Eysenck’s broader three-factor model of personality (extraversion, neuroticism, psychoticism; e.g., Barrett et al., 1998) has been used to predict academic performance (see Poropat, 2011a). Given at least some differences in the theoretical nature, organization, and aggregation of personality traits, questions regarding the criterion-related validity of personality scales from alternative theoretical models may reflect different magnitudes and patterns of validity than those reported here for Big Five measures. Thus, future meta-analytic comparative-validity studies might include scales from different personality models (cf. Grucza & Goldberg, 2007; Steel et al., 2008; Thalmayer et al., 2011).

A third research possibility that would extend the current meta-analysis is to focus on the comparative validity of measures of the narrower personality facets underlying Big Five factors for predicting academic performance, as has been recommended by others (Chamorro-Premuzic & Furnham, 2003a; Corker et al., 2012; De Fruyt & Mervielde, 1996; de Vries et al., 2011; Gray & Watson, 2002; O’Connor & Paunonen, 2007; Wolfe & Johnson, 1995). Our meta-analysis investigated the validity of Conscientiousness, but not achievement-striving or dutifulness, which are facets of Conscientiousness. The NEO-PI–R was the only measure examined in the current meta-analysis that includes reliable facet-level scales; therefore, we could not compare facets across measures. A meta-analysis comparing reliable facet-level personality measures that are available might show informative differences in their criterion-related validity for academic performance. Facet-level validity might explain what “drives” factor-level validity, or conversely, certain facet-level validities may demonstrate higher validity for predicting academic performance than the overall factor (see Nofle & Robins, 2007; O’Connor & Paunonen, 2007).

A fourth direction for future research is the need to examine more closely the item content of individual Big Five scales. Note that as a practical matter, it is useful to compare measures for their validity regardless of their overlap in content. Theoretically, however, convergence may be a function of specific content overlap in the items that are unrelated to personality traits just as much as it might be related to convergence in tapping underlying constructs that the items intend to measure. The issue of the trait-relevant and trait-irrelevant content overlap between measures—and the criterion-related validity of these sources of variances within individual personality items or scales—could be further explored in future comparative-validity studies.
As a final research direction, we echo previous researchers (e.g., Ackerman et al., 2011; Poropat, 2009) who call for the development of personality measures that are more sensitive to academic outcomes, not only traditional GPA or graduation criteria, but also critical criteria of student success found in college mission statements, such as multicultural appreciation, community involvement, ethics, and leadership (Oswald, Schmitt, Kim, Ramsay, & Gillespie, 2004). A longitudinal component to this type of research might also be important for measure development, as the relationship between personality traits and diverse academic outcomes may change over students’ tenure in college (e.g., Lievens et al., 2009; Schmitt et al., 2009). As personality and academic outcome measures improve, future research and meta-analyses on personality and academic outcomes will also improve as a consequence.

Concluding Comments

The current meta-analysis estimated fairly similar levels of validity across measures—and we claim that although this finding might feel both intuitive and reassuring, it is not guaranteed without empirical examination. In fact, this meta-analysis is among the first to heed the call of previous research to attend to the effect of differences in specific personality scales on the range and efficacy of their validity (e.g., Pace & Brannick, 2010).

References

References marked with an asterisk indicate studies included in the meta-analyses.


Komaraju, M., Karau, S. J., & Schmeck, R. R. (2009). Role of the Big Five personality traits in predicting college students’ academic motiva-
tion and achievement. Learning and Individual Differences, 19, 47–52. doi:10.1016/j.lindif.2008.07.001


Received May 22, 2012
Revision received October 31, 2012
Accepted January 2, 2013