Psychological Assessment

Comparative Validity of Brief to Medium-Length Big Five and Big Six Personality Questionnaires

Amber Gayle Thalmayer, Gerard Saucier, and Annemarie Eigenhuis

Online First Publication, August 22, 2011. doi: 10.1037/a0024165

CITATION

Thalmayer, A. G., Saucier, G., & Eigenhuis, A. (2011, August 22). Comparative Validity of Brief to Medium-Length Big Five and Big Six Personality Questionnaires. *Psychological Assessment*. Advance online publication. doi: 10.1037/a0024165

Psychological Assessment

1040-3590/11/$12.00 DOI: 10.1037/a0024165

Comparative Validity of Brief to Medium-Length Big Five and Big Six Personality Questionnaires

Amber Gayle Thalmayer and Gerard Saucier
University of Oregon

Annemarie Eigenhuis
University of Amsterdam

A general consensus on the Big Five model of personality attributes has been highly generative for the field of personality psychology. Many important psychological and life outcome correlates with Big Five trait dimensions have been established. But researchers must choose between multiple Big Five inventories when conducting a study and are faced with a variety of options as to inventory length. Furthermore, a 6-factor model has been proposed to extend and update the Big Five model, in part by adding a dimension of Honesty/Humility or Honesty/Propriety. In this study, 3 popular brief to medium-length Big Five measures (NEO Five Factor Inventory, Big Five Inventory [BFI], and International Personality Item Pool), and 3 six-factor measures (HEXACO Personality Inventory, Questionnaire Big Six Scales, and a 6-factor version of the BFI) were placed in competition to best predict important student life outcomes. The effect of test length was investigated by comparing brief versions of most measures (subsets of items) with original versions. Personality questionnaires were administered to undergraduate students (N = 227). Participants' college transcripts and student conduct records were obtained 6–9 months after data was collected. Six-factor inventories demonstrated better predictive ability for life outcomes than did some Big Five inventories. Additional behavioral observations made on participants, including their Facebook profiles and cell-phone text usage, were predicted similarly by Big Five and 6-factor measures. A brief version of the BFI performed surprisingly well; across inventory platforms, increasing test length had little effect on predictive validity. Comparative validity of the models and measures in terms of outcome prediction and parsimony is discussed.

Keywords: inventories, psychometrics, psychological assessment, test validity, five factor personality model

Supplemental materials: http://dx.doi.org/10.1037/a0024165.supp

The Big Five is a structural model of personality attributes that emerged from a variety of factor-analytic studies conducted with diverse temperament and personality scales and with lexical studies in the latter half of the twentieth century (Digman, 1996). The convergence of previous studies on five core factors became apparent to personality researchers in the 1980s (Digman, 1996; Goldberg, 1993). The contemporary prominence of the model is in part due to the development of the NEO personality inventory (NEO-PI-R) by Costa and McCrae (1989) and to lexical studies conducted in German (Ostendorf, 1990) and Dutch (De Raad, Hendriks, & Hofstee, 1992) that gave initial cross-cultural support to the Big Five model. Extraversion, Agreeableness, Conscientiousness, Emotional Stability (or Neuroticism), and Intellect (or Openness) are now widely accepted as five relatively independent factors that account for phenotypic personality variation between people. The common language for personality description provided by the Big Five has been highly generative for personality psychologists. Reliable measurement tools have been established to measure the five factors in self- and peer reports, and studies demonstrate the robustness of the model for many Western populations.

An abundance of ongoing research on personality structure, however, has also brought some weaknesses of the model to light. Whether a Big Five structure emerges in the factor analyses within indigenous lexical studies appears to depend on two principal constituents of method. One is variable-selection restrictions, established by Allport and Odbert (1936): Highly evaluative terms, temporary states (including many emotion descriptors), social roles, physical descriptors, and social attitudes or belief-dispositions were excluded from lists of person-descriptive traits by these authors. Many lexical studies have followed these variable-restriction practices quite closely, while others have not. One reason for the inconsistency is that such restrictions do not entirely conform to contemporary conceptions of how personality is defined (Saucier, 1997; Funder, 2007). Studies including a wider selection of variables have not produced good replications of the Big Five (e.g., Benet-Martinez & Waller, 1997; Church, Reyes,

Amber Gayle Thalmayer, Gerard Saucier, and Annemarie Eigenhuis, Department of Psychology, University of Oregon.

Annemarie Eigenhuis is now at the Department of Psychology, University of Amsterdam, Amsterdam, the Netherlands.

This study would not have been possible without the help of hardworking undergraduate research assistants, in particular Phil Getty and Jennifer Anne Beltz. Jason Isbell, Shanee Mali, Thomas Bradley, Lily Pon, Katrina Friedrich, and Caitlin Sweeney also provided many hours of assistance. Carl Yeh, Director of Student Conduct and Community Standards, provided crucial assistance by facilitating our retrieval of student conduct records used in this study.

Correspondence concerning this article should be addressed to Amber Gayle Thalmayer, Department of Psychology, 1227 University of Oregon, Eugene, OR 97403. E-mail: athalmay@uoregon.edu

Katigbak, & Grimm, 1997; Saucier, 1997, 2003, 2009; Saucier, Ole-Kotikash, & Payne, 2006).

Another aspect of method contributing to likelihood of a Big Five structure is linguistic/cultural setting. The Big Five rose to prominence after the structure was found in three closely related languages (English, German, and Dutch). McCrae and Costa (1997) translated the NEO-PI-R into half a dozen languages, found that items coalesced into similar factors in each language, and judged the model to be cross-culturally universal. But indigenous lexical studies in Italian (De Raad, DiBlas, & Perugini, 1998), Hungarian (Szirmak & De Raad, 1994), Greek (Saucier, Georgiades, Tsaousis, & Goldberg, 2005), and Chinese (Zhou, Saucier, Gao, & Liu, 2009) failed to find the Big Five in the five-factor solution where one would expect it.

Could the Big Five benefit from a 21st century upgrade to correct for these weaknesses? Could findings from lexical studies with broader selection criteria, and those collected in more diverse languages, be incorporated into the current model, without forcing researchers to start from scratch? A Big Six model, which is based on convergences among lexical study results when more factors and a wider selection of variables is allowed, would seem to fit the bill. In practice, this would primarily mean allowing a sixth factor to join the Big Five, with only minor adjustments to the content of other dimensions. Thus it would be straightforward to integrate previous findings based on Big Five questionnaires with new findings using Big Six inventories.

In order to justify an upgrade from the Big Five to the Big Six model, however, it is necessary to establish that such an upgrade leads to measurable improvement in validity. To yield evidence on this count, we arranged a comparative validity competition, or "race," including the most commonly used brief, short, and medium-length Big Five and Big Six inventories.

Aside from the "how many factors" question, the present race addresses the comparative validity of diverse Big Five inventories. With an abundance of Big Five inventories now available, there can be uncertainty among researchers about which to choose. Choice of inventory may be consequential for results, since inventories are not perfectly matched in the content-conceptualization of all factors. As Goldberg (1999) pointed out, the dearth of comparative validity tests, directly pitting popular instruments against one another, makes it hard for those who use such measures to make informed choices between them. This dearth also slows scientific progress in improving measures. Instead, published personality measures tend to remain static over long periods of time, and in some cases authors have a financial incentive to avoid direct comparison with other measures.

Finally, in addition to issues of model and inventory, this comparative validity study was designed to address the ongoing debate on optimal test length. We (as well as Rammstedt & John, 2007) observe that the predominant direction in the development of Big Five inventories over the past two decades has been toward increasingly briefer measures. Typically, use of abbreviated measures has been seen as a compromise, in which one maximizes convenience but gives up reliability (at least internal-consistency reliability) and, presumably, predictive validity. But it has also been argued that short measures might have advantages in terms of validity, for example by reducing respondent boredom and fatigue (Burisch, 1984). The design of the current study allowed us to compare brief (as few as 10 items), short (26–60 items), and medium-length (around 100 items) versions of popular inventories in terms of predicting important life outcomes.

Measures Associated With a Big Five Model

The NEO-Five Factor Inventory (NEO-FFI)

The NEO-FFI is a shorter (60-item) version of the 240-item Revised NEO Personality Inventory (NEO-PI-R), both developed by Costa and McCrae (1992). The first version of the measure, published in 1985, included scales measuring facets of Neuroticism, Extraversion, and Openness, but scales to measure facets of Agreeableness and Conscientiousness were added due to increasing scientific consensus on five trait factors. The shorter version of the measure is designed to measure five dimensions without the facet scores assessed in the long version (although Saucier, 1998, did isolate replicable subcomponents in the NEO-FFI). It includes the items most closely associated with the five factors in the longer version, which means that the facets are not equally represented in the short version. Agreeableness in the short version, for example, does not include items related to modesty and tender-mindedness, two facets in the longer version. Costa and McCrae (1989) stated that the NEO-FFI scales account for about 75% as much variance as the full-scale measures on convergent criteria (adjective self-reports from 3 years previous, and spouse and peer ratings). The NEO-FFI has been translated into many languages, and the translated items typically group into the Big Five dimensions (McCrae & Allik, 2002). The NEO-FFI is a popular measure with a wide base establishing the validity of its scores, but its status as a proprietary, commercial measure (with substantial per-use costs) inhibits some research uses.

The Big Five Inventory (BFI)

The 44-item BFI (John, Donahue, Kentle, 1991) was designed to be a short, efficient and noncommercial research measure of the Big Five. It includes short phrases, based on adjectives demonstrated to be prototypical for each of the five dimensions by expert ratings and factor analytic studies (John & Srivastava, 1999). This measure has been used frequently in research, and has been translated into at least eight languages (http://www.ocf.berkeley.edu/~johnlab/bfi.htm). In response to increasing demands for shorter measures, a 10-item version of the BFI was developed in German and English (Rammstedt & John, 2007) for research settings with severe time constraints. Because the 10 items are a subset of the 44-item BFI, the BFI-10 can be scored wherever the longer BFI has been administered.

An alternative six-factor version of the BFI was created for the purposes of this study. The second author created an Honesty/Propriety (H/P) scale with 10 International Personality Item Pool (IPIP; http://ipip.ori.org) items, based entirely on data from a community sample in which all IPIP items (as well as the BFI) were administered. Content for this dimension (degree of socially disapproved risk-taking, deceit, and instrumental use of others) matches content for a sixth factor as discussed below. A change in the content of the Agreeableness (A) dimension was necessary to reduce correlation between A and H/P. Thus the original set of BFI A items were replaced with IPIP items with content focused on patience, trust, forgiveness, and lack of anger, grudge-holding, and

Table 1
Agreeableness and Honesty/Propriety Items for the BFI and Alternative Six-Factor BFI

BFI Agreeableness	BFI-6 Agreeableness	BFI-6 Honesty/Propriety
Tend to find fault with others (R)	Am usually a patient person	Take risks that could cause trouble for me (R)
Start quarrels with others (R)	Trust what people say	Use others for my own ends (R)
Have a forgiving nature	Get angry easily (R)	Don't enjoy taking risks
Am considerate and kind to almost everyone	Get back at people who insult me (R)	Avoid activities that are physically dangerous
Can be cold and aloof (R)	Am inclined to forgive others	Use flattery to get ahead (R)
Am generally trusting	Hold grudges (R)	Have bad manners (R)
Am helpful and unselfish with others	Become frustrated and angry with people when they don't live up to my expectations (R)	Would never take things that aren't mine
Am sometimes rude to others (R)		Am not good at deceiving other people
Like to cooperate with others	Distrust people (R)	Misrepresent the facts (R)
		Stick to the rules

Note. BFI = Big Five Inventory; BFI-6 = six-factor BFI; (R) = reverse-scored items.

vindictiveness, content that tends to be more orthogonal to H/P content than is much of the original BFI A scale. Table 1 presents items for the original BFI A scale, the corresponding "A6" scale and the H/P scale added to the BFI in this study.

The International Personality Item Pool Big-Five Marker Scales (IPIP 50)

This inventory was developed using as benchmarks the Big Five as captured by 100 factor-marker adjectives presented by Goldberg (1992), but using short phrases that are intended to provide greater context than adjectives alone can convey (Goldberg, 1999). The inventory was developed as part of an international collaborative effort to develop broad-bandwidth, noncommercial measurement instruments which can be freely compared to other instruments, and refined over time (Goldberg, 1999). This inventory differs from the NEO and BFI in that the fifth factor is defined as Intellect/Imagination rather than Openness to Experience, and this scale includes several items referencing perceived abilities of a cognitive nature. IPIP Agreeableness also differs notably from NEO and BFI conceptualizations in its emphasis on empathy and interest in others, and lack of items referring to quarrelsomeness.

A briefer 20-item version of the IPIP Big Five was developed by Donnellan, Oswald, Baird, and Lucas (2006). The authors found convergent, discriminant, and criterion-related validity for scores on this measure comparable to that found with the longer version of the measure. Test–retest correlations were also comparable between brief and longer versions.

Big Six Measures

The HEXACO Personality Inventory (HEXACO-PI)

Ashton and colleagues (2004) integrated the factor structures of eight previous independent lexical studies. They noted that authors of lexical studies have tended to look for the Big Five in their results, and have reported results in terms of similarities to it. There are, however, consistencies in the divergences from the Big Five these diverse studies describe. In particular, a dimension with content related to ethical behavior (honesty, humility, and integrity vs. greed) often appears. Some of this content is occasionally included in dimensions of Conscientiousness or Agreeableness but more often is left out of the Big Five. (DeRaad et al., 2010, claimed that six factors [indeed, like five factors] are not replicable across languages in lexical studies. However, in a rebuttal, Ashton & Lee, 2010, showed that DeRaad et al. had handled some of the data sets in a way that suppressed the estimate for the average replicability of the six-factor models.)

Ashton et al.'s (2004) six-factor model also diverges from common Big Five conceptualizations in that anger and ill-temper are prominently referenced at the low end of Agreeableness (A) rather than being a peripheral part of Neuroticism; this conception of A is more orthogonal to the Honesty dimension than are some other conceptions of A. And unlike Big Five Neuroticism, the dimension of Emotionality includes a negative pole with content such as fearlessness and self-assurance, rather than being conceived of solely as an absence of negative emotions. Ashton and colleagues noted that the dimension of Intellect/Openness/Unconventionality was the least consistent dimension across the lexical studies reviewed, and its emergence depended strongly on the terms included in the study. However, this dimension appeared more consistently when six factors (rather than five) were extracted, giving the six-factor model another advantage in terms of cross-cultural replicability.

Lee and Ashton (2004; Ashton & Lee, 2007) developed the HEXACO-PI to operationalize this six-factor model. The questionnaire includes 24 facet scales that define the six factors: Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). Abbreviated versions have followed: first a half-length (96-item) version, and then a shorter, 60-item version of the measure developed by selecting 10 items for each dimension, including at least two from each of the facets. Items were selected based on high primary loadings and low secondary loadings and to maintain an even selection of forward- and reverse-scored items (Ashton & Lee, 2009).

The Questionnaire Big Six Scales (QB6)

Saucier (2009) compared five-, six- and seven-factor models from eight lexical studies that had used very broad variable selection criteria, and concluded that a Big Six model improves on the Big Five in terms of cross-cultural replicability and that, moreover, adjective Big Six scales predict important criteria (many related to psychopathology) better than do Big Five scales. The Big Six dimensions are termed Conscientiousness, Honesty/Propriety, Agreeableness (Kindness & Even Temper), Resiliency versus Internalizing Negative Emotionality, Extraversion (Gregariousness and Positive Emotionality), and Originality/Talent.

Big Six dimensions are close enough to HEXACO dimensions that it is fair to consider them variants of a single model (Saucier, 2009). As in the HEXACO, Big Six Agreeableness contains some content found in Big Five Neuroticism (aggressiveness and irritability). With hostility and irritability relocated to A, and a better defined favorable pole, Neuroticism is redefined as Resiliency versus Internalizing Negative Emotionality. This arrangement allows Big Six scales to better map onto temperament dimensions of Neuroticism versus Resiliency, Extraversion/Positive Emotionality, and Disinhibition (a combination of low Agreeableness and Conscientiousness; alternatively conceived of as "Constraint" in Tellegen and Waller's, 2008, Multidimensional Personality Questionnaire). These dimensions of temperament are theorized to precede both mental disorders and personality traits expressed in adulthood (Clark, 2005). Drawing from broader lexical studies means that the questionnaire measures of the Big Six (Questionnaire Big Six or QB6), particularly in longer versions, include a fuller representation of internalizing affect (depression, anxiety, tendencies toward panic and phobias) than other measures discussed in this article; this fuller representation should allow instruments based on this model to better correlate with and predict mental disorders.

The "negative valence" content typically found in inclusive-variable-selection lexical studies is included in the Honesty/Propriety factor (which was shown to correlate with externalizing disorder tendencies; Saucier, 2009). Originality/Talent encompasses perceived talents and abilities, originality, and intellectual and aesthetic interests, as well as some "positive valence" terms typically found in broader variable selection studies, although these are generally excluded in narrower selection lexical studies.

In developing QB6 scales, an initial pool of 120 promising IPIP items was selected to measure the six factors described in Saucier (2009), from data collected over a decade-long period from the Eugene-Springfield Community Sample. An optimal subset of 36 items (the 36QB6) was selected based on a second single-wave administration of the 120 items to the same sample. A slightly longer subset of 48 items (the 48QB6), including all of the first 36, was then selected. In order to arrive at the shortest version of the inventory, while simultaneously examining the effects of item validity as a criterion in scale construction, a 24QB6 was also created. The 24QB6 consists of the better half of the 48QB6 items selected based not only on criteria used in selection of items for other QB6 inventories (fidelity to Big Six factors from Saucier, 2009, internal consistency, unidimensionality, and having as much as possible equal numbers of forward- and reverse-keyed items) but also mean correlation with several dozen criterion variables in the Eugene-Springfield sample. This "better half" 24QB6 can be compared with the "other half" 24-item subset of the 48QB6 to examine effects on the scores' predictive validity in new samples as an item-selection criterion in the derivation sample. Finally, based on the same second administration, progressively longer subsets of 60 and 96 items were selected for the purposes of this study, to enable comparison with longer Big Five measures.

A Comparative Validity Competition

Grucza and Goldberg (2007) compared nine personality inventories administered to the Eugene Springfield Community Sample between 1993 and 2000 in terms of predictive ability for frequency of behavioral acts (drug and alcohol use, undependability, friendliness, erudition, communication and creativity), correspondence with peer reports, and relation to measures of mental disorders. The measures compared included Big Five measures (NEO-PI-R, 100 unipolar adjective markers of the Big-Five factor structure developed by Goldberg, 1992, and the 485 items of the IPIP-AB5C Inventory), one Big Six measure (the 180-item version of the HEXACO), and various non-Big Five inventories (Six Factor Personality Questionnaire [6FPQ], California Psychological Inventory [CPI], Cattell's Sixteen Personality Factor Questionnaire, Hogan Personality Inventory, Temperament and Character Inventory, Multidimensional Personality Questionnaire, Jackson Personality Inventory). Overall the inventories were similar in their correlation with the domains of interest (with the possible exception of the 6FPQ and the CPI, which trailed the field in this race).

Johnson (2000) compared the CPI, the Hogan Personality Inventory, and the NEO-PI-R in terms of their ability to predict acquaintance ratings from four peers. Peers rated targets on four variants of the Five-Factor Model. The NEO-PI-R surpassed its competitors, but (as with a similar comparison on observer ratings in the Grucza and Goldberg, 2007, study) this does not seem to be a fair comparison: It is only logical that if the criterion is ratings on the Big Five, a Big Five measure would outperform measures from different models in predicting those ratings.

While these studies have provided some initial comparison between measures, they do not include the comparisons we think would best aid contemporary psychologists in choosing an instrument for research studies. Grucza and Goldberg (2007) only included very long measures, ranging from 100 to 485 items. Johnson (2000) also used long measures, and did not include life-outcome or behavioral criteria, but only peer-ratings on a Big Five instrument.

The current study seeks to compare brief, short, and medium-length Big Five and Big Six self-report measures in terms of their ability to predict important student outcomes later in the academic year. These criteria, grades and student conduct violations, fall into the category of "L data," or life outcomes (Cattell, 1957), a criterion domain that should be of paramount interest to personality psychologists. Student-conduct-code violations were chosen as a novel way to look at broad rule-breaking versus rule-following behavior, somewhat analogous to an arrest or criminal record, but with a larger base-rate. While previous studies have demonstrated that Big Five and other personality trait measures correlate with retrospective delinquent acts or arrest records (Alalehto, 2003; Clower & Bothwell, 2001; van Dam, Janssens & De Bruyn, 2005), none, to our knowledge, use traits to predict infractions arising after the assessment. The demonstration of a capability to forecast important life events at a later date might be the gold standard for the validity of scores on an inventory.

The current study also looked at the trait measures' ability to predict future grade point average (GPA), as well to correlate with current GPA. GPA is an excellent summary of academic achievement, given the sustained levels of performance over time and across domains and raters required to obtain high values (Hirsh &

Peterson, 2008). Noftle and Robins (2007) summarized 20 previous studies, most of which found significant correlations between Big Five (or readily comparable) personality trait inventories and college GPA (an exception was the IPIP-50, which did not correlate with GPA). In a series of studies, the authors compared college students' NEO-FFI, HEXACO-PI, and BFI scores with SAT scores, high school GPA, and, in one case, college GPA. One study, of a 4-year longitudinal design, found predictive ability for all three inventories for college GPA (after accounting for SAT scores and gender).

Secondarily, the current study used behavioral observations ("B data"; Funder, 2007), such as punctuality, use of Facebook, and text messaging as outcomes. These variables are intended to be illustrative (but in no way comprehensive) of broad behavior patterns that can be assessed with personality inventories. Of paramount importance in this study was to use outcome criteria that did not rely on self-report data ("S data"; Funder, 2007), in order to avoid shared error (from sources such as acquiescence bias) that can not be ruled out as an alternative explanation for shared variance in data collected using the same methodology.

Effects of Model

This study was designed as a comparative-validity competition, or "race," to compare Big Five and Big Six inventories in terms of correlation with and prediction of important student life outcomes. The Big Six model, based on convergences among lexical study results when more factors and a wider selection of variables is allowed, is similar in many ways to the Big Five model, but it differs in ways that extend beyond the simple addition of a sixth dimension, as described above. A comparative-validity race allows for a comparison of sets of dimensions, which differ in their conceptualizations, in terms of their power to predict outcomes in the real world.

Furthermore, Big Five inventories also differ from each other in their conceptualization of the different dimensions, (e.g., NEO-FFI Openness to Experience vs. IPIP Intellect), and these differences may lead to real differences in the performance of inventories. This study will allow us to gather data relevant to the comparative-validity impact of the different conceptualizations.

Effects of Test Length

This study also sought to explore the effect of test length on the predictive validity of scores. In general, more items would be expected to increase score reliability (internal consistency) and reduce measurement error and therefore, by conventional expectations, should predict outcomes more effectively. A possible limiting factor would be an "attenuation paradox" (Loevinger, 1954), whereby increasing reliability may sometimes decrease validity. Loevinger's paradox points to the need for testing the assumption that scores on longer, more reliable measures have more predictive validity. A general trend toward shorter trait measures is evident in the literature (Rammstedt & John, 2007), and by comparing different versions of the same instrument, we can assess costs and benefits in the trade-off between efficiency and comprehensiveness.

Method

Participants

Undergraduate students in introductory psychology and linguistics courses were recruited using the psychology department's human subjects pool (N = 227, 65% female). The majority (62%) were freshman. Age ranged from 17 to 35 years, with a mean of 19.24 (SD = 2.09). The sample was predominantly White (72%; 11% were Asian or Asian American, 4% Hispanic, 2% Black, 1% American Indian, 4% "more than one," and 5% "other"). Some 54% reported that their mother had finished at least a 4-year college degree, and 60% reported that their father had at least a 4-year college degree.

Measures

NEO-FFI. The 60-item NEO-FFI was administered in its published test booklet. Costa and McCrae (1992) reported average scale-score reliabilities of .78 on this inventory. Mean score, standard deviation, Cronbach's alpha, and the mean and variance of interitem correlations for each scale of each measure in the current sample are provided in Table 2. (The variance of interitem r is included, since it tends to be an indicator of unidimensionality—if this variance is zero, a factor analysis will only find one nonsinglet factor. Ideally, correlations between items in measuring a single construct should range from .15 to .50 [Clark & Watson, 1995], but a large range in correlations within a scale will indicate multidimensionality.)

BFI. We used the standard 44-item BFI. For the extension into a six-factor model, the item set was augmented with 18 IPIP items, eight of which replaced the nine original Agreeableness items and 10 of which formed an Honesty/Propriety factor—thus the BFI-6 has 53 items. The 10-item version was derived as a subset of the original 44 items, following Rammstedt and John (2007). In general, we did not replace or impute missing data in this study, because there was very little of it, but one of the BFI-44 (not BFI-10) items, "Am helpful and unselfish with others," had an unusually high number (20) of missing data points; we replaced missing responses, on this item only, with the middle option on the response scale.

IPIP. Goldberg (http://ipip.ori.org/newBigFive5broadTable.htm) reported alpha values of .79–.87 for scores on each of the five scales in the 50-item version. Donnellan et al. (2006) reported alpha values of .65–.77 for scores on the five scales of the 20-item version (N = 2,663 college freshmen).

HEXACO-PI. Lee and Ashton (2004) reported internal consistency values ranging from .81–.84 for the six scales in the 96-item version (N = 1,126 college students; the authors refer to this as a 100-item version, because it also includes four items "interstitial" between A and H which do not belong to one of the Big Six scales and were not included in the current study; descriptive statistics are available at http://www.hexaco.org/). For the 60-item version, internal consistency ranged from .73 to .80 across two samples (N = 734 community members, N = 936 college students; Ashton & Lee, 2009).

QB6. Twenty-four-, 36-, 48-, 60-, and 96-item versions of the QB6 were used. Each progressively longer version contains all of the items in the shorter versions, with the exception that three

Table 2
Descriptive Statistics for Questionnaire Scales

Scale	*M*	*SD*	α	*M* interitem *r*	Variance interitem *r*
NEO-FFI					
Neuroticism	2.03	.48	.85	.31	.009
Extraversion	2.71	.41	.83	.29	.013
Openness	2.64	.41	.78	.23	.025
Agreeableness	2.83	.37	.77	.22	.010
Conscientiousness	2.65	.42	.84	.30	.018
IPIP-20					
Extraversion	3.31	.89	.83	.56	.006
Agreeableness	4.07	.61	.72	.39	.008
Conscientiousness	3.23	.79	.72	.39	.015
Neuroticism	3.49	.72	.62	.29	.011
Intellect	3.77	.73	.72	.40	.008
IPIP-50					
Extraversion	3.35	.77	.90	.48	.009
Agreeableness	4.06	.54	.84	.35	.009
Conscientiousness	3.31	.60	.81	.29	.015
Stability	3.32	.68	.84	.34	.016
Openness	3.62	.56	.80	.29	.020
BFI-10					
Extraversion	3.39	.94	.72	.56	
Agreeableness	3.63	.78	.43	.28	
Conscientiousness	3.46	.80	.54	.34	
Stability	2.90	.98	.60	.42	
Intellect	3.66	.91	.51	.37	
BFI					
Extraversion	3.38	.81	.90	.51	.012
Agreeableness	3.82	.55	.77	.28	.008
Conscientiousness	3.46	.60	.79	.31	.011
Neuroticism	2.82	.70	.81	.35	.018
Openness	3.62	.62	.82	.33	.011
BFI 6 factor additions					
Agreeableness-6	3.54	.64	.76	.28	.009
Honesty	3.35	.55	.71	.19	.017
HEXACO-60					
Honesty	3.36	.64	.74	.22	.013
E. Stability	3.27	.65	.77	.25	.010
Extraversion	3.54	.75	.74	.29	.012
Agreeableness	3.28	.61	.77	.25	.012
Conscientiousness	3.40	.62	.78	.26	.011
Openness	3.49	.73	.82	.31	.017
HEXACO-96					
Honesty	3.35	.59	.81	.21	.018
E. Stability	3.35	.59	.82	.23	.012
Extraversion	3.55	.66	.83	.29	.015
Agreeableness	3.11	.57	.83	.24	.013
Conscientiousness	3.41	.59	.84	.25	.015
Openness	3.42	.64	.84	.24	.019
24QB6					
Conscientiousness	3.03	.80	.67	.35	.007
Honesty/Propriety	3.32	.72	.55	.24	.008
Agreeableness	3.34	.80	.68	.35	.012
Resiliency	3.36	.73	.62	.29	.007
Extraversion	4.07	.58	.54	.25	.017
Originality/Talent	3.40	.63	.58	.25	.004
36QB6					
Conscientiousness	3.20	.67	.70	.28	.018
Honesty/Propriety	3.38	.73	.70	.29	.017
Agreeableness	3.37	.69	.70	.28	.011
Resiliency	3.10	.78	.79	.39	.010
Extraversion	3.98	.57	.60	.22	.017
Originality/Talent	3.39	.55	.59	.20	.009
48QB6					
Conscientiousness	3.34	.62	.75	.27	.015
Honesty/Propriety	3.43	.67	.75	.28	.012
Agreeableness	3.37	.64	.73	.25	.010
Resiliency	3.32	.71	.81	.34	.010
Extraversion	3.90	.54	.69	.23	.017
Originality/Talent	3.57	.52	.65	.19	.010
60QB6					
Conscientiousness	3.35	.60	.79	.27	.012
Honesty/Propriety	3.43	.63	.77	.25	.011
Agreeableness	3.29	.58	.74	.22	.009
Resiliency	3.30	.63	.80	.28	.023
Extraversion	3.75	.57	.75	.24	.021
Originality/Talent	3.56	.50	.69	.19	.009
96QB6					
Conscientiousness	3.31	.57	.83	.24	.014
Honesty/Propriety	3.59	.54	.82	.22	.010
Agreeableness	3.22	.47	.79	.19	.011
Resiliency	3.38	.58	.85	.26	.024
Extraversion	3.69	.55	.82	.23	.018
Originality/Talent	3.46	.51	.79	.20	.012

Note. NEO-FFI = NEO Five Factor Inventory; IPIP = International Personality Item Pool; BFI = Big Five Inventory; HEXACO = HEXACO Personality Inventory; QB6 = Questionnaire Big Six Scales.

of the 24QB6 items are not in 36QB6 (the 24QB6 being derived from the 48QB6).

Procedure

Participants came into our lab during fall and winter terms to complete self-report questionnaires either in paper-and-pencil form (first half of participants) or electronically on a desktop computer (second half of the participants). Punctuality was recorded in terms of minutes arrived before versus after the scheduled participation time. After completing questionnaires and a short wait in a waiting room, participants were led to a third room for an interview, conducted by either the first or third author. A research assistant was present in the room as an observer. After being reminded of confidentiality guarantees, participants were asked whether they had a Facebook account and, if so, if they would show it to the interviewer and observer on a laptop in the room. Participants were also asked if they had a cell phone with them and, if so, whether they would show us the time of their last sent and last received text messages (text messages themselves were not viewed). Finally, written permission for access to the participants' transcripts and student conduct records at the end of the school year was requested.

Criterion Variables

Grades. Of the 227 participants, 217 (96%) filled out a transcript request form, giving us permission to access their academic records at the end of summer term. In order to assess the predictive validity of inventory scale-scores as forecasting tools, we looked at GPA for terms after the term of participation (Fall and Winter quarter of 2008–2009 academic year), in addition to GPA overall. Transcripts were obtained at the end of summer term, and post-study GPA was calculated for the one to three terms of coursework on record for each student after and not including the term of participation in the study. Descriptive statistics for criterion variables are included in Table 3.

Student conduct records. Two hundred eighteen participants (96%) gave us permission to access their student conduct records (these were the same participants who agreed to let us access their transcript, excepting one student who did not agree to the transcript request but did agree to the student conduct record request). Records were obtained from the Office of the Dean of Students in August, 5 to 10 months after participants' questionnaires were completed. Because violations were more likely to exist for those students who were currently living in student housing (55% of participants) or had lived in student housing at some previous point (81%), regressions accounted for months spent in university housing. Some of the less frequent types of violations, however (academic dishonesty, behavior on campus and at sporting events), could occur for any enrolled student.

The Office of the Dean of Students provided us with information about complaints involving our participants and, for each complaint, whether or not the student was found responsible for the charges. In many cases, a single incident led to more than one complaint or "charge," for example, a student might have been charged with both underage alcohol possession and disorderly conduct after a single run-in with campus housing or security officers. For our analyses, we looked at both the number of events and the number of charges for which participants were found responsible (analogous to a "conviction"). Eighty-six students (39%) had experienced at least one (and as many as eight) incident(s) that led to one or more complaints being lodged against them. For 55 of these students (25%), complaints resulted in at least one (and up to 18) charge(s) of responsibility. Because the distribution of both variables was very positively skewed (skew = 2.81 and 4.30, $SE = .16$), log10 ($y + 1$) transformations were performed, which reduced the skew to 1.18 and 1.74, respectively.

Table 3
Descriptive Statistics for Criterion Variables

Criterion	N	M	SD
Academic performance			
GPA overall	218	2.95	0.59
GPA post study	212	2.93	0.71
Student conduct records			
Incidents total	219	0.75	1.32
log10 ($y + 1$) transformed		0.17	0.23
Responsible total	219	0.90	2.24
log10 ($y + 1$) transformed		0.15	0.28
Incidents forecast	219	0.25	0.66
log10 ($y + 1$) transformed		0.06	0.15
Responsible forecast	219	0.08	0.50
log10 ($y + 1$) transformed		0.02	0.10
Observed behavioral indices			
Punctuality	227	5.13	7.32
Facebook contacts	205	339.95	215.72
People in photo (median)	205	2	1.13
Minutes since text received	220	620.44	4,934.41
log10 ($y + 1$) transformed		1.72	0.89
Minutes since text sent	219	353.9	1,453.22
log10 ($y + 1$) transformed		1.74	0.83

Note. GPA = grade point average.

We assessed forecasting power by looking at violation complaints reported to the Office of the Dean of Students after participation in our study. Between one and four incidents were reported after participation in the study for 37 students in our sample. This variable also had an extreme positive skew (skew = 3.10), reduced by a log10 ($y + 1$) transformation (transformed skew = 2.28). Seven individuals were found responsible for the complaints received after participation (skew = 7.05, $SE = .16$; after log10 [$y + 1$] transformation skew = 5.91).[1]

Behavioral observations. Personality attributes reflect broad patterning in behavior. Measures of these attributes should predict specific behavioral instances, to a degree. We tracked several specific behavioral indicators that can be easily collected in any laboratory setting.

Punctuality was calculated by subtracting the actual arrival time from the time students had scheduled themselves to participate. Positive values indicate an early arrival and negative values a late arrival. The average arrival was about 5 min prior to appointment time.

Of the 226 participants interviewed (due to a late appointment, one participant was not interviewed) 214 (95%) acknowledged having a Facebook account. Possession of a Facebook account was negatively correlated with age ($r = –.20$, $p < .01$) but was not predicted by scores on the personality measures. Of the 214 participants with a Facebook account, 205 (96%) agreed to log into their account and show us their profile page. Contacts (noted by the interviewer) ranged from 18 to 1,086, with a mean of 346.1 ($SD = 214.32$). The researchers noted the number of people in the profile picture, on an ordinal scale (0 = *not a picture of the participant*; 1 = *only the participant*; 2 = *participant with one other person*; 3 = *participant with 2–3 others*; 5 = *a group larger than four*). The median score was 2 ($SD = 1.13$); the most common scores were 1 (80 participants) and 2 (75 participants).

Two hundred twenty-two participants (98%) had a cell phone with which they had ever sent or received a text message. One participant had lost his phone, and three participants reported either no phone or no use of texting. All participants with a phone agreed to get it out to determine the exact time of their most recent

[1] We also looked at specific charges/complaints (the base rate of "convictions" was too small). The most common charge was underage alcohol possession: 60 students (27%) had one or more alcohol related incidents on their records. Other charges included disorderly conduct (29 students had one or more charge), noise (25), and marijuana use (20). Specific prediction of safety violations (e.g., candles in a dorm room; 12), property damage (6), academic dishonesty (2), other drug use (1), and interpersonal issues (1) were not analyzed due to low base rates.

messages received and sent (in four cases, the cell phone was not in the room and participants self-reported the time of their last messages; in one case, the participant had received a text once but had never sent one). The latency since a message was sent or received was recorded in minutes. Possession of a cell phone was not correlated with age, but age was correlated with latency since a text was sent ($r = .29$, $p < .001$) and since a text was received ($r = .61$, $p < .001$). In six cases, a message was received within 1 min of asking, and in one case, a message was sent while the question was asked. In a few cases, it had been weeks since a message was received, such that these variables were extremely positively skewed (skew for sent messages = 11.91; skew for received messages = 12.92; *SE* = .164). A log 10 ($y + 1$) transformation was performed on these variables before running the regressions to reduce the effect of outliers (transformed variables skew = .35 and .28, respectively).

Analyses. For each criterion variable, age and gender were entered at Step 1 (months spent in university housing was also included in this step for the student conduct variables), and the set of scales making up an inventory were entered at Step 2. Thus, prediction is at the level of scores on each inventory (in part to minimize Type I error) in Table 4. Results are detailed in terms of specific scales within each inventory in Table 5.

Results

Prediction of Grades

The personality inventories significantly correlated with overall GPA and significantly predicted post-study GPA, with the exception of the two IPIP inventories. Table 4 provides *R*-change coefficients. Table 5 reproduces this table with indications as to specific scales within each inventory that provided significant prediction of the criterion. Figure 1 displays differences across the inventories, grouping questionnaires by their length. In each figure, the four leftmost positions on the horizontal axis represent the *L*-data outcomes (which are arguably more important than the behavioral criteria on the right).

As Table 5 demonstrates, for the Big Five inventories, Conscientiousness is the major predictor of GPA. But among the six-factor inventories, Honesty and Originality/Talent scales come forward as significant predictors.

Student Conduct Records

Regressions on student conduct data accounted for months lived in university housing (months lived in UO housing and number of conduct violations reported $r = .21$, $p < .01$), age, and gender. Overall complaints and charges for which students were found responsible were significantly correlated with most of the inventories (excepting the BFI-6 for total complaints and the BFI-44 and IPIP-50 for convictions), with the highest coefficients found for the Big Six inventories. For incidents/complaints that occurred after participation, *R* values for the inventories' set of scale-scores were in the .20 to .29 range, with about half (not including any QB6 versions) obtaining statistical significance. None of the inventories significantly predicted responsibility for charges resulting from incidents after study participation (which occurred at a lower base-rate).

Note that due to redundancy in the student conduct variables (charges for which students were found responsible overall and after the study are nested within complaints overall and after the study), only two are included in figures. The figures include the number of charges overall, for which participants were found

Table 4
R Change for Criterion Variables After Accounting for Age and Gender

	Academic performance		Student conduct records variables[a]				Behavioral observations				
Scale	GPA	GPA post study	Complaints total	Responsible total	Complaints forecast	Responsible forecast	Punctuality	Facebook contacts	People in photo	Min. text rcvd	Min. text sent
NEO FFI	.28**	.27*	.27*	.27*	.26*	.11	.22	.38**	.31**	.26**	.24*
IPIP-20	*.20*	*.20*	.30**	.24*	.28**	.09	.18	**.39****	.20	.29**	.30**
IPIP-50	*.18*	*.22*	.31**	*.22*	.28**	.11	.22	.35**	.19	.28**	.28**
BFI-10	**.35****	.30**	.25*	.24*	.29**	.17	.28**	.35**	.19	.24*	.26*
BFI-44	.25*	.24*	.25*	*.22*	.26*	.13	.27**	.38**	.26*	.23*	.25*
BFI-6	**.35****	.35**	.22	.27*	.24	.15	.28**	**.43****	.28*	.27**	.27*
HEXACO-60	**.36****	.35**	.28*	.33**	.25*	.09	.23	.36**	.20	.33**	.25*
HEXACO-96	.33**	.33**	.27*	.32**	.27*	.12	.21	.37**	.22	.34**	.27*
24QB6	**.38****	**.42****	.27*	.32**	.20	.18	.19	.27*	.22	.21	.21
36QB6	.33**	**.37****	.32**	**.35****	.22	.18	.19	.28*	.25*	.24*	.23
48QB6	**.35****	**.39****	.31**	**.35****	.23	.15	.23	.30**	.31**	.24*	.23
60QB6	.34**	**.38****	.26*	.31**	.23	.14	.21	*.26**	.29**	.23	.24
96QB6	**.37****	**.37****	.32**	**.36****	.25	.15	.23	.31**	.27*	.28**	.29**
Range of *N*	205–210	198–204	192–197	192–197	192–197	192–197	211–219	192–199	192–199	204–212	203–211

Note. GPA = grade point average; Min. text rcvd = minutes since last text message received; Min. text sent = minutes since last text message sent; NEO-FFI = NEO Five Factor Inventory; IPIP = International Personality Item Pool; BFI = Big Five Inventory; HEXACO = HEXACO Personality Inventory; QB6 = Questionnaire Big Six Scales. For a column, any bold coefficients are each substantially larger (difference in R^2 of at least .07) than any italicized coefficients.

[a] Also accounts for months lived in student housing.

* $\alpha < .05$. ** $\alpha < .01$.

Table 5
R Change for Criterion Variables After Accounting for Age and Gender, Scales With Significant Coefficients Noted

	Academic performance		Student conduct records variables[a]				Behavioral observations				
Scale	GPA	GPA post study	Complaint total	Responsible total	Complaints forecast	Responsible forecast	Punctuality	Facebook contacts	People in photo	Min. text rcvd	Min. text sent
NEO-FFI	$.28^{C}$	$.27^{C}$	$.27^{E}$	$.27^{\mathbf{E}C}$	$.26^{A}$	.11	$.22^{-\mathbf{E}}$	$.38^{E}$	$.31^{\mathbf{E}-O}$	$.26^{-\mathbf{E}}$	$.24^{-\mathbf{E}}$
IPIP-20	$.20^{S}$	.20	$.30^{EA}$	$.24^{E}$	.28	.09	$.18^{-E}$	$.39^{E}$	$.20^{E}$	$.29^{-E}$	$.30^{-E}$
IPIP-50	.18	.22	$.31^{A}$	$.22^{-C}$	$.28^{A-C}$	.11	$.22^{-\mathbf{E}}$	$.35^{\mathbf{E}-C-I}$	.19	$.28^{-\mathbf{E}}$	$.28^{-\mathbf{E}}$
BFI-10	$.35^{C}$	$.30^{C}$	$.25^{E}$	$.24^{E-C}$	$.29^{A-CO}$	$.17^{-C}$	$.28^{\mathbf{C}-E}$	$.35^{\mathbf{E}A}$	$.19^{-O}$	$.24^{-\mathbf{E}}$	$.26^{-\mathbf{E}}$
BFI-44	$.25^{C}$	$.24^{C}$	$.25^{EA}$	$.22^{E-C}$	.26	.13	$.27^{-\mathbf{EC}}$	$.38^{\mathbf{E}-C-O}$	$.26^{-\mathbf{O}E}$	$.23^{-\mathbf{E}}$	$.25^{-\mathbf{E}}$
BFI-6	$.35^{\mathbf{H}C}$	$.35^{CHA}$	$.22^{E}$	$.27^{-H}$	.24	.15	$.28^{-\mathbf{E}C}$	$.43^{\mathbf{E}-H-O}$	$.28^{-\mathbf{O}E}$	$.27^{-\mathbf{E}}$	$.27^{-\mathbf{E}}$
HEXACO-60	$.36^{CH}$	$.35^{\mathbf{CH}}$	$.28^{-\mathbf{C}E}$	$.33^{-\mathbf{C}}$	$.25^{-C}$	.09	$.23^{\mathbf{C}-E}$	$.36^{\mathbf{E}-O}$	.20	$.33^{-\mathbf{E}}$	$.25^{-\mathbf{E}}$
HEXACO-96	$.33^{\mathbf{H}C}$	$.33^{\mathbf{CH}}$	$.27^{-\mathbf{C}E}$	$.32^{-\mathbf{C}E}$	$.27^{-CE}$	.12	$.21^{C-E}$	$.37^{\mathbf{E}-O}$	$.22^{E}$	$.34^{-\mathbf{EH}}$	$.27^{-\mathbf{E}}$
24QB6	$.38^{\mathbf{HO}}$	$.42^{\mathbf{HAO}}$	$.27^{E}$	$.32^{E-O}$	.20	.18	.19	$.27^{-H-O}$	.22	$.21^{H}$	.21
36QB6	$.33^{\mathbf{HO}}$	$.37^{\mathbf{HO}}$	$.32^{E}$	$.35^{\mathbf{E}-C-A-O}$	$.22^{E}$	.18	.19	$.28^{E-O}$	$.25^{E}$	$.24^{-EH}$	$.23^{-E}$
48QB6	$.35^{\mathbf{HO}}$	$.39^{\mathbf{HO}A}$	$.31^{E}$	$.35^{\mathbf{E}-C-O}$	$.23^{E}$	.15	.23	$.30^{\mathbf{E}-O}$	$.31^{\mathbf{E-O}}$	$.24^{-E}$	$.23^{-E}$
60QB6	$.34^{\mathbf{HO}}$	$.38^{\mathbf{HO}A}$	$.26^{E}$	$.31^{\mathbf{E}-C}$	.23	.14	.21	.26	.29	.23	.24
96QB6	$.37^{\mathbf{HO}C}$	$.37^{\mathbf{O}H}$	$.32^{-\mathbf{CE}}$	$.36^{-\mathbf{C}E-O}$	$.25^{-\mathbf{C}}$	.15	.23	$.31^{E-O}$	$.27^{-\mathbf{O}E}$	$.28^{-\mathbf{E}}$	$.29^{-\mathbf{E}}$

Note. GPA = grade point average; Min. text rcvd = minutes since last text message received; Min. text sent = minutes since last text message sent; NEO-FFI = NEO Five Factor Inventory; IPIP = International Personality Item Pool; BFI = Big Five Inventory; HEXACO = HEXACO Personality Inventory; QB6 = Questionnaire Big Six Scales. Superscripts ([C]onscientiousness; [H] Honesty or Honesty/Propriety; [A]greeableness; [E]xtraversion, even on HEXACO; [N]euroticism, [S]tability or [R]esiliency; [O]penness or Originality/Talent or [I]ntellect) are noted for scales with β significant at $\alpha < .05$; superscripts are in bold where $\alpha < .01$. A minus (−) indicates that the scale noted after the minus had a negative direction of effect.
[a] Also accounts for months lived in student housing.

responsible, and the number of complaints lodged against participants after participation in study.

As Table 5 demonstrates, high Extraversion (notated, for consistency, as E, even for the HEXACO, where E typically means Emotionality and X Extraversion—the Emotionality scale significantly predicted no criterion) was the most frequent predictor of conduct-code violations across inventories. Low Conscientiousness was also a frequent predictor.

Behavior Observations

Punctuality was significantly predicted only by scores on the three versions of the BFI (although these values were not substantially different from the slightly lower *R*-change values found for the remaining inventories). All the inventories significantly predicted the number of contacts on Facebook, and most predicted the number of people in the Facebook profile photo. Most predicted latency since last text message received and sent, although the QB6 versions appeared weaker on this criterion.

Table 5 documents that Extraversion scores were far and away the best predictors of observed behaviors. This is unsurprising: Most of the behaviors reflect sociability. For the one criterion not related to sociability—punctuality—Extraversion scores were typically negative predictors.

In Figure 1, one can see that for life outcome variables, where a QB6 inventory is pitted against a Big Five inventory of comparable length, the set of scores on the QB6 inventory tends to show a higher R value. However, QB6 inventories show no such advantage on observed-behavior criteria.

Comparisons Across Inventories (Effects of Model)

How much of a difference in *R* values can be considered substantial? One conservative way of comparing setwise *R* values for two inventories is to place them in hierarchical regression with the lesser *R* inventory's scales entered in an earlier block and the greater *R* inventory's scales entered in a later block. In such a framework, a statistically significant difference occurs when the *R*-change for the later block is significant. There are 858 possible comparisons of two inventories on one validity criterion for the coefficients in Table 4, too many to report. However, we did find that for the most unfavorable value for degrees of freedom that would be found in hierarchical-regression analyses for inventories in this study (i.e., 6, 183), two *R* values, after being squared, would have to differ by at least .066 for the later block to show a significant ($p < .05$) boost in prediction. Therefore, we set an R^2 difference of .07 as an effect-size threshold for "a substantial difference in *R*." Observing this threshold, Table 4 indicates that for four of the criteria the highest-*R* inventories did generate a substantially larger *R* than a group of lowest-*R* inventories. For these four criteria in particular, one cannot conclude that all inventories performed roughly equally well. Where highest- and lowest-*R* inventories were substantially different in *R*, most (over 80%) of the highest-*R* inventories on a given criterion represented a six-factor model, while most (over 80%) of the lowest-*R* inventories were IPIP Big Five inventories. Since the QB6 inventories use IPIP items, the difference in prediction appears to be attributable to the model (Big Six versus Big Five) or the scale-construction strategy, not the item format.

Table 6 provides a tally of the comparisons made in Table 4. We assigned a point for "better" if it demonstrated statistically significantly higher *R*-change value over any other inventory for an outcome in Table 4. A point for "worse" was tallied if, for any outcome in Table 4, this inventory had a significantly smaller *R*-change value than any other inventory. Inventories were ordered by their provisional rank on the resulting indices. This demon-

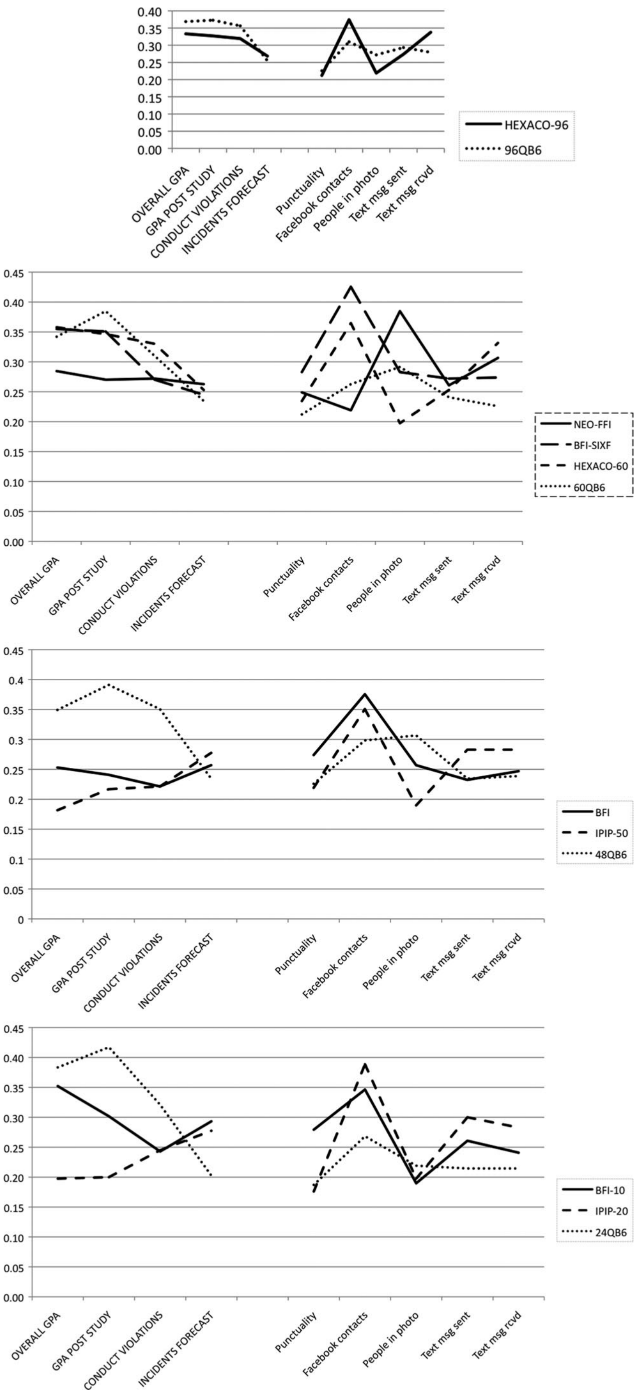

Figure 1. Predictive ability (*R* change) by length of measure for each questionnaire type. GPA = grade point average; HEXACO = HEXACO Personality Inventory; BFI = Big Five Inventory; NEO-FFI = NEO Five Factory Inventory; QB6 = Questionnaire Big Six Scales; IPIP = International Personality Item Pool.

strates the tendency for six-factor models to perform especially well and for the IPIP to perform least well.

Questionnaire Length

Figure 2 graphs the differences in predictive ability of scores from the four sets of inventories depending on inventory length. The graphs in Figure 2 have more uniform lines than those in Figure 1, indicating that shorter versus longer versions of the same inventory usually predict quite similarly across the criteria. The largest exception to this is when the BFI is converted from a five- to a six-factor inventory—the predictive ability increases for most criteria but not substantially for any one. However the 10-item BFI, the shortest measure in our study, demonstrated impressive prediction, for half the criteria actually showing a detectably higher *R* than the 44-item BFI from which it was derived.

Likewise, the 50-item IPIP showed no predictive advantage over its 20-item counterpart. And, the 96-item HEXACO-PI demonstrated no predictive advantage over its 60-item version. The five versions of the QB6 were also highly comparable in their predictiveness; virtually all of the predictive capabilities of the longer versions are encapsulated in the shortest versions.

To establish that the strength of the brief BFI and IPIP compared with their longer versions was due to selecting the best possible items, and to rule out the alternate hypothesis that *any* shortened set of items from the longer measures would have equal strength, random alternate sets of both short measures were constructed. Items for three additional short sets for each measure were selected randomly from those items not used in the published short versions, maintaining balanced keying where it existed in the original scale and avoiding unbalanced overlap of items drawn between the different scales. Regressions of the criterion variables on these alternate 10-item BFI and 20-item IPIP inventories are reported in Table 7. Also reported in Table 7 is the comparative validity of scores from the 24QB6 with the other, unselected half of the 48QB6.

Table 6
Tally of Better and Worse Performance in Significance of R Change Values

Relative rank	Inventory	Better	Worse
1	48QB6	3	0
2	96QB6	3	0
3	24QB6	2	0
4	36QB6	2	0
5	BFI-6	2	0
6	BFI-10	1	0
7	HEXACO-60	1	0
8	60QB6	1	1
9	NEO-FFI	0	0
10	HEXACO-96	0	0
11	BFI-44	0	1
12	IPIP-20	1	2
13	IPIP-50	0	3

Note. QB6 = Questionnaire Big Six Scales; BFI = Big Five Inventory; HEXACO = HEXACO Personality Inventory; NEO-FFI = NEO Five Factor Inventory; IPIP = International Personality Item Pool. A point for "better" was given to an inventory if it demonstrated statistically significantly higher *R*-change value over any other inventory for an outcome in Table 4. A point for "worse" was tallied if, for any outcome in Table 4, this inventory had a significantly smaller *R*-change value than any other inventory. Where scores were equal, a higher rank was given to the shorter inventory.

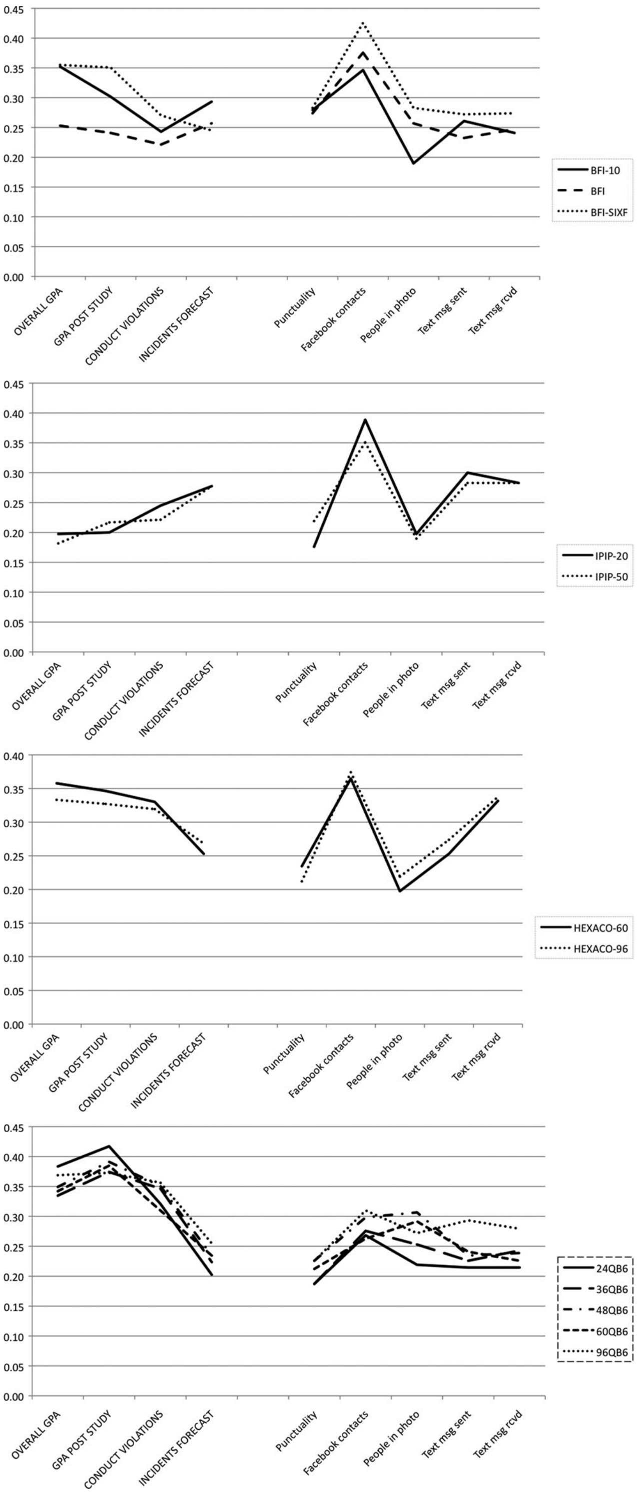

Figure 2. Predictive ability (*R* change) by questionnaire length. GPA = grade point average; BFI = Big Five Inventory; IPIP = International Personality Item Pool; HEXACO = HEXACO Personality Inventory; QB6 = Questionnaire Big Six Scales.

The original BFI-10's *R* values well exceeded the average *R* from the three random 10-item abbreviations (.35, .30, .26, .30 vs. .19, .19, .22, .22), which would tend to indicate that the creators of the BFI-10 were successful in selecting the best (most valid) items. For the IPIP-20 this difference was less dramatic (.20, .20, .22, .27 vs. .18, .22, .18, .24); this may indicate that for purposes of score validity the IPIP-20 may still not draw on the most optimal selection of items from the IPIP-50 (although the selection may be optimal for preserving internal consistency).

As noted earlier, the 24QB6 was selected in large part as being the better half of the longer 48QB6 with respect to item validity in another, very different sample. And indeed, in Table 7, the *R*-change values for the 24QB6 exceeded those for the unselected other half of the 48QB6. Thus, item-score validity from a much older community sample did appear to generalize to a student sample to some extent where a different set of criteria was used.

Discussion

In the current study, a "race" was orchestrated between popular Big Five and Big Six measures of moderate length, and their comparative ability to predict outcomes was assessed. The criterion variables included outcomes with real life significance for the population sampled: GPA (an amalgam of academic effort over multiple settings, topics, and raters) and student conduct violations (an indicator of rule-breaking versus rule-following behavior). Values on both variables were obtained 5–10 months after participation in the study, allowing us to examine prospective prediction. Secondarily, this study looked at students' behaviors ascertained on the day of participation: punctuality, Facebook profile picture

Table 7
R Change for Alternate Short Versions of BFI, IPIP, and QB6

Inventory	GPA overall	GPA post-study	Conduct: Total responsible	Complaints forecast
BFI-10	**.35**	**.30**	*.26*	**.30**
BFI-10 (2)	.17	.12	.15	.22
BFI-10 (3)	.18	.20	*.27*	.21
BFI-10 (4)	*.23*	*.24*	*.25*	.22
IPIP-20	.20	.20	.22	*.27*
IPIP-20 (2)	.15	.18	.10	.22
IPIP-20 (3)	*.24*	**.28**	.20	*.28*
IPIP-20 (4)	.15	.19	.19	.22
24-QB6	**.38**	**.42**	.34	.18
24-QB6 (2)	**.31**	**.34**	*.28*	.19

Note. GPA = grade point average; BFI = Big Five Inventory; IPIP = International Personality Item Pool; QB6 = Questionnaire Big Six Scales. The first version of each inventory listed is the published shortest version of the inventory (coefficients from Table 4). The alternates that follow were constructed by randomly selecting remaining items from the long versions of the inventory. In the case of the BFI, this meant randomly selecting BFI-44 items not used in Rammstedt and John's (2007) 10-item BFI without replacement to create alternate two-item scales. In the case of the IPIP, items from the 50-item versions that were not used in Donnellan et al. (2006) 20-item version were randomly selected with the constraint that only two items per four-item scale could overlap with any other alternate scale. In the case of the QB6, the alternate version includes all items from the 48-QB6 not included in the 24QB6. In all cases balanced keying was retained where it existed in the original long version of scales. Coefficients are italicized where $p < .05$ and bold where $p < .01$.

and contacts, and cell phone text-messaging usage, all indicators of the participants' typical mode of interaction with the social environment.

A surprise in the results of this study was the impressive predictive validity of scores on very short scales. This is especially remarkable given that an advantage of short scales, that they circumvent fatigue (Burisch, 1984), was not operative in the current study, where respondents spent an hour responding to about 400 items. While reliability (alpha) of scores on the scales decreases as scale length decreases (see Table 2), scores on the abbreviated measures in this study maintained criterion validity impressively well, often even surpassing their longer counterparts. Holding questionnaire and number of scales constant, and examining scores on briefer versus longer versions for relative predictive ability, our results provide no support for using longer versions. These results may speak to a ceiling effect in the predictive power of self-report questionnaires. Some important variation in psychological differences can be captured by self-report inventories, but it may be of a finite amount, sufficiently obtainable with a few high-validity items. Adding additional items to gauge the same tendency in different settings or with varied wording may fail to improve a scale from a predictive-validity standpoint.

The strong validity of scores on the brief inventories in the study is of course partly explained by the fact that developers of short inventories are able to draw on bodies of previous research using longer versions of the inventories, in order to choose the strongest, core items for each scale. Our results suggest that the best scale-construction strategy may involve a large pool of initial items reduced (if practicality dictates) to a medium-length measure, later culminating in a brief inventory based on analysis of multiple data sets. The most impressive measure on a per-item basis in these results, the BFI-10, is also the one whose item-selection was based on the largest collection of pre-existing data. From the standpoint of efficiency-maximization, many widely used longer personality inventories might best be considered unfinished—prematurely treated as endpoints, when they could be better viewed as way stations in the direction of even more efficient measures.

Our results also challenge the common assumption that reliability is a prerequisite for validity and suggest that the attenuation paradox may come into play even in inventories of modest length (40–60 items). The BFI-10 had four of five coefficient alpha values in the .43–.60 range. Despite weaker internal consistency than the other Big Five inventories, the predictive capability of its scores was detectably highest for nearly half of the criteria (although not substantially higher by the standard used here). It is possible that beyond the 10 core BFI items selected by Rammstedt and John (2007), the additional 34 items may be a mixed blessing, adding to score reliability while also adding noise to the signal from a predictive standpoint.

Cronbach (1990, p. 213) wrote that "other things being equal, the more accurate a test [the more observations and internal consistency, the less measurement error], the stronger its correlation with other variables." That higher internal consistency facilitates higher predictive validity is a conventional assumption embodied in the use of corrections for attenuation due to unreliability. However, note Cronbach's crucial qualifier, "other things being equal." The situation may be different if items are unequal in their validity; adding less valid items to a set of highly predictive items will actually lower the longer scale's predictive capability (Thorndike, 1967). An old psychometric axiom is that one maximizes internal consistency by selecting items with high correlations among one another but maximizes score validity by selecting items that correlate highly with criteria and at a low level with each other. The results in this report support that axiom: The path toward internal consistency (by padding a scale with more items) is not necessarily the path toward predictive validity. Thorndike (1967, p. 214) commented that "exclusive preoccupation with item internal consistency may lead to an undue narrowing of the scope" of a measure, a decrease of validity consistent with the attenuation paradox. This view was also recently supported by McCrae, Kurtz, Yamagata, and Terracciano (2011), who found that in a large dataset of NEO facet scales, internal consistency, despite being the most widely used measure of reliability, was "essentially unrelated to differential validity" (p. 42).

While versions of the same inventories tended to show strong similarity on predictive capability, and thus a lack of advantage for lengthier inventories, comparisons across different inventories showed more variation. Each inventory appears to have strengths and weaknesses. For the two 96-item inventories, the 96QB6 demonstrated a marginal advantage for the life outcomes, and the HEXACO-PI for behavioral observations (see Figure 1).

Three inventories (NEO-FFI, HEXACO-60, 60-QB6) had 60 items. Here, again, predictive capability varied. Scores on the HEXACO-60 and 60-QB6 had marginally higher R values than the NEO-FFI in terms of life outcomes, but the NEO–FFI did just as well in predicting observed behavior.

Three inventories had 44 to 50 items. The relevant QB6 version showed higher R values for the life outcomes, substantially so in comparison to the IPIP-50. But the BFI most often had the highest R values for behavioral observations.

The shortest inventories were not well matched on number of items but bear a brief mention. The 24QB6 and BFI-10 had a detectable advantage over the IPIP-20 on life outcomes, but they had no evident advantage over the IPIP-20 for behavioral observations. It is worth noting that scores on the efficient BFI-10 did not have substantially lower predictive validity than these longer competitor measures (see Table 6).

Among the Big Five inventories, the shortest (the BFI-10) was a substantially better predictor of GPA than some longer ones (i.e., IPIP inventories). For the behavioral observations, it is difficult to discern a predictive advantage for any Big Five inventory over another.

Did the Big Six outperform the Big Five? There seemed to be some overall predictive advantage for six-factor inventories for *L*-data criteria. That is, the highest R values for predicting GPA and being found responsible for conduct-code violations were derived almost entirely from six-factor inventories. Furthermore, the rank orders in Table 6 might suggest that in the current study, the Big Six measures won the race to predict student life outcomes: Six-factor inventories hold the top five positions in the rank order. This serves to support the contention of some investigators (e.g., Saucier, 2009; Almagor, Tellegen, & Waller, 1995) that a more inclusive variable selection leads downstream to measures that perform more capably; Saucier's model of the Big Six—like the Big Seven of Tellegen, Waller, and colleagues—is based on lexical studies with inclusive variable selection. (Given this finding, it would be interesting to include the Big Seven in future studies of this kind.)

Why did the Big Six inventories achieve higher rankings in Table 6 compared with Big Five inventories? Some might argue that we are comparing apples and oranges, because the models differ in several ways, beyond the addition of the Honesty/Propriety or Honesty/Humility dimension. But as an advantage of our design, the nested BFI and the BFI-six allow for a relatively direct comparison between models, as four of the scales (N, E, C, and O) are identical across the two inventories. Table 4 indicates that in the current data, the BFI-six had observably higher *R* values than its five-factor counterpart. The BFI-six had the advantage of being nine items longer, but length is not likely the sole explanation of its advantage, as the 10-item BFI, sheared of 34 items, also had higher *R* values than the BFI-44. As described above, the BFI-six includes 10 Honesty/Propriety (H) items and eight alternative Agreeableness items that replace the nine BFI-44 Agreeableness items. Note that in Table 5, for the four validity criteria on which the largest *R* change difference exists between the BFI and BFI-six factor versions (GPA, GPA post study, student conduct responsible total, and Facebook contacts), H is a significant predictor. The two Agreeableness scales appear similar in their predictive abilities in this dataset.

The Honesty/Propriety or Honesty/Humility scale was significantly related to both GPA variables for every six-factor inventory. What might give this dimension its power to predict this complex behavioral outcome? As described above, this dimension has emerged in diverse lexical studies and includes content related to ethical behavior (honesty, humility, and integrity vs. greed), some of which is occasionally included in dimensions of Conscientiousness or Agreeableness but more often is left out of the Big Five. This content relates to regulating behavior by expectations related to sociomoral norms and, as such, is more interpersonal in nature than the content comprised by Conscientiousness. While Conscientiousness contributes to punctuality and organization, Honesty/Propriety or Honesty/Humility is also logically related to GPA, as grades are partly a reflection of attending to normative expectations set out in a social context, such as the expectations defined in a syllabus, and to following through with a commitment.

The Originality/Talent dimension of the QB6 inventories contributed substantially to prediction on many criteria, including GPA, in contrast to the Openness or Intellect dimensions of the Big Five and HEXACO inventories, in which this dimension was in general only significantly negatively related to number of Facebook contacts. One might argue that the predictive advantage for the QB6 stemmed from use of "talent" descriptors not found in the Big Five. There are two reasons to discount this interpretation, however. First, the IPIP-50 has "talent" items (e.g., "Am quick to understand things," "Have excellent ideas") on the Intellect scale, in higher proportion than the QB6 Originality/Talent scales, yet showed no predictive advantages for GPA. Secondly, for the QB6 scales, the proportion of talent descriptors (e.g., "Am considered to be a wise person," "Have a rich vocabulary") is always less than half the total number of Originality scale items. It is more likely that differences in content emphasis between the QB6 and IPIP items for originality/talent/intellect (e.g., references to curiosity, insight, and wisdom, in the QB6, versus references to imagination, reflection, and "having excellent ideas" in the IPIP) are responsible for differences between them in prediction of GPA.

Relation to psychopathology was not explored in the current study, but based on results with adjective scales (correlation with externalizing disorder tendencies; Saucier, 2009) and with inventories that add dimensions of positive and negative valence to the Big Five (Simms, Yufik, & Gros, 2010), we expect the Big Six to demonstrate stronger relations and superior prediction in this domain. While the NEO-PI-R (Costa & McCrae, 1992) has been the instrument of choice when researchers have compared traits to personality disorders or *DSM* Axis I disorders (e.g., Lowe & Widiger, 2008; Trull & Durrett, 2005), the Big Six may better elucidate these relations due to better articulated relations with internalizing and externalizing spectra. Big Six questionnaires do not mix impulsivity with anxiety in the dimension of Neuroticism (these are more likely opposite predictors of externalizing versus internalizing problems; Carver, 2005). Neither does the Big Six mix angry hostility with other negative emotions, instead placing this content distinctly in the domain of low Agreeableness. Furthermore, positive and negative valence content (to some extent included in the honesty/propriety dimension of Big Six inventories) has been shown to incrementally increase prediction of personality disorders, beyond the Big Five (Simms et al., 2010).

Decision-making and risk-taking were explored in a general sense by looking at student conduct records but could be a fruitful direction in which to further extend Big Five and Big Six comparisons in the future. Weller and Tikir (2011) found the HEXACO model useful in elucidating domain specific aspects of risk-taking, beyond what would be possible with the Big Five. In that study, HEXACO Emotionality was correlated with heightened risk-perceptions and Conscientiousness with less perceived benefits in all risk domains studied, whereas Openness predicted more risk-taking in social and recreational areas, and low Honesty/Humility was associated with increased health, safety, and ethical risk-taking.

A few caveats to our conclusions are in order. The remarkable performance of very short measures in this study may be a tendency that does not generalize beyond the particular predictor measures, outcomes, and population examined here. It was impractical in the current study to compare full-length inventories (such as the NEO-PI-R) with one another and thus to determine (a) whether our findings stem from the peculiar weaknesses of inventories in the 36- to 96-item range, or (b) how much "facet-level" prediction might add to predictive capability.[2] Furthermore, while the mean scores on brief, short and medium-length scales of inventories were highly similar in this study (see Table 2), comparison with the longest measures would allow for exploration of intercept bias: The shortest measures might differ from the longest versions in mean or difficulty level, and this could have consequences in applied assessment contexts. Finally, our sample size was substantial but not huge. Better statistical power might lead to a more refined understanding of predictive differences across inventories.

[2] As mentioned above, Costa and McCrae (1989) estimated that the NEO-FFI has 75% the predictive validity of the full scale measure. Using this ratio, hypothetical *R* values for the NEO-PI-R were estimated. For three criteria (complaints forecast, Facebook contacts, and people in photo) this was the highest *R* in the set. In the latter two cases, the hypothetical values appeared to be slightly larger than those for other inventories in the study. In the first case, however, it was not significantly larger than other large values. For the remaining eight criteria, the hypothetical NEO-PI-R R value had no advantage over the values of much shorter measures.

Conclusions

For basic research purposes that call for a Big Five measure, the BFI-10 requires a minimum of participant time and seems, based on these results, at least as predictively capable as the standard 44-item BFI. Our study indicates, however, that a six-factor extension of the BFI may have slightly higher predictive capability than the five-scale BFI (and a very short version of this extended BFI can be developed). The Big Six draws on a larger base of lexical research than does the Big Five, incorporating studies from more diverse languages and studies that use more inclusive variable selection criteria. It is thus more likely to replicate well in diverse settings and to be reproduced within a greater diversity of variable selection strategies. The Honesty scale of six-factor inventories was related to important outcomes in this study, including grades and student conduct charges, and contributed to the prediction and interpretation of behavioral observations, sometimes evidencing a negative relation to number of Facebook contacts and text messaging. Furthermore, it is hypothesized that inclusion of content related to honesty and propriety, and the better separation and articulation of internalizing emotion and externalizing behavior content will lead inventories based on the Big Six model to demonstrate advantages when predicting important mental health and decision making outcomes, among other real-life criteria. Such a model may give the Big Five a meaningful upgrade, enriching and expanding this useful model of personality attributes for the 21st century.

References

Alalehto, T. (2003). Economic crime: Does personality matter? *International Journal of Offender Therapy and Comparative Criminology, 47,* 335–355.

Allport, G. W., & Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs, 47*(Serial No. 211).

Almagor, M., Tellegen, A., & Waller, N. (1995). The Big Seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language trait descriptors. *Journal of Personality and Social Psychology, 69,* 300–307. doi:10.1037/0022-3514.69.2.300

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11,* 150–166. doi:10.1177/1088868306294907

Ashton, M., & Lee, K. (2009). The HEXACO-60: A short measure of the major dimensions of personality. *Journal of Personality Assessment, 91,* 340–345. doi:10.1080/00223890902935878

Ashton, M., & Lee, K. (2010). On the cross-language replicability of personality factors. *Journal of Research in Personality, 44,* 436–441. doi:10.1016/j.jrp.2010.05.006

Ashton, M., Lee, K., Perugini, M., Szarota, P., de Vries, R. E., Di Blas, L., . . . De Raad, B. (2004). A six-factor structure of personality-descriptive adjectives: Solutions from psycholexical studies in seven languages. *Journal of Personality and Social Psychology, 86,* 356–366. doi:10.1037/0022-3514.86.2.356

Benet-Martinez, V., & Waller, N. G. (1997). Further evidence for the cross-cultural generality of the Big Seven factor model: Indigenous and imported Spanish personality constructs. *Journal of Personality, 65,* 567–598.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39,* 214–227. doi:10.1037/0003-066X.39.3.214

Carver, C. S. (2005). Impulse and constraint: Perspectives from personality psychology, convergence with theory in other areas, and potential for integration. *Personality and Social Psychology Review, 9,* 312–333. doi:10.1207/s15327957pspr0904_2

Cattell, R. B. (1957). *Personality and motivation structure and measurement.* New York, NY: World Book.

Church, A. T., Reyes, J. A. S., Katigbak, M. S., & Grimm, S. D. (1997). Filipino personality structure and the Big Five model: A lexical approach. *Journal of Personality, 65,* 477–528.

Clark, L. A. (2005). Temperament as a unifying basis for personality and psychopathology. *Journal of Abnormal Psychology, 114,* 505–521. doi:10.1037/0021-843X.114.4.505

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319. doi:10.1037/1040-3590.7.3.309

Clower, C. E., & Bothwell, R. K. (2001). An exploratory study of the relationship between the Big Five and inmate recidivism. *Journal of Research in Personality, 35,* 231–237. doi:10.1006/jrpe.2000.2312

Costa, P. T., Jr., & McCrae, R. R. (1989). *NEO PI/FFI manual supplement.* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1990). *Essentials of psychological testing.* New York, NY: Harper & Row.

De Raad, B., Barelds, D. P. H., Levert, E., Ostendorf, F., Mlačić, B., Di Blas, L., . . . Katigbak, M. S. (2010). Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. *Journal of Personality and Social Psychology, 98,* 160–173. doi:10.1037/a0017184

De Raad, B., DiBlas, L., & Perugini, M. (1998). Two independently constructed Italian trait taxonomies: Comparisons among Italian and between Italian and Germanic languages. *European Journal of Personality, 12,* 19–41.

De Raad, B., Hendriks, A. J., & Hofstee, W. K. (1992). Towards a refined structure of personality traits. *European Journal of Personality, 6,* 301–319. doi:10.1002/per.2410060405

Digman, J. M. (1996). The curious history of the Five Factor model. In J. S. Wiggins (Ed.), *The Five Factor model of personality: Theoretical perspectives* (pp. 1–20). New York, NY: Guilford Press.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. *Psychological Assessment, 18,* 192–203. doi:10.1037/1040-3590.18.2.192

Funder, D. (2007). *The personality puzzle.* New York, NY: Norton.

Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment, 4,* 26–42. doi:10.1037/1040-3590.4.1.26

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American Psychologist, 48,* 26–34. doi:10.1037/0003-066X.48.1.26

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

Grucza, R. A., & Goldberg, L. R. (2007). The comparative validity of 11 modern personality inventories: Predictions of behavioral acts, informant reports, and clinical indicators. *Journal of Personality Assessment, 89,* 167–187.

Hirsh, J. B., & Peterson, J. B. (2008). Predicting creativity and academic success with a “fake-proof” measure of the Big Five. *Journal of Research in Personality, 42,* 1323–1333. doi:10.1016/j.jrp.2008.04.006

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The Big Five Inventory* (Versions 4a and 54). Berkeley, CA: University of California, Berkeley, Institute of Personality and Social Research.

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). New York, NY: Guilford Press.

Johnson, J. A. (2000). Predicting observers' ratings of the Big Five from the CPI, HPI, and NEO-PI-R: A comparative validity study. *European Journal of Personality, 14,* 1–19.

Lee, K., & Ashton, M. (2004). Psychometric properties of the HEXACO personality inventory. *Multivariate Behavioral Research, 39,* 329–358. doi:10.1207/s15327906mbr3902_8

Loevinger, J. (1954). The attenuation paradox in test theory. *Psychological Bulletin, 51,* 493–504. doi:10.1037/h0058543

Lowe, J. R., & Widiger, T. A. (2008). Personality disorders. In J. E. Maddux & B. A. Winstead (Eds.), *Psychopathology: Foundations for a contemporary understanding* (pp. 223–250). Mahwah, NJ: Erlbaum.

McCrae, R. R., & Allik, J. (Eds.). (2002). *The Five-Factor model of personality across cultures. International and cultural psychology series.* New York, NY: Kluwer Academic/Plenum Press.

McCrae, R. R., & Costa, P. T. (1997). Personality trait structure as a human universal. *American Psychologist, 52,* 509–516.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review, 15,* 28–50. doi:10.1177/1088868310366253

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big Five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology, 93,* 116–130. doi:10.1037/0022-3514.93.1.116

Ostendorf, F. (1990). Sprache und persönlichkeitsstruktur: Zur Validität des Fünf-Faktoren-Modells der Persönlichkeit [Language and personality structure: Toward the validation of the Five-Factor model of personality]. Regensberg, Germany: Verlag.

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41,* 203–212.

Saucier, G. (1997). Effects of variable selection on the factor structure of person descriptors. *Journal of Personality and Social Psychology, 73,* 1296–1312. doi:10.1037/0022-3514.73.6.1296

Saucier, G. (1998). Replicable item-cluster subcomponents in the NEO Five-Factor Inventory. *Journal of Personality Assessment, 70,* 263–276. doi:10.1207/s15327752jpa7002_6

Saucier, G. (2003). An alternative multiple-language structure of personality attributes. *European Journal of Personality, 17,* 179–205. doi:10.1002/per.489

Saucier, G. (2009). Recurrent personality dimensions in inclusive lexical studies: Indications for a Big Six structure. *Journal of Personality, 77,* 1577–1614. doi:10.1111/j.1467-6494.2009.00593.x

Saucier, G., Georgiades, S., Tsaousis, I., & Goldberg, L. R. (2005). The factor structure of Greek personality adjectives. *Journal of Personality and Social Psychology, 88,* 856–875.

Saucier, G., Ole-Kotikash, L., & Payne, D. L. (2006). *The structure of personality and character attributes in the language of the Maasai.* Unpublished manuscript, University of Oregon.

Simms, L. J., Yufik, T., & Gros, D. F. (2010). Incremental validity of positive and negative valence in predicting personality disorder. *Personality Disorders: Theory, Research, and Treatment, 1,* 77–86. doi:10.1037/a0019752

Szirmak, Z., & De Raad, B. (1994). Taxonomy and structure of Hungarian personality traits. *European Journal of Personality, 8,* 95–117.

Tellegen, A., & Waller, N. G. (2008). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *The SAGE handbook of personality theory and assessment. Vol. 2: Personality measurement and testing* (pp. 261–292). Thousand Oaks, CA: Sage.

Thorndike, R. L. (1967). The analysis and selection of test items. In D. N. Jackson & S. Messick (Eds.), *Problems in human assessment* (pp. 201–216). New York, NY: McGraw-Hill.

Trull, T. J., & Durrett, C. A. (2005). Categorical and dimensional models of personality disorder. *Annual Review of Clinical Psychology, 1,* 355–380. doi:10.1146/annurev.clinpsy.1.102803.144009

van Dam, C., Janssens, J. M. A. M., & De Bruyn, E. E. J. (2005). PEN, Big Five, juvenile delinquency and criminal recidivism. *Personality and Individual Differences, 39,* 7–19. doi:10.1016/j.paid.2004.06.016

Weller, J. A., & Tikir, A. (2011). Predicting domain-specific risk taking with the HEXACO personality structure. *Journal of Behavioral Decision Making, 24,* 180–201.

Zhou, X., Saucier, G., Gao, D., & Liu, J. (2009). The factor structure of Chinese personality descriptors. *Journal of Personality, 77,* 363–400.

Received August 12, 2010
Revision received April 19, 2011
Accepted April 20, 2011 ■