JOURNAL OF PERSONALITY ASSESSMENT, *89*(2), 167–187

The Comparative Validity of 11 Modern Personality Inventories: Predictions of Behavioral Acts, Informant Reports, and Clinical Indicators

Richard A. Grucza
Department of Psychiatry
Washington University School of Medicine

Lewis R. Goldberg
Oregon Research Institute
Eugene, Oregon

In science, multiple measures of the same constructs can be useful, but they are unlikely to all be equally valid indicators. In psychological assessment, the many popular personality inventories available in the marketplace also may be useful, but their comparative validity has long remained unassessed. This is the first comprehensive comparison of 11 such multiscale instruments against each of three types of criteria: clusters of behavioral acts, descriptions by knowledgeable informants, and clinical indicators potentially associated with various types of psychopathology. Using 1,000 bootstrap resampling analyses from a sample of roughly 700 adult research participants, we assess the relative predictability of each criterion and the comparative validity of each inventory. Although there was a wide range of criterion predictability, most inventories exhibited quite similar cross-validities when averaged across all three types of criteria. On the other hand, there were important differences between inventories in their predictive capabilities for particular criteria. We discuss the factors that lead to differential validity across predictors and criteria.

> *No technology of which we are aware—computers, telecommunications, televisions, and so on—has shown the kind of ideational stagnation that has characterized the testing industry. Why? Because in other industries, those who do not innovate do not survive. In the testing industry, the opposite appears to be the case. Like Rocky I, Rocky II, Rocky III, and so on, the testing industry provides minor cosmetic successive variants of the same product where only the numbers after the names substantially change. These variants survive because psychologists buy the tests and then loyally defend them.* (Sternberg & Williams, 1998, p. 577)

Before the founding of the Consumers Union and the publication of its journal, *Consumer Reports*, retail marketers could say just about anything about their products and no one would be the wiser. The consumer movement changed history by applying the scientific method to the comparison of competing goods and services.

Among the competing products developed by psychologists, perhaps the most important are their assessment instruments. Unfortunately, in psychology we have no Consumers Union to test competing claims and to compare these products on their overall effectiveness. In this article, we hope to persuade readers to join us in applying the lessons of the consumer movement to the evaluation of modern personality inventories. Psychological test manuals typically include dozens of tables listing correlations with other inventory scales and with various criterion indices. In the hundreds of tables included in the manuals of modern personality inventories, however, there are few that include any *comparative* analyses of the merits of that inventory with its competitors. Why is that?

COMPARING SOAPS AND COMPARING TESTS

Just as different brands of handsoaps differ from one another in their color, smell, and feel, personality inventories differ from one another in a host of ways, including the number and nature of their items and the number and nature of the scales they provide. A latent assumption of many test developers seems to be that these differences make comparisons impossible, and every inventory has its unique niche. As with soaps, however, some features of inventories are more crucial than others. Soaps should clean as effectively as possible, while not harming the user's skin. And inventories should predict important life behaviors and outcomes as effectively as possible, for any given amount of testing time.

Indeed, the analogy may be even stronger: Soaps differ in the variety of waste materials they can remove, and in their effectiveness in removing each of those types of materials. And, inventories differ in the variety of different outcomes they can predict (their bandwidth) and in the extent to which they can predict each outcome (their fidelity). In psychometrics as in audio engineering, bandwidth and fidelity are normally negatively related to each other, but the correlation is far from perfect. Some soaps do a lousy job of cleaning most waste products, and some inventories may predict near nothing of interest or value. What we need are empirical comparisons among inventories, similar in spirit to the product tests found in *Consumer Reports*. Instead, the best that we seem to have been able to come up with are the subjective and noncomparative test reviews found in such sources as the *Buros Mental Measurements Yearbooks*.

EVALUATING PERSONALITY INVENTORIES

Inventories might be compared on indices reflecting their bandwidth, their fidelity, or both. Bandwidth refers to the degree of diversity of the behaviors that can be predicted from an inventory, and fidelity refers to the ability of the inventory to predict each type of behavior within its range. The narrower the inventory's bandwidth, the more testing time can be devoted to the measurement of a single attribute, and, other things being equal, the more validly that attribute can be measured.[1] For example, inventories developed to measure only employee honesty and integrity (Goldberg, Grenier, Guion, Sechrest, & Wing, 1991) might be able to measure those aspects of conscientiousness with greater fidelity in the same amount of testing time than can inventories of broader bandwidth that contain content intended to measure many other personality attributes and employee behaviors. Because the developers of most inventories aspire to measure an extremely broad array of personality attributes, it makes sense to compare them on their breadth.

We can index fidelity by the size of the cross-validated multiple correlation between a set of inventory scales and a particular criterion. In turn, the *average* cross-validity coefficient across a broad array of criteria (Ashton & Goldberg, 1973; Goldberg, 1972; Hase & Goldberg, 1967) can be used as an evaluative index that combines bandwidth and fidelity. It will be used in this role in the present series of studies.

WHAT SHOULD ONE USE FOR CRITERIA?

One prominent approach to criterion specification is focused on the *frequency* with which individuals perform behavioral acts classified within socially relevant categories (e.g., Buss & Craik, 1983, 1984, 1985, 1987). This "act-frequency" approach to criterion measurement is limited only by our ingenuity in measuring the relative frequencies of individuals' behavioral acts. Within some industrial and other organizational settings, it may be possible to find objective indices of act performance, such as the dollar amount of sales, the number of products produced, or the number of accidents or other negative incidents, per unit of time. In most settings, however, human observers—including the target individual as a self-observer—must be used as recorders and transducers of such behavioral acts. For certain categories of behavioral acts, the self-observer may be the only person who is privy to the information needed to assess the act frequencies. In the first study of this series, we use as our criteria six highly reliable clusters of behavioral acts, two composed of relatively undesirable behaviors, two of relatively desirable behaviors, and two of relatively neutral types of behaviors.

In addition to their ability to predict specific behavioral acts, personality inventories are often touted as predictors of broad personality traits, as filtered through the eyes of those who know the target best. In many ways, these aggregate personality descriptions have come to represent the "gold standard" of the assessment enterprise, providing a composite portrait of an individual as seen by others with whom he or she interacts. Moreover, whereas in some contexts such as personnel selection self-descriptions might be distorted by the target's desires for impression management, the descriptions by knowledgeable others should not normally be subject to these kinds of pressures. In the second study in this series, we use as criteria descriptions by knowledgeable informants about the target's personality traits, employing two different measures of the broad Big-Five factor constructs (e.g., Goldberg, 1992).

In our third study, we shift from criteria reflecting essentially normal acts and traits to behavioral tendencies reflecting more disordered types of characteristics. The association between measures of broad personality traits and characteristics of clinical interest, including both symptom measures and personality disorders, has been a topic of increasingly active investigation. Moreover, the availability of genetic data has spurred interest in personality as an endophenotype for psychopathology. Hence, there is a need to determine which personality constructs are most reliably associated with measures of clinical interest. The third study in this series marks a first attempt at addressing this issue in a comparative-validity context. Criteria for Study 3 include two measures related to impulsive personality disorders, two measures of neurotic symptoms, and two measures that might be described as indicators of elevated risk for psychosis. One is a screen for dissociative disorders and the other as a screen for some precursors to schizophrenia.

[1] This classic view of the bandwidth–fidelity issue, however, has been questioned by Burisch (1984b) and more recently by McGrath (2005).

OVERVIEW OF THE PRESENT ARTICLE

Using adults who have been participating over the past 13 years in an ongoing community sample, self-reported act frequencies (Study 1), reports from knowledgeable others (Study 2), and measures of clinical interest (Study 3) here are employed as criterion indices for empirical comparisons of the relative validities of 11 popular modern personality inventories. The goal of this article is to provide some initial answers to questions such as, (1) How do some of the most popular modern personality inventories compare in their cross-validities as predictors of these three diverse types of criteria? (2) Using items from the same inventory, is it better to employ a few relatively broad higher-level scales (thus maximizing measurement reliability) or is it better to use a larger number of narrow lower-level scales (thus utilizing the specific variance associated with each of the individual scale scores)? (3) What are the specific personality scales that correlate most highly with individual criteria and therefore may contribute to the ability of some inventories to predict criteria with higher fidelity than others?

METHODS COMMON TO ALL THREE STUDIES

The Research Participants

In 1993, roughly 500 men and 500 women from one metropolitan area in Oregon were recruited for participation in the Eugene–Springfield Community Sample (ESCS). To facilitate administration of a series of assessments over several years, recruitment was limited to homeowners who agreed to participate for at least 5 to 10 years. The sample is diverse on most personality and demographic variables except for race/ethnicity (the county is about 94% Caucasian and less than 1% African American). In 1993, the sample ranged in age from 18 to 85, with a mean age of 51 (standard deviation = 13). For the subsamples used in the present analyses, approximately 98% were White, 58% were female, and 55% had college degrees. All inventories were mailed to the participants who returned the completed forms by mail in return for honorarium checks that ranged from $10 to $25 over the years. Only identification code numbers were included on each inventory, and participants always were instructed not to include their names. For 10 of the 11 inventories, data were missing for only a small subset of participants; the size of the missing subset ranged from 3 to 133. Because the IPIP-AB5C items (Goldberg, 1999a) were administered over several mailings, data were missing for 257 participants.

The Inventories to Be Compared

Nine commercial personality inventories were administered periodically between 1993 and 2000 as follows.

a. The revised NEO inventory (NEO-PI-R: Costa & McCrae, 1992) included 240 items with five-step response options, and provides 30 lower-level "facet" scales, six associated with each of its five higher-level "domains," labeled Neuroticism (N), Extraversion (E), Openness to Experience (O), Agreeableness (A), and Conscientiousness (C). The NEO-PI-R was administered during the summer of 1994.
b. The California Psychological Inventory (CPI: Gough & Bradley, 2002) included 462 True/False items. For the present project, 36 CPI scales were scored, including all 20 of its basic scales, the three highest-level "vector" scores, and 13 special scales. The CPI was administered in the fall of 1994.
c. The 5th edition of Cattell's Sixteen Personality Factor Questionnaire (16PF: Conn & Rieke, 1994; Russell & Karol, 1994) included 185 items with three-step response options. The inventory provides scores for each of 16 scales, one of which is a short intelligence test; in addition, five "second-order" or "global" factors also are available. The 16PF was administered in the fall of 1996.
d. The Hogan Personality Inventory (HPI: Hogan & Hogan, 1995) included 206 True/False items. Its seven higher-level scales are labeled Ambition, Sociability, Likeability, Prudence, Adjustment, Intellectance, and School Success, each of which is derived from subsets of 44 lower-level homogeneous item clusters (HICs). The HPI was administered during the winter of 1997.
e. The Temperament and Character Inventory (TCI-R: Cloninger, Przybeck, Svrakic, & Wetzel, 1994) was administered in the spring of 1997. A revised version of this 240-item inventory with five-step response options measured four higher-level "temperament" scales labeled Novelty-seeking, Harm-Avoidance, Reward-Dependence, and Persistence, each of which was scored from four lower-level subscales. In addition, there were three higher-level "character" scales, labeled Self-directedness, Cooperativeness, and Self-transcendence, each scored from three to five subscales.
f. The Multidimensional Personality Questionnaire (MPQ: Tellegen, in press; Tellegen & Waller, in press) was administered in the summer of 1999. Based on 276 True/False items, the MPQ included 11 content scales plus a measure of "Unlikely Virtues" (social desirability response bias). Each of the scales of the MPQ contains between two and four subscales, for a total of 30 subscales.
g. The revised version of the Jackson Personality Inventory (JPI-R: Jackson, 1994) included 300 True/False items from which 15 content scales are scored. The JPI-R was administered in the fall of 1999.
h. The Six-Factor Personality Questionnaire (6FPQ: Jackson, Paunonen, & Tremblay, 2000) was also administered in the fall of 1999. This inventory included 108 items with a 5-point response scale. Six higher-level

constructs, labeled Extraversion, Agreeableness, Methodicalness, Independence, Openness to Experience, and Industriousness are derived from 18 lower-level subscales.

i. The HEXACO Personality Inventory (HEXACO-PI: Lee & Ashton, 2004) included 192 items with five-step response options, and contained 24 lower-level "facet" scales, four subsumed within each of its six higher-level "domains," labeled Honesty–Humility (H), Emotionality (E), eXtraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O). The HEXACO-PI was administered during the spring of 2003.

In addition, we included two sets of measures that are freely available in the public domain:

j. The set of 100 unipolar adjective markers of the Big-Five factor structure developed by Goldberg (1992) was administered during the summer of 1993. Its five scales are labeled Extraversion, Agreeableness, Conscientiousness, Emotional Stability (versus Neuroticism), and Intellect/Imagination.

k. The 485 items in the IPIP-AB5C Inventory (Goldberg, 1999a) were administered on different occasions between 1994 and 2000. These 45 scales are targeted at each of the bipolar facets in the Abridged Big Five-dimensional Circumplex (AB5C) model of Hofstee, de Raad, and Goldberg (1992).

Analytic Methodology

Our general methodology, which is the same in all three studies, both corrects for the overfitting that is endemic to multiple regression analyses and simultaneously provides a means for testing the statistical significance between pairs of cross-validated multiple correlation coefficients. To accomplish these two aims, we employed a method that combines bootstrap resampling procedures with split-sample cross-validation (Browne & Cudeck, 1989; Efron, 1983; Efron & Tibshirani, 1986). For each criterion and each inventory in turn, we used the set of inventory scales as predictors in a stepwise multiple regression analysis. In each analysis, 1,000 random resamplings were drawn from the data set for which complete predictor-criterion information was available, with each resample including the same number of individuals as in the original sample. Each resampling was conducted *with replacement*; individuals selected for the resample are not removed from the original sample, and therefore individuals can be selected on more than one occasion.

As a result of resampling with replacement, in each analysis about two thirds of the participants are included at least once, and about one third are left out. It is this latter portion of participants who then serve as a "hold-out" sample for cross-validation purposes. The predicted criterion values for the hold-out subjects are calculated using the regression weights from the derivation equation. The correlation between the predicted and actual criterion values in the hold-out subsample then is used as our index of cross-validity. Although sampling with replacement is somewhat counterintuitive, this is the standard procedure used in modern bootstrap-resampling analyses; its purpose is to ensure that the outcome of each selection is independent of the outcomes of all other selections.

In each of these multiple regression analyses, we always controlled for the demographic variables of gender, age, and educational level by entering those variables first in each regression equation, prior to including the personality inventory scale scores. Then, the most predictive set of *five* scales was selected using a stepwise regression model.[2] In reporting our findings from each of the three studies, we provide the correlations between each of the three demographic variables and the relevant criterion measures, as well as the corresponding partial correlations controlling for the effects of the other two demographic variables.

Finally, as a result of our 1,000-fold resampling procedure, the distributions of the cross-validity coefficients can be compared, inventory by inventory, for each criterion, allowing a test of whether the difference between any two cross-validity coefficients is statistically significant. To conduct these tests, the distributions of the differences between the squares of the cross-validity coefficients were computed. The mean and standard deviation of these distributions were then used in t tests to assess the significance of the differences between pairs of coefficients. For each criterion, these tests were conducted for all possible inventory pairs.

Obviously, reporting the t-test results for each pairwise comparison would be impracticably cumbersome. Instead, to get an estimate of the minimum value of the difference that is significant, we conducted a regression analysis of the resulting t values against the actual squared differences between each pair of inventories. This was done separately for each criterion. In every case, the equation accounted for more than 90% of the variance of the values from the t tests. Using the regression equation, we computed that value of a pairwise difference for which the t test value was equal to the customary critical value of 1.96. When the actual difference for a pair of inventories exceeds the critical value, then the t test for that pair of inventories would be expected to yield $t \geq 1.96$, and the inventories would differ at $p < .05$. A table providing these critical pairwise differences is included in our Appendix. Because this value is a close approximation of an explicit t test, it serves as a useful metric for evaluating the magnitude of differences between pairs of cross-validity coefficients.

[2] Preliminary analyses were conducted in which a criterion was predicted from all of the inventories, using the bootstrap cross-validation procedure that we described. In general, no substantial improvement in the cross-validated multiple correlation coefficients occurred after five steps of variable selection, but five-variable models were generally more effective than less exhaustive models. The choice of five steps is not necessarily optimal for all inventories and all criteria, but improvements in the coefficients beyond five steps are likely to be very small, on the order of .01 or less. Hence, we chose five step variable selection as a balance between more complex models, which tend to overfit the data, and simpler models that do not include enough of the predictive variance.

TABLE 1
Clusters of Behavioral Acts Used as Criterion Variables

Drug Use (14 acts: Mean Interitem $r = .38$; Alpha = .89; Skew = .23)	
Smoked marijuana.	Became intoxicated.
Drank wine.	Had a hangover.
Drank alcohol during working hours.	Drank in a bar.
Went to a nightclub.	Drank beer.
Drank alcohol or used other drugs to make myself feel better.	Took a hard drug (for example, cocaine, LSD, or heroin).
Drove a car after having a few alcoholic drinks.	Drank whiskey, vodka, gin, or other hard liquor.
Had an alcoholic drink before breakfast or instead of breakfast.	
Undependability (7 acts: Mean Interitem $r = .28$; Alpha = .72; Skew = .08)	
Changed or canceled an appointment.	Did not return a phone call.
Arrived at an event more than an hour late.	Broke a promise.
Borrowed something and lost it, broke it, or never returned it.	Misplaced something important (glasses, car keys, etc.).
Let work pile up until just before a deadline.	
Friendliness (8 acts: Mean Interitem $r = .29$; Alpha = .76; Skew = −.24)	
Hugged someone.	Complimented someone.
Started a conversation with strangers.	Apologized to someone.
Made a new friend.	Did a favor for a friend
Shared a problem with a close friend or relative.	Was consulted for help or advice by someone with a personal problem.
Erudition (6 acts: Mean Interitem $r = .31$; Alpha = .72; Skew = −.30)	
Read an entire book in one sitting.	Read in bed before going to sleep.
Read a book.	Bought a book.
Went to a public library.	Had an overdue fine for a movie rental or library book.
Communication (8 acts: Mean Interitem $r = .27$; Alpha = .74; Skew = .33)	
Wrote poetry.	Wrote a thank-you note.
Read poetry.	Wrote a handwritten letter.
Made an entry in a diary or journal.	Put pictures in a photo album.
Wrote a postcard.	Worked on a scrapbook.
Creativity (11 acts: Mean Inter-item $r = .18$; Alpha = .70; Skew = .49)	
Produced a work of art.	Wrote poetry.
Painted a picture (oil, watercolor, pastel, etc.).	Acted in a play.
Took music lessons (voice or instrument).	Asked questions in a meeting or lecture.
Talked in a language other than English.	Played a piano or other instrument.
Sang in or conducted a choir or small ensemble.	Gave a prepared talk or public recital (vocal, instrumental, etc.).
Played in or conducted a band or orchestra.	

STUDY 1: ANALYSES OF BEHAVIORAL ACTS

The Criterion Set of Act Clusters

In our first study, we use self-reported frequencies of various categories of specific activities to evaluate the comparative validity of the 11 inventories. Included in a questionnaire administered during the fall of 1997 was a set of 400 behavioral acts (e.g., Played chess, Shot a gun, Polished my toenails, Gave money to a panhandler, Cut my own hair, Acted in a play, Took a sleeping pill, Rode a bicycle, Bought a book, Signed a petition, Slept past noon), to each of which the research participants reported the frequency with which they had carried out that activity, using the following response options: (1) *Never* in my life. (2) *Not in the past year*. (3) *Once or twice* in the past year. (4) *Three or more times* in the past year, but not more than 15 times (such as once or twice a month). (5) *More than 15 times* in the past year.

To develop a reasonably comprehensive pool of activity descriptors, we began with the 324 acts that had been included in the Objective Behavior Inventory used in the classic study by Loehlin and Nichols (1976). Items that would not be suitable for a heterogeneous adult community sample were either reformulated or omitted, and new items were constructed to tap additional aspects of daily living. Findings from a number of different types of cluster analyses of the responses to these 400 activities led to the development of 60 multi-act clusters of related behaviors.

For the present study, six of these multi-act clusters were selected as likely to be of the most utility as criteria to be predicted by personality inventories. Clusters were selected on the basis of their reliabilities (as assessed by Coefficient Alpha) and their relatively low associations with the demographic variables of gender, age, and educational level. Within these constraints, the set included two clusters of relatively undesirable activities (here labeled Drug use and Undependability), two clusters of relatively desirable activities (Friendliness and Creativity), and two clusters that seem relatively neutral in their desirability (Communication and Erudition). Of the six criteria, five were selected from the most reliable act clusters; in contrast, the criterion of Creativity was developed rationally by selecting those acts that appeared to tap achievement-oriented content.

The specific behaviors included in each of these six act clusters are listed in Table 1, along with the mean intercorrelation of the behaviors in each cluster and its Coefficient Alpha

TABLE 2
Intercorrelations Among the Six Act Clusters, and Between Each Act Cluster and Its Orthogonalized Counterpart (Diagonal), as Well as the Correlations of Each of the Orthogonal Act Clusters With Gender, Age, and Educational Level

	Drug Use	*Undependability*	*Friendliness*	*Erudition*	*Communication*	*Creativity*
Drug Use	.[99]					
Undependability	.12	[.95]				
Friendliness	.02	.20	[.93]			
Erudition	.09	.15	.15	[.95]		
Communication	–.06	.07	.18	.17	[.95]	
Creativity	.09	.16	.17	.16	.19	[.94]
Gender (1=M)	**–.28**	–.10	.16	.17	**.35**	–.11
(2=F)	**(–.27)**	(–.08)	(.18)	**(.22)**	**(.38)**	(–.04)
Age	**–.26**	**–.24**	**–.28**	–.15	.13	–.11
	(–.26)	**(–.25)**	**(–.28)**	(–.15)	(.14)	(–.12)
Education	.12	.14	.02	**.20**	.09	**.33**
	(.07)	(.13)	(.05)	**(.24)**	(.17)	**(.31)**

Note. $N = 759$. Correlations with absolute values greater than .20 are listed in bold. Correlations with absolute values greater than .11 are significant at $p <$ .001. Values in parentheses are partial correlations, corrected for the other two demographic variables.

reliability estimate. The act clusters vary in their size from 6 behaviors (Erudition) to 14 (Drug use), and in their reliabilities from .70 (Creativity) to .89 (Drug Use). Table 2 provides the intercorrelations among the six act clusters. Although Drug use was relatively independent of the other clusters, all of the remaining intercorrelations were positive, suggesting some individual differences in overall activity level. To provide independent measures of each of these behavioral tendencies, the six clusters were factor analyzed, and the orthogonal factor scores from a varimax rotation then were used as criterion variables, along with the original (untransformed) act frequencies. Included in the diagonal of Table 2 are the correlations between the original and the orthogonalized clusters, which range from .93 (Friendliness) to .99 (Drug Use).

Also included in Table 2 are the correlations between each of the six orthogonalized act clusters and the demographic indices of gender, age, and educational level. As would be expected, men and younger persons report more drug-related acts; young persons also report more acts of undependability and unfriendliness; better-educated persons report more creative achievements; and women report higher frequencies of communicative acts.

Even in a heterogeneous sample such as this one, some types of behavioral acts could be of such low base rates that they might suffer from restriction of range. Consequently, we examined the frequency distributions for each of the six act clusters. As would be expected, the frequencies of acts of Drug use and Creativity, and to some extent Undependability, tend to be relatively low, whereas the frequencies of acts of Friendliness and Erudition tend to be relatively high. In all six cases, however, there is a substantial range of individual differences available for predictive purposes. In Table 1, we have included the values of a traditional index of skew as an indicator of the shape of these criterion distributions.

RESULTS

How Predictable Are the Different Act Criteria?

Table 3 presents the mean cross-validity correlations across the 1,000 resampling analyses for both the original and the orthogonalized act criteria and for various sets of scales included in each of the 11 inventories under comparison. Results are provided for scale sets at each of three hierarchical levels—Higher level (3 to 7 scales), Middle level (12 to 16 scales), and Lower level (18 to 45 scales). Within each level, the rows (inventories) are ordered by the size of their mean cross-validity correlations across the six criteria; these mean values are provided in the far right-hand columns.

The column means at the bottom of the table indicate the predictability of each criterion across the inventories for both the original and the orthogonalized act clusters. Invariably, the orthogonalized criteria were less predictable than the original act clusters; these differences ranged from .06 (Drug Use) to .16 (Creativity), and averaged .12. To understand this finding, one must realize that the orthogonal transformation removes aspects of overall activity level from each of the criterion clusters, and this general activity factor is quite predictable from the inventory scales; this is analogous to the use of orthogonalized factors derived from cognitive ability tests where a general factor of intelligence is removed by the orthogonalization process. We will expand on this explanation in our General Discussion later in this article. Interestingly, in both the original and the orthogonalized data the act criterion of Undependability turned out to be the least predictable (.35 versus .38 to .45 for the orthogonalized act clusters and .48 versus .51 to .54 for the original ones).

The Comparative Validity of the Personality Inventories

Perhaps the most interesting feature of Table 3 is the narrow range of cross-validities within sets of scales at each of the three hierarchical levels. The cross-validated multiple correlations tend to be remarkably similar between the best and the worst of the scale sets at each level, the extremes typically differing from each other by a mere .04 or .05. The statistical significance of these differences will be addressed in our General Discussion, but it can be seen that the practical differences between inventories are not substantial. Seemingly, there are scales within each of the 11 personality inventories

TABLE 3
Cross-Validated Multiple Correlation Coefficients for Predicting Each of the Act Clusters from Each of the Personality Inventories.

			Drug Use		*Undependability*		*Friendliness*		*Erudition*		*Communication*		*Creativity*		*Mean*	
Inventory	*k*	*N*	*Orthog*	*Obl.*	*Orthog*	*Obl.*	*Orthog*	*Obl.*	*Orthog*	*Obl.*	*Orthog*	*Obl.*	*Orthog*	*Obl.*	*Orthog*	*Obl.*
Higher-Level Constructs																
NEO Domains	5	680	.44	.51	**.38**	**.50**	.41	.52	.37	.52	**.44**	.52	.39	.53	**.40**	**.52**
16PF Globals	5	633	.41	.49	.30	.44	.46	.55	.42	.55	.42	.52	.39	**.54**	.40	.52
TCI	7	675	**.47**	.52	.35	.49	**.49**	**.58**	.30	.45	.42	**.52**	.38	.52	.40	.51
HPI	7	685	.46	**.52**	.33	.48	.43	.53	.41	.53	.41	.49	.37	.51	.40	.51
6FPQ	6	670	.42	.47	.37	.48	.40	.50	**.46**	**.58**	.40	.50	.35	.51	.40	.51
HEXACO	6	669	.43	.50	.31	.43	.40	.50	.37	.52	.43	.52	**.42**	.55	.39	.50
Big-5 Markers	5	754	.40	.46	.35	.48	.35	.47	.36	.49	.40	.48	.39	.52	.37	.48
CPI Vectors	3	658	.40	.48	.29	.43	.39	.50	.37	.50	.37	.45	.34	.48	.36	.47
Middle-Level Constructs																
JPI-R	15	667	**.50**	**.56**	**.42**	**.54**	.45	.56	.39	.55	.40	**.52**	.37	.54	**.42**	**.54**
16PF	16	633	.48	.54	.33	.44	**.45**	**.56**	**.40**	**.56**	.40	.50	**.39**	**.55**	.41	.53
MPQ	12	685	.43	.51	.32	.45	.42	.52	.36	.49	**.42**	.50	.36	.51	.38	.50
Lower-Level Constructs																
HPI-HICs	44	685	.44	.51	.35	**.51**	.40	.52	**.57**	**.66**	.41	.53	.42	**.58**	**.43**	**.55**
NEO Facets	30	680	**.51**	**.56**	**.38**	.50	.42	.53	.33	.49	**.44**	**.54**	.41	.55	.41	.53
6FPQ	18	670	.39	.46	.37	.49	.38	.49	.53	.63	.39	.51	.38	.53	.41	.52
MPQ	30	685	.48	.54	.36	.50	.41	.51	.32	.49	.41	.53	.34	.53	.39	.51
TCI	29	675	.45	.50	.32	.46	**.50**	**.57**	.27	.43	.40	.49	.38	.52	.39	.50
CPI	33	658	.49	.54	.31	.45	.37	.51	.47	.59	.33	.45	.34	.52	.38	.51
AB5C	45	472	.38	.45	.32	.45	.46	.54	.37	.53	.36	.51	.36	.56	.38	.51
HEXACO	24	669	.43	.49	.31	.44	.37	.48	.33	.49	.40	.50	**.45**	.57	.38	.50
Mean*			**.45**	**.51**	**.35**	**.48**	**.41**	**.52**	**.39**	**.54**	**.39**	**.51**	**.38**	**.54**	**.40**	**.52**

Note. For each criterion, the highest cross-validated multiple correlation coefficient within each hierarchical level is listed in bold. For the individual act clusters, on average, critical differences in r^2 (the square of each table entry) of .09 or more between any two inventories are significant at $p < .05$. For the mean validity coefficients across the act clusters, on average, critical differences of .04 or greater between any two inventories are significant at $p < .05$. These significance values apply to predictions of the orthogonalized criteria. k = number of scales. N = number of subjects. Inventory abbreviations are provided in the text.

that are sufficiently highly related to one or more of the criterion variables so that interinventory differences are minimal. To provide some evidence for this assertion, Table 4 lists those inventory scales that are most highly associated with each of the six orthogonalized act criteria. One cannot help but be struck by both the content-homogeneous patterns of association between inventory scales and behavioral acts and by the degree of redundancy among those prototypical scales from different inventories.

Potential Administration Timing Effects

Although all of the criterion act clusters were assessed from data obtained on a single occasion, the personality inventories were administered over a 10-year period. It is possible that the predictive validity of the inventories may depend on the time difference between its administration and that of the criteria. To check on this possibility, in all three of our studies we use correlation and regression analysis to predict the average cross-validity coefficients from the absolute value of the time difference between the inventory and the criterion administrations; the IPIP-AB5C is excluded from these analyses because its items were administered over several occasions. Based on the cross-validity coefficients displayed in Table 3, the correlation between cross-validity and distance in time from the criterion administration is −.135, which translates into a regression weight of −.004 correlation points per year. The standard error for the regression coefficient is 0.0019. Hence, at 95% confidence, the decay is no more than 0.008 correlation points per year and may be considerably smaller.

DISCUSSION OF STUDY 1

This first in our series of three studies used self-reported behavioral act clusters as criteria, and discovered relatively small differences in the mean cross-validities of the 11 inventories across the six act criteria. On the other hand, we found that orthogonalized criteria were less predictable than the original act clusters, and the act cluster labeled Undependability turned out to be relatively less predictable than the other five criteria, with both types of data. How general are these findings? To answer this question, we turn to some analyses of a quite different type of criteria.

TABLE 4
The Scales With the Highest Correlations with Each of the Orthogonalized Act Clusters

Inventory	*Scale*	*Full Name*	*r*
		Drug Use	
MPQ	TR	Traditionalism	–.44
JPI-R	Tra	Traditional Values	–.40
JPI-R	Ris	Risk-Taking	.38
CPI	Sp	Social Presence	.37
AB5C	II+/I-	Cooperation	–.36
NEO-PI-R	E5	Excitement-Seeking	.34
HPI	PRU	Prudence	–.34
16PF	Factor G	Rule Consciousness	–.34
HPI	SOC	Sociability	.33
JPI-R	Res	Responsibility	–.32
		Undependability	
AB5C	III+/III+	Conscientiousness	–.37
AB5C	II+/I+	Warmth	–.33
JPI-R	Org	Organization	–.32
AB5C	III+/IV+	Organization	–.32
6FPQ	Ord	Order	–.32
6FPQ	MET	Methodicalness	–.31
NEO-PI-R	C	Conscientiousness	–.30
MPQ	UV	Unlikely Virtues	–.30
HEXACO	C: Orga.	Organization	–.30
NEO-PI-R	C5	Self-Discipline	–.29
		Friendliness	
TCI	RD2	Warmth	.40
TCI	RD	Reward Dependence	.40
AB5C	I+/II+	Friendliness	.38
AB5C	II+/I+	Warmth	.38
TCI	RD3	Attachment	.38
16PF	EXT	Extraversion	.37
TCI	C2	Empathy	.36
AB5C	I+/I+	Gregariousness	.34
16PF	Factor A	Warmth	.34
16PF	Factor N	Privateness	–.34
		Erudition	
HPI	Read	Reading	.57
6FPQ	Und.	Understanding	.48
6FPQ	OPN	Openness to Experience	.40
HPI	SCH	School Success	.36
CPI	Ie	Intellectual Efficiency	.36
JPI-R	Com	Complexity	.34
16PF	Factor I	Sensitivity	.34
16PF	Factor G	Tough-Mindedness	–.34
AB5C	V+/V+	Intellect	.32
CPI	Ai	Achievement via Independence	.31
		Communication	
16PF	Factor I	Sensitivity	.38
NEO-PI-R	O2	Openness to Aesthetics	.36
16PF	Factor G	Tough-Mindedness	–.30
AB5C	V+/II+	Reflection	.30
MPQ	AB	Absorption	.29
TCI	ST2	Transpersonal Identification	.28
HEXACO	O: AesA	Aesthetic Appreciation	.28
HPI	Cul	Culture	.28
TCI	ST	Self-Transcendence	.27
HEXACO	E: Sent	Sentimentality	.27
		Creativity	
HEXACO	O Domain	Openness to Experience	.40
HEXACO	O: Crea	Creativity	.38
AB5C	V+/V+	Intellect	.35
AB5C	V+/II−	Creativity	.35
AB5C	V+/III−	Imagination	.33
100 Markers	Factor V	Intellect/Imagination	.33
NEO-PI-R	O Domain	Openness to Experience	.33
HPI	Educ	Education	.33
NEO-PI-R	O2	Openness to Aesthetics	.32
CPI	Em	Empathy	.32

Note. $N = 759$. All correlation coefficients are significant at $p < .001$.

STUDY 2: ANALYSES OF DESCRIPTIONS BY KNOWLEDGEABLE INFORMANTS

Among all of the potential criteria for the comparative evaluation of personality inventories, the most important are measures that go beyond self-reports, and especially those that are based on ratings of the targets' personality traits by persons who know the targets well. Indeed, the developers of at least two of the inventories under comparison (Gough's California Psychological Inventory and the Hogan Personality Inventory) have argued that the primary objective of their measures is to predict individuals' reputations in the eyes of others. As a consequence, in Study 2 we use as our standard of comparison the relative validity of each of the 11 inventories in predicting the ways that our research participants were described by knowledgeable others.

Informant Measures

In the fall of 1998, the participants in the ESCS were requested to help us obtain information about them from three individuals who knew them quite well. The cover letter to our participants included the following instructions:

> Many of you have asked how we can accurately assess your personality characteristics with information provided entirely by you. This mailing is our opportunity to gather information about your personality from people who know you well. This is a very important part of our study, and we hope that you will help us with it. . . . You will find three booklets that we are asking you to pass along to three people who know you *very* well. You can distribute these to your significant other, your friends, your co-workers, even your boss—anyone who knows you *very* well (even if they don't like you, or if you don't like them). . . . To ensure that these are honest evaluations of you, we ask that you not talk to the people completing the questionnaires about their responses.

The ESCS participants were paid for completing the survey themselves, and they were promised double payment if we received responses from their three informants.

Each informant received a four-page booklet, the first page of which was a cover letter providing the name of the ESCS participant whom they were asked to describe, and assuring the informant that his or her "responses to this survey are completely confidential, and *the person you are describing will not see them;*" the informants returned their completed surveys directly to the investigators in stamped preaddressed envelopes. On the last page of the booklet were some questions about the relationship between the informant and the target, how well he or she was known to the informant, how well the person was liked by the informant, and the informant's gender and age.

The middle two pages of the booklet presented the items in two different five-factor instruments. These included all 44 items in John's Big-Five Inventory (BFI: Benet-Martinez & John, 1998; John & Srivastava, 1999) and all 40 items in Saucier's Mini-Markers of the Big-Five factor structure (SMM: Saucier, 1994). The 44 BFI items are verbal phrases (e.g., "Is considerate and kind to almost everyone"), which were listed together on one page, and the SMM items are single adjectives (e.g., "Sympathetic"), which were listed separately on another page. Also included in the survey (but not relevant to the present study) were two items of each type asking about the target's physical attractiveness. The SMM items are a subset of Goldberg's (1992) 100 unipolar Big-Five factor markers, specifically those items that proved to be the most factor univocal across diverse samples of research participants (Saucier, 1994). In contrast to these personality-descriptive adjectives derived from lexical studies, the BFI was developed to reflect some aspects of the Five-Factor model of Costa and McCrae (1992), and its scale labels are those used for the NEO domains.

Table 5 presents the factor loadings of each of the 88 items from a varimax (and thus orthogonal) rotation of six principal components. In this solution, the five personality factors are of roughly equal size, and they correspond identically to their expected locations as Big-Five factor markers. Agreeableness (e.g., Sympathetic and Warm versus Harsh and Rude) includes 18 items most highly associated with it; Conscientiousness (e.g., Organized and Efficient versus Disorganized and Careless) includes 17 items; Extraversion (e.g., Talkative and Bold versus Quiet and Shy) includes 16 items; Neuroticism versus Emotional Stability (e.g., Fretful and Touchy versus Relaxed and Stable) includes 15 items; and Intellect/Openness (e.g., Imaginative and Creative versus Unintellectual) includes 18 items. Finally, the physical attractiveness factor, which is not a focus of the present analyses, includes the remaining four items.

TABLE 5
Factor Structure of the 88 Items Used for the Informant Descriptions

	A	*C*	*E*	*N*	*O*	*PA*
Is considerate and kind to almost everyone	**.76**	.16	.05	–.13	.03	.13
Is helpful and unselfish with others	**.74**	.16	.09	–.09	.00	.11
Sympathetic	**.73**	.04	.07	.16	.15	.05
Has a forgiving nature	**.72**	.00	.03	–.17	.04	.07
Warm	**.71**	.07	.23	.03	.13	.15
Is sometimes rude to others	**–.71**	–.16	.09	.29	.01	-.05
Kind	**.71**	.08	.06	–.04	.06	.03
Harsh	**–.70**	.04	.18	.27	–.01	–.04
Can be cold and aloof	**–.70**	.02	–.23	.17	.04	.02
Likes to cooperate with others	**.69**	.22	.03	–.19	.02	.07
Unsympathetic	**–.68**	–.08	–.15	.01	–.21	–.03
Rude	**–.68**	–.13	.14	.28	–.04	–.08
Cooperative	**.66**	.25	–.01	–.22	.06	.03
Cold	**–.62**	–.06	–.18	.05	–.04	–.01
Starts quarrels with others	**–.61**	–.08	.17	**.37**	–.07	–.04
Tends to find fault with others	**–.59**	.00	.12	**.44**	–.05	.01
Is generally trusting	**.51**	.11	.02	–.22	–.11	.10
Unenvious	.28	.10	–.14	–.19	–.04	–.04
Organized	.00	**.85**	.03	–.04	.03	.06
Tends to be disorganized	–.03	**–.82**	–.01	.03	.07	–.05
Does things efficiently	.12	**.82**	.06	–.13	.05	.15
Disorganized	.02	**–.80**	.01	.07	.01	–.06
Efficient	.05	**.79**	.10	–.08	.07	.12
Inefficient	–.04	**–.76**	–.05	.10	–.06	–.06
Does a thorough job	.14	**.72**	.03	–.03	.06	.01
Makes plans and follows through with them	.09	**.71**	.11	–.13	.11	.07
Can be somewhat careless	–.20	**–.70**	.11	.06	.01	–.04
Systematic	–.07	**.70**	–.02	–.14	.10	–.04
Sloppy	–.20	**–.68**	–.04	–.04	.02	–.15
Perseveres until the task is finished	.14	**.66**	.09	–.08	.04	.12
Careless	–.24	**–.60**	.13	.05	–.08	.00
Is easily distracted	–.02	**–.58**	.11	**.33**	–.01	–.05
Is a reliable worker	.09	**.52**	.08	–.18	.00	.14
Practical	.20	**.51**	–.04	–.25	–.02	–.04
Tends to be lazy	–.22	**–.43**	–.24	.15	–.12	–.15
Tends to be quiet	.03	.08	**–.85**	–.13	.04	.00
Quiet	.03	.10	**–.82**	–.11	.07	.00
Is sometimes shy, inhibited	.03	–.07	**–.76**	.20	–.02	.06
Has an assertive personality	–.13	.18	**.76**	–.06	.21	.04
Extraverted	.07	–.02	**.76**	–.03	.09	.06
Shy	.07	–.04	**–.75**	.19	–.05	.05
Is outgoing, sociable	**.30**	.08	**.74**	–.08	.09	.11
Is talkative	.14	–.10	**.74**	.23	.07	–.05
Talkative	.10	–.08	**.73**	.24	.06	–.04
Is reserved	–.02	.09	**–.73**	–.04	.01	.01
Bashful	.08	–.07	**–.66**	.18	–.08	.10
Withdrawn	–.29	–.16	**–.65**	.21	–.01	.02
Bold	–.24	.05	**.64**	–.07	.19	.07
Generates a lot of enthusiasm	**.30**	.12	**.60**	–.05	.26	.26
Is full of energy	.14	.27	**.48**	–.22	.13	.27
Energetic	.11	**.33**	**.43**	–.19	.16	.30
Gets nervous easily	.00	–.09	–.14	**.79**	–.07	–.03
Worries a lot	.02	–.08	–.17	**.78**	.01	–.03
Is relaxed, handles stress well	.19	.17	.05	**–.76**	.06	.08
Fretful	–.01	–.12	–.12	**.76**	–.09	–.07
Remains calm in tense situations	.19	.23	.01	**–.71**	.13	.10
Is emotionally stable, not easily upset	.29	.20	.02	**–.71**	.02	.09
Can be tense	–.25	.00	.00	**.68**	.03	.00

(*Continued*)

The correlations between each of these orthogonalized personality factors and the three demographic variables of gender, age, and educational level are provided in Table 6, along with the residual correlations with the other two demographic variables partialed out. Women tended to be

TABLE 5
Factor Structure of the 88 Items Used for the Informant Descriptions (Continued)

	A	*C*	*E*	*N*	*O*	*PA*
Relaxed	.22	.10	−.11	**−.66**	.02	.04
Is depressed, blue	−.23	−.17	−.26	**.63**	.10	−.06
Touchy	**−.42**	−.16	.01	**.61**	−.03	.02
Moody	**−.44**	−.12	−.16	**.59**	.01	.03
Can be moody	**−.45**	−.12	−.15	**.58**	−.01	.00
Temperamental	**−.49**	−.08	.10	**.56**	−.01	.05
Jealous	**−.35**	−.13	.13	**.43**	.02	.01
Envious	**−.33**	−.12	.11	**.42**	−.02	.04
Has an active imagination	.05	−.13	.18	.03	**.75**	.07
Imaginative	.02	−.05	.16	.03	**.75**	.19
Likes to reflect, play with ideas	.03	.07	.13	−.10	**.74**	.01
Creative	.05	.01	.11	.10	**.70**	.19
Deep	.01	.07	−.04	.00	**.70**	−.02
Is sophisticated in art, music, literature	.10	.02	−.02	.10	**.70**	.00
Is ingenious, a deep thinker	−.11	.17	−.08	−.12	**.69**	−.07
Is original, comes up with new ideas	−.01	.13	.28	−.14	**.69**	.17
Is curious about many different things	.09	−.01	.12	−.06	**.68**	.01
Is inventive	.02	.11	.15	−.18	**.66**	.13
Values artistic, aesthetic experiences	.25	.02	.06	.22	**.66**	.08
Uncreative	−.08	−.06	−.13	.01	**−.63**	−.22
Philosophical	.12	.06	−.06	−.09	**.63**	−.12
Intellectual	−.07	.16	−.14	−.12	**.60**	−.13
Has few artistic interests	−.14	.05	−.03	−.20	**−.60**	−.10
Unintellectual	−.02	−.14	.04	.15	**−.55**	.14
Complex	−.29	−.06	−.05	.12	**.47**	−.05
Prefers work that is routine	.03	.21	−.26	.13	**−.47**	−.07
Good-Looking	.10	.20	.02	−.04	.07	**.85**
Is physically attractive	.14	.22	.03	−.07	.11	**.83**
Is not good-looking	−.12	−.23	.02	.05	−.08	**−.80**
Unattractive	−.19	−.25	−.01	.10	−.13	**−.78**

Note. A = Agreeableness; C = Conscientiousness; E = Extraversion; N = Neuroticism; O = Intellect/Openness; PA = Physical Attractiveness. Correlations with absolute value greater than .30 are listed in bold. The single words are from the Saucier Mini-Markers, and the phrases are from the Big-Five Inventory. The size of the first 10 eigenvalues in this solution were 17.5, 9.7, 7.3, 6.7, 4.8; 2.7, 2.1, 1.8, 1.7, and 1.5. The proportion of variance explained by each of the rotated components above are 11.8, 10.8, 10.1, 9.4, 9.3, and 3.9.

described by their informants as more Agreeable and more Neurotic than men. And, as would be expected, individuals with more education tended to be described as higher on the Intellect/Openness factor.

RESULTS

How Predictable are the Different Big-Five Criteria?

Table 7 presents the mean cross-validity correlations across the 1,000 resampling analyses for sets of scales at each of the three hierarchical levels from each of the 11 inventories. Included in the bottom row of the table are the mean values across all of the scale sets under study, thus providing information about the relative predictability of each of the Big-Five factors when based on descriptions by knowledgeable informants. Clearly, these informant-based measures of the five factors differ substantially in their predictability by self-reports: By far the most predictable criterion was Extraversion, with an average cross-validity of .57, and the least predictable was Conscientiousness with an average cross-validity of only .39. In between the two extremes were

TABLE 6
Correlations of Each of the Informant Criteria With Gender, Age, and Educational Level

	A	*C*	*E*	*N*	*O*	*PA*
Gender (1=M)	**.29**	.05	−.15	**.34**	.02	.10
(2=F)	**(.29)**	(.05)	(−.16)	**(.31)**	(.09)	(.08)
Age	.17	.03	.03	−.04	.05	−.12
	(.17)	(.03)	(.03)	(−.05)	(.05)	(−.12)
Education	−.03	.01	−.01	−.19	**.32**	−.15
	(.04)	(.02)	(−.04)	(−.13)	**(.33)**	(−.13)

Note. Correlations with absolute values greater than .20 are listed in bold. Values in parentheses are partial correlations, corrected for the other two demographic variables.

TABLE 7
Cross-Validated Multiple Correlation Coefficients for Predicting Each of the Informant Factors From Each of the Personality Inventories

Inventory	*k*	*n*	*A*	*C*	*E*	*N*	*O*	*Mean*
Higher-Level Constructs								
Big-5 Markers	5	587	.49	.49	**.66**	.47	**.56**	**.53**
HEXACO Domain	6	541	.48	**.50**	.65	.47	.52	.52
NEO Domains	5	542	**.49**	.45	.57	**.51**	.54	.51
16PF Globals	5	494	.46	.41	.57	.44	.53	.48
6FPQ	6	532	.41	.38	.55	.39	.49	.44
HPI	7	542	.45	.26	.52	.48	.45	.43
TCI	7	533	.41	.30	.50	.48	.43	.42
CPI Vectors	3	523	.35	.27	.49	.35	.40	.37
Middle-Level Constructs								
JPI-R	15	551	.39	**.43**	**.59**	**.48**	**.54**	**.49**
16PF	16	494	**.42**	.41	.58	.45	.52	.48
MPQ	12	554	.41	.30	.55	.46	.47	.44
Lower-Level Constructs								
HEXACO	24	541	.49	.48	**.63**	.47	.55	**.52**
AB5C	45	378	.49	**.48**	.57	.48	**.57**	**.52**
NEO Facets	30	542	**.50**	.46	.57	**.51**	.50	.51
6FPQ	18	532	.45	.41	.54	.42	.48	.46
MPQ	30	554	.40	.26	.54	.44	.45	.42
TCI	29	533	.41	.30	.55	.46	.40	.42
HPI-HICs	44	542	.45	.25	.50	.44	.43	.42
CPI	33	523	.40	.29	.50	.43	.44	.41
Mean			**.44**	**.39**	**.57**	**.46**	**.49**	**.44**

Note. For each criterion, the highest cross-validated multiple correlation coefficient within each hierarchical level is listed in bold. For each of the individual informant factors, on average, critical differences in r^2 (the square of each table entry) of .11 or higher between any two inventories are significant at $p < .05$. For the mean validity coefficients across all factors, on average, critical differences in r^2 of .05 or greater between any two inventories are significant at $p < .05$. k = number of scales. n = number of subjects. Inventory abbreviations are provided in the text.

Intellect/Openness (.49), Neuroticism (.46), and Agreeableness (.44). The relatively low predictability of Conscientiousness is reminiscent of the finding from Study 1, where the behavioral act cluster labeled Undependability (which is a common label for the negative pole of the Conscientiousness factor) was the least predictable of those six criteria.

The Comparative Validity of the Personality Inventories

Table 7 reveals that the strongest predictions of these informant-based criteria were from Goldberg's (1992) Big-Five factor markers, the 6 HEXACO composite scores, and the NEO-PI-R domain scores—all at the highest hierarchical level—as well as the 24 HEXACO, the 45 AB5C, and the 30 NEO-PI-R facet scales—at the lowest level. All of these six scale sets produced cross-validities in the narrow range from .51 to .53. And, at the other extreme, the scales with the lowest relative predictability were the 3 CPI vectors (.37) and 33 CPI scales (.41), a paradoxical finding given that some of the CPI scales (but not the CPI structural vectors) were explicitly developed to predict informant-based criteria. On the other hand, the finding becomes less paradoxical when one takes into account the degree of content overlap between the informant-based Big-Five criterion factors and the corresponding self-report versions.

Table 8 lists those inventory scales that were most highly associated with each of the six informant factors (the five personality traits, plus—for fun—the observer-based physical-attractiveness factor). Again, as in Study 1, one cannot help but be struck by both the content-homogeneous patterns of association between inventory scales and informant descriptions and by the degree of redundancy among those prototypical scales from different inventories.

Potential Inventory-Criterion Timing Effects

Based on the cross-validity coefficients presented in Table 7, the correlation between inventory validity and distance from the time that the criteria were administered was .29, which translates into a regression weight of .0027 correlation points per year (*SE* = .006). In this study (unlike Study 1) the timing correlation was positive, rather than negative as would be predicted, however, suggesting that the further in time from the criteria that the inventories were administered the more valid they were. Given that there seems to be no strong rationale for such an effect, it suggests that any such relation between inventory validity and administration time is likely to be spurious.

DISCUSSION OF STUDY 2

In our second study in this series, we used informant reports as criteria and found substantial differences in the predictability of the five informant-based personality factors: Extraver-

TABLE 8
The Scales With the Highest Correlations With Each of the Informant Factors

Inventory	*Scale*	*Scale Name*	*r*
		Agreeableness	
NEO-PI-R	A	Agreeableness domain	.47
AB5C	II+/IV+	Empathy	.45
AB5C	II+/V-	Nurturance	.43
100 Markers	Factor II	Agreeableness	.42
AB5C	II+/III-	Sympathy	.42
HEXACO	A:Gent	Gentleness	.42
NEO-PI-R	A4	Compliance	.41
AB5C	II+/I+	Warmth	.41
HPI	LIK	Likeability	.38
CPI	B-FM	Femininity	.38
		Conscientiousness	
AB5C	III+/III+	Conscientiousness	.51
NEO-PI-R	C2	Order	.50
100 Markers	Factor III	Conscientiousness	.50
AB5C	III+/V-	Orderliness	.50
HEXACO	C: Orga	Organization	.49
HEXACO	C Domain	Conscientiousness	.47
JPI-R	Org	Organization	.45
16PF-Global	SC/U	Self-Control	.45
NEO-PI-R	C Domain	Conscientiousness	.44
6FPQ	Ord	Order	.44
		Extraversion	
100 Markers	Factor I	Extraversion	.67
HEXACO	EXT	Extraversion	.63
AB5C	I+/I+	Extraversion	.59
HEXACO	X: Expr	Expressiveness	.56
JPI-R	Soc	Social Confidence	.56
NEO-PI-R	E Domain	Extraversion	.55
16PF	Factor H	Social Boldness	.55
6FPQ	Ext	Extraversion	.53
HEXACO	X:Soci	Sociability	.52
6FPQ	Exh	Exhibition	.51
		Neuroticism	
AB5C	IV+/IV+	Stability	–.48
HEXACO	E Domain	Emotionality	.47
JPI-R	Anx	Anxiety	.45
NEO-PI-R	N Domain	Neuroticism	.45
AB5C	IV+/V-	Tranquility	–.45
NEO-PI-R	N1	Anxiety	.44
HEXACO	E: Anxi	Anxiety	.43
AB5C	IV+/II−	Imperturbability	–.42
HPI	ADJ	Adjustment	–.42
AB5C	IV+/V+	Toughness	–.42
		Openness/Intellect	
AB5C	V+/III−	Imagination	.53
HEXACO	O Domain	Openness	.51
100 Markers	Factor V	Intellect/Imagination	.51
NEO-PI-R	O Domain	Openness to Experience	.50
AB5C	V+/V+	Intellect	.47
6FPQ	OPN	Openness to Experience	.47
16PF	Factor Q1	Openness to Change	.46
JPI-R	Inn	Innovation	.46
16PF	T-M	Tough-Mindedness	–.46
HEXACO	O:Crea	Creativity	.45
		Physical Attractiveness	
MPQ	WB1	Cheerfulness	.21
MPQ	WB	Well-being	.21
16PF	Factor F	Liveliness	.19
NEO-PI-R	E Domain	Extraversion	.19
NEO-PI-R	E4	Activity	.18
NEO-PI-R	E6	Positive Emotions	.18
AB5C	III+/I−	Efficiency	–.18
AB5C	II+/I+	Warmth	.18
AB5C	II+/IV-	Tenderness	.18
NEO-PI-R	E5	Excitement-seeking	.17

Note. All correlation coefficients are significant at $p < .001$.

sion turned out to be the most predictable, and Conscientiousness the least predictable, of these criterion dimensions. Differences in average cross-validities among the inventories were not substantial, with one exception: The CPI scales at both the highest and lowest levels turned out to be less valid predictors than the other inventories, perhaps because the CPI scales and its higher-level vectors do not reflect Big-Five factor variance to the same degree as most of the other inventories under comparison. As with Study 1, the statistical significance of these differences will be addressed in our general Discussion.

STUDY 3: ANALYSES OF ABNORMAL TENDENCIES

Associations between personality and mental health have long been under investigation. Indeed, constructs such as Neuroticism have been studied as both indicators of mental health and as measures of normal individual differences (Woodworth, 1919). Moreover, the association is not merely tautological; longitudinal studies have shown that some personality traits are premorbid risk factors for substance abuse, antisocial behavior, and other mental health problems (e.g., Cloninger, Sigvardsson, & Bohman, 1988; Krueger, 1999; Tremblay, Pihl, Vitaro, & Dobkin, 1994). Twin and family studies have suggested that common genetic factors may influence both personality measurements and psychiatric outcomes, bolstering the argument that measures of normal personality are of clinical interest (e.g., Fanous, Gardner, Prescott, Cancro, & Kendler, 2002; Farmer, Mahmood, Redman, Harris, Sadler, & McGuffin, 2003; Krueger, Hicks, Patrick, Carlson, Iacono, & McGue, 2002).

The objective of Study 3 was to compare broad-bandwidth personality inventories developed to assess normal traits in their validity as predictors of criteria related to mental health.[3] From the total set of psychological measures, we chose the six that were most likely to be of clinical interest. These scales all have been associated with psychiatric or personality disorders as direct measures, screening instruments, or theoretically conceived premorbid indicators of susceptibility. We will refer to these constructs collectively as "clinical indicators."

The Selection of Clinical Indicators

Unlike Studies 1 and 2, the criterion measures chosen for Study 3 were not all administered on the same occasion. The ESCS measures of clinical interest that were chosen as criteria are as follows:

a. Borderline Personality Inventory (BPI: Leichsenring, 1999). In the first report on the development of this instrument, the BPI showed acceptable sensitivity and specificity against clinical diagnosis of borderline personality disorder.
b. Levenson Self-Report Psychopathy Scale (LSRP: Levenson, Kiehl, & Fitzpatrick, 1995). Developed to assess a "protopsychopathic interpersonal philosophy," the LSRP has shown strong relations with self-reported measures of serious antisocial behaviors and with psychophysiological measures known to characterize psychopathic offenders (Lynam, Whiteside, & Jones, 1999). In a study of prisoners, the correlation between LSRP scores and violent criminal activity was comparable with, or slightly better than, that for the widely used Hare Psychopathy Checklist-Revised (Brinkley, Schmitt, Smith, & Newman, 2001).
c. Magical Ideation Scale (MIS: Eckblad & Chapman, 1983). Magical ideation is defined as beliefs in forms of causation that by conventional standards are invalid. Subjects with high MIS scores have been shown to exhibit more psychotic and psychotic-like experiences during an initial interview and also at a 10-year follow-up (Chapman, Chapman, Kwapil, Eckblad, & Zinser, 1994); moreover, subjects who scored high on the MIS and on the highly correlated Perceptual Aberration Scale (Chapman, Chapman, & Raulin, 1978) were characterized by symptoms of schizotypal personality disorder and borderline personality disorder.
d. Center for Epidemiological Studies Depression Scale (CES-D: Radloff, 1977). The CES-D is a classic measure of depressive symptoms. Widely used as a screen or a proxy for major depressive disorder, CES-D scores are also correlated with generalized anxiety disorder, depressive personality disorder, and other depression-related psychiatric conditions (Breslau, 1985; Myers & Weissman, 1980; Roberts & Vernon, 1983). Hence, the CES-D is best thought of as a well-characterized measure of subjective dysphoria and psychological distress, as well as a fallible indicator of more serious clinical depression.
e. Goldberg's (1999b) Curious Experiences Survey, a revised version of the Dissociative Experiences Scale (DES: Bernstein & Putnam, 1986). The DES is a thoroughly studied and highly reliable measure of dissociative experiences. Although a few dissociative experiences are relatively common in normal populations, most of them are prominent features of dissociative disorders, as well as post-traumatic stress disorders, and they are correlated with other psychotic-like symptoms (Bremner et al., 1992; Moskowitz, Barker-Collo, & Ellson, 2005; Putnam et al., 1996). The revision by Goldberg (1999b) upgrades the psychometric characteristics of this classic measure.

[3] Obviously, in predicting these clinical criteria, instruments developed specifically to measure various aspects of psychopathology, such as the Minnesota Multiphasic Personality Inventory (MMPI), should have an advantage over the broad-bandwidth inventories here under comparison. On the other hand, we assume that the MMPI and other such inventories may not be particularly appropriate for use in many nonclinical settings.

TABLE 9
Intercorrelations Among the Six Clinical Indicators, and the Correlations of Each of the Criteria With Gender, Age, and Educational Level

	LSRP	*MIS*	*OCI*	*CES-D*	*DES*	*BPI*
Sociopathy (LSRP) [24 items; $\alpha = .82$; skew = .55]						
Magical Ideation (MIS) [28 items; $\alpha = .93$; skew = 1.06]	.33					
Obsessive-Compulsive (OCI) [16 items; $\alpha = .88$; skew = .60]	.30	.27				
Depression (CES-D) [24 items; $\alpha = .93$; skew = 1.47]	.21	.22	.35			
Dissociation (DES) [31 items; $\alpha = .91$; skew = 2.27]	.23	**.52**	.33	**.40**		
Borderline Personality (BPI) [47 items; $\alpha = .91$; skew = 1.11]	**.58**	**.63**	**.46**	**.45**	**.58**	
Gender (1 = M; 2 = F)	**–.26** **(–.28)**	.02 (.01)	–.10 (–.14)	.13 (.11)	–.03 (–.02)	–.10 (–.13)
Age	–.07 (–.06)	.03 (–.03)	.09 (.09)	–.12 (–.13)	–.15 (–.15)	–.08 (–.08)
Education	–.07 (–.13)	–.03 (–.03)	–.16 (–.19)	–.17 (–.13)	.00 (.00)	–.13 (–.16)

Note. $N = 633$. Intercorrelations among the clinical indicators with absolute values greater than .40 are listed in bold, and correlations with the demographic variables with absolute values greater than .20 are also listed in bold. Values in parentheses are partial correlations, corrected for the other two demographic variables.

f. Obsessive-Compulsive Inventory (OCI: Foa, Kozak, Salkovskis, Coles, & Amir, 1998). The OCI includes items tapping the major symptoms of obsessive-compulsive disorder, and it is intended for diagnostic screening, symptom profiling, and severity determination. The measure exhibited strong convergent validity with other such instruments, both in an initial validity study and in non-clinical samples (Hajcak, Huppert, Simons, & Foa, 2004).

The CES-D and the revised DES were administered to the ESCS participants concurrently with the Behavioral Report Inventory (used as criteria in Study 1) in 1997. The OCI was administered as part of a Comprehensive Health Survey in 1999. The BPI, LSRP and MIS were administered as part of a survey of Personality, Emotions, and Attitudes in 2000.

Table 9 provides the number of items in each of the clinical indicators (ranging from 16 [OCI] to 47 [BPI]), their coefficient alpha reliability estimates (ranging from .82 [LSRP] to .93 [MIS and CES-D]), and their skewness coefficients (ranging from .60 [OCI] to 2.27 [DES]). In addition, Table 9 includes the intercorrelations among the indicators. As would be expected for measures related to psychopathology, there were a number of moderate-to-large correlations among the six criteria, and the BPI was highly related to all of the other five indicators. Also included in Table 9 are the correlations between each of the six clinical indicators and the demographic indices of gender, age, and educational level. Most of these correlations were quite small, with one unsurprising exception: Men were more likely than women to obtain high scores on the LSRP measure of antisocial tendencies.

Because we are using clinical indicators in an essentially normal community sample, some of the distributions could be so highly skewed that they might suffer from restriction of range. As shown in Table 9, however, only two distributions exhibited skewness coefficients markedly above 1.0—Depression (1.47) and Dissociation (2.27). The remaining four coefficients ranged from .55 to 1.11, indicating that these distributions were reasonably normal in shape. In all six cases, however, there was a substantial range of individual differences available for predictive purposes.

RESULTS

How Predictable Are the Different Criteria of Abnormal Tendencies?

Table 10 presents the mean cross-validity correlations across the 1,000 resampling analyses for sets of scales at each of the three hierarchical levels from each of the 11 inventories. In the bottom row of this table are the mean values across the scale sets for each of the six clinical indicators. As in Study 2, the six criteria differed substantially in their overall predictability. The indicators of Sociopathy (LSRP) and Borderline Personality (BPI) were the most predictable criteria (.56 and .51, respectively); at the other extreme, Magical Ideation (MIS) and Dissociation (DES) were the least predictable (.24 and .33, respectively).

The Comparative Validity of the Personality Inventories

There was substantial variation among the inventories in their prediction of this set of criteria. Two inventories, the MPQ and the TCI, were unusually strong predictors of the clinical indicators, in large part because of their prowess in the prediction of magical ideation and their relatively strong validity in predicting dissociative experiences. Another factor contributing to the wide range of cross-validities was poor performance of the 6FPQ; at either hierarchical level, this was the least predictive set of constructs across all six criteria. Table 11 lists the scales with the highest correlations with each of the six clinical indicators. Of the 10 highest correlating scales with magical ideation, 7 were from the TCI or the MPQ, and all correlations with absolute values above .30 were from the TCI or the MPQ. For dissociation, there were fewer strong correlates, but among the 10 highest correlating

TABLE 10
Cross-Validated Multiple Correlation Coefficients for Predicting Each of the Clinical Indicators From Each of the Personality Inventories

Inventory	*k*	*N*	*LSRP*	*MIS*	*OCI*	*CES-D*	*DES*	*BPI*	*Mean*
Higher-Level Constructs									
TCI	7	583	**.68**	**.48**	.50	.46	**.41**	**.64**	**.53**
NEO Domains	5	579	.60	.25	.49	.48	.33	.53	.45
HPI	7	591	.54	.24	.48	**.49**	.35	.54	.44
CPI Vectors	3	561	.53	.27	**.51**	.39	.30	.50	.42
16PF Globals	5	548	.57	.22	.43	.40	.33	.50	.41
HEXACO	6	593	.64	.26	.37	.33	.23	.40	.37
Big-5 Markers	5	630	.44	.15	.45	.40	.29	.44	.36
6FPQ	6	585	.47	.00	.34	.28	.22	.34	.27
Middle-Level Constructs									
MPQ	12	604	.53	**.50**	**.53**	**.54**	**.47**	**.64**	**.53**
16PF	16	548	.55	.23	.48	.47	.35	.51	.43
JPI-R	15	602	**.58**	.26	.44	.41	.31	.46	.41
Lower-Level Constructs									
MPQ	30	604	.53	**.50**	.49	**.52**	**.48**	**.60**	**.52**
TCI	29	583	**.66**	.44	.46	.46	.35	.59	.50
CPI	33	561	.59	.27	**.56**	.47	.36	.58	.47
AB5C	45	420	.65	.12	.55	.43	.38	.55	.45
NEO Facets	30	579	.58	.19	.42	.45	.32	.53	.41
HEXACO	24	593	.63	.25	.43	.41	.30	.44	.41
HPI	44	591	.51	.20	.43	.53	.25	.55	.41
6FPQ	18	585	.48	.00	.36	.29	.23	.39	.29
Mean*			**.56**	**.24**	**.46**	**.44**	**.33**	**.51**	**.42**

Note. For each criterion, the highest cross-validated multiple correlation coefficient within each hierarchical level is listed in bold. For individual criteria, on average, critical differences in r^2 of .09 or greater between any two inventories are significant at $p < .05$. For the mean validity coefficients across the criteria, on average, critical differences in r^2 of .04 or greater between any two inventories are significant at $p < .05$. k = number of scales. N = number of subjects. Inventory abbreviations are provided in the text.

scales, two were from the MPQ, one from the TCI, and two from the AB5C.

Potential Inventory-Criterion Timing Effects

The correlation between the cross-validity coefficients presented in Table 10 and the corresponding time differences between inventory and criterion administration were remarkably low: −0.0008, which corresponds to a regression coefficient of −0.000057 regression points per year. The standard error of the regression coefficient is .01 and so, even if the nearly-zero correlation is coincidental, it is highly unlikely that true decay is more than .02 points per year, which is the upper 95% confidence limit. Coupled with the near zero timing effects reported for the other two studies, the decay is probably much smaller and may be negligible.

Relations Between the Criteria From the Three Different Studies

One final finding is included in Table 12, which provides the correlations between the criteria across our three studies,

TABLE 11
The Scales With the Highest Correlations With Each of the Clinical Indicators

Inventory	*Scale*	*Full Name*	*r*
Sociopathy (LSRP)			
TCI	C	Cooperativeness	–.64
HEXACO	H: Fair	Fairness	–.56
HEXACO	H Domain	Honesty/Humility	–.56
TCI	SD	Self-Directedness	–.54
TCI	C5	Conscience	–.54
AB5C	II+/II+	Understanding	–.54
NEO-PI-R	A Domain	Agreeableness	–.53
AB5C	II+/I-	Cooperation	–.53
AB5C	II+/III+	Morality	–.52
CPI	Ami	Amicability	–.52
Magical Ideation (MIS)			
MPQ	AB2	Imaginative/Altered states	.48
MPQ	AB	Absorption	.46
TCI	ST1	Self-forgetfulness	.41
TCI	ST	Self-transcendence	.40
TCI	ST2	Transpersonal Identification	.37
MPQ	AB1	Sentient	.36
CPI	WO	Work Orientation	–.32
CPI	Sc	Self-control	–.32
CPI	Ami	Amicability	–.29
MPQ	CO2	Planfulness	–.28
Obsessive-Compulsive (OCI)			
CPI	V 3	Vector 3	–.52
AB5C	IV+/I+	Impulse Control	–.47
CPI	Ie	Intellectual Efficiency	–.47
AB5C	IV+/V+	Toughness	–.44
MPQ	SR	Stress Reaction	.44
CPI	Py	Psychological-Mindedness	–.43
CPI	Ai	Achievement via Independence	–.42
CPI	Wb	Well-being	–.42
16PF	Anx	Anxiety	.42
CPI	Mp	Managerial Potential	–.41
Depression (CES-D)			
HPI	Depr.	No depression	–.52
HPI	ADJ	Adjustment	–.50
NEO-PI-R	N Domain	Neuroticism	.49
NEO-PI-R	N3	Depression	.49
16PF	Factor C	Emotional Stability	–.49
AB5C	IV+/I+	Impulse Control	–.49
CPI	Wb	Well-being	–.48
MPQ	SR	Stress Reaction	.48
MPQ	SR1	Negative Emotions	.45
CPI	LP	Leadership Potential	–.45
Dissociation (DES)			
MPQ	AB2	Imaginative/Altered states	.41
MPQ	AB	Absorption	.38
AB5C	III+/II+	Dutifulness	–.38
CPI	WO	Work Orientation	–.38
CPI	Sc	Self-control	–.38
CPI	Gi	Good Impression	–.37
CPI	Ami	Amicability	–.37
TCI	ST1	Self-forgetfulness	.36
AB5C	V+/II+	Reflection	–.35
CPI	Wb	Well-being	–.35
Borderline Personality (BPI)			
CPI	WO	Work Orientation	–.54
TCI	SD	Self-Directedness	–.54
CPI	Ami	Amicability	–.53
NEO-PIR	N Domain	Neuroticism	.52
CPI	Wb	Well Being	–.52
NEO-PI-R	N3	Depression	.52
CPI	MP	Managerial Potential	–.50
HPI	ADJ	Adjustment	–.49
AB5C	IV+/III+	Moderation	–.48
CPI	Lp	Leadership Potential	–.48

Note. $N = 604$. All correlations coefficients are significant at $p < .001$.

TABLE 12
Criterion Correlations Across the Three Studies

	Study One						*Study Two*				
	DU	*Un*	*Fr*	*Er*	*Com*	*Crea*	*A*	*C*	*E*	*N*	*O*
Study Two											
A	**−.26**	.04	.08	.06	**.25**	−.13					
C	−.05	−.17	−.01	−.06	.07	−.03					
E	.05	−.03	**.26**	.00	.15	.03					
N	−.04	.00	.14	.03	.13	−.02					
O	.02	.06	.07	**.29**	.13	**.35**					
Study Three											
LSRP	**.20**	**.20**	−.18	−.10	**−.21**	−.08	**−.34**	−.08	−.06	.07	−.11
MIS	.17	.05	.00	−.08	.09	.09	−.08	−.04	.01	.10	.08
OCI	−.02	.14	−.06	**−.24**	−.01	−.08	−.07	.02	−.06	**.21**	−.18
CES-D	.00	**.23**	−.01	−.03	.04	−.02	.00	−.13	−.14	**.33**	−.04
DES	.10	**.24**	.09	−.03	.06	.18	−.12	−.13	−.04	.18	.10
BPI	.05	**.22**	−.04	−.12	−.04	.05	−.15	−.12	−.13	**.21**	.01

Note. Correlations with absolute values of .20 or greater are listed in bold. DU = Drug Use, Un = Undependability, Fr = Friendliness, Er = Erudition, Com = Communication, Crea = Creativity (Study 1); A = Agreeableness, C = Conscientiousness, E = Extraversion, N = Neuroticism, O = Intellect/Openness (Study 2); LSRP = Levenson Sociopathy, MIS = Magical Ideation Scale, OCI = Obsessive-Compulsive Inventory, CES-D = CES Depression Scale, DES = Dissociation, BPI = Borderline Personality Inventory (Study 3).

showing that in general the three types of criteria were relatively independent of one another. Among the highest cross-study criterion correlations were those between informant-rated Intellect/Openness (Study 2) and the frequency of acts involving Creativity ($r = .35$) and Erudition ($r = .29$) in Study 1. Other relatively high cross-study correlations were between informant-based ratings of Agreeableness and low scores on LSRP sociopathy ($r = -.34$), and between informant-rated Extraversion and the frequency of acts of Friendliness ($r = .26$). Finally, there was a correlation of .33 between informant-based ratings of Neuroticism in Study 2 and the CES-D Depression criterion from Study 3. Interestingly, there were no correlations greater than .13 between informant-rated Conscientiousness and any of the clinical indicators, and there were no correlations higher than .17 between the MIS and any of the criteria used in Studies 1 and 2.

DISCUSSION OF STUDY 3

The TCI and the MPQ were among the best-predicting inventories for all criteria examined in study 3, but their advantages were much stronger for some criteria than others; they stand out most in the prediction of magical ideation and, to a lesser degree, dissociation. The scales from the TCI and MPQ that correlated most strongly with these criteria were TCI Self-Transcendence, MPQ Absorption, and the lower-level constructs that contribute to these domains. Interestingly, the Self-Transcendence construct incorporates elements of Absorption (Cloninger, Svrakic, & Przybeck, 1993), and these scales are strongly intercorrelated with each other ($r = .59$).

GENERAL DISCUSSION

The Relative Predictability of Each of the Three Sets of Criteria

Are any of those criterion sets substantially more predictable from the 11 personality inventories than the others? In our first study, we used for criteria six clusters of behavioral acts; in our second study we used five broad traits as assessed by knowledgeable informants; and in our last study we used six clinical indicators that may have psychopathological implications. Surprisingly, across the 11 inventories under comparison, the three quite diverse classes of criteria did not differ much in their overall predictability, with mean cross-validity correlations ranging from .40 for the orthogonalized act clusters in Study 1 to .44 for the five traits assessed by informants in Study 2.

Rather, the major differences among the three sets of criteria stemmed from the variance in their predictability within each set. The act clusters from Study 1 turned out to be the least differentially predictable, with only the cluster labeled Undependability slightly less predictable than the other five. At the other extreme were the clinical indicators, with a substantial difference in predictability between the index of Magical Ideation (.24) and that of Sociopathy (.56). In between the two extremes were the Big-Five factors as assessed by informant reports, with the five traits varying from .39 for Conscientiousness to .57 for Extraversion.

Finally, in our Discussion of Study 1 we noted that the orthogonalized act-frequency criteria were generally less predictable than the original somewhat-intercorrelated act

TABLE 13
Mean Cross-Validity Coefficients for All Inventories, Across Studies

					Across-Study
Inventory	*k*	*Study 1*	*Study 2*	*Study 3*	*Mean*
		Higher-Level Constructs			
NEO Domains	5	**.40**	.51	.45	.45
TCI	7	.40	.42	**.53**	.45
16PF Globals	5	.40	.48	.41	.43
HEXACO	6	.39	.52	.37	.43
HPI	7	.40	.43	.44	.42
Big-5 Markers	5	.37	**.53**	.36	.42
CPI Vectors	3	.36	.37	.42	.38
6FPQ	6	.40	.44	.27	.37
		Middle-Level Constructs			
MPQ	12	.38	.44	**.53**	.45
JPI-R	15	**.42**	**.49**	.41	.44
16PF	16	.41	.48	.43	.44
		Lower-Level Constructs			
AB5C	45	.38	**.52**	.45	.45
MPQ	30	.39	.42	**.52**	.44
NEO Facets	30	.41	.51	.41	.44
HEXACO	24	.38	**.52**	.41	.44
TCI	29	.39	.42	.50	.44
CPI	33	.38	.41	.47	.42
HPI-HICs	44	**.43**	.42	.41	.42
6FPQ	18	.41	.46	.29	.39

k = Number of scales.

clusters. Because this finding may seem counterintuitive to some readers, it merits additional discussion. The intercorrelations among criterion variables can arise from a higher-level attribute, common to one or more of them. In our Study 1, the most likely such attribute is a broad "general activity" factor, reflecting differences between the participants in their activity or energy levels or both. In our Study 3, the predictable intercorrelations among the clinical indicators might be traced to individual differences in general psychopathology or at least to differences in individuals' willingness to report symptoms of psychological deviance or pain or both. In both cases, these broad individual differences are quite predictable from the personality inventories under comparison, and therefore the process of orthogonalization (which removes this common variance) serves to attenuate the validities of predictions from the inventory scales. In Study 1, we reported our findings for both the original and the orthogonalized act clusters. In Study 3, we confirmed this general result in unreported analyses, and therefore we did not report any findings based on orthogonalized scores from the clinical indicators.

Comparing Higher-Level, Middle-Level, and Lower-Level Constructs

Turning from the criteria to the predictors, how well do the relatively few but more reliable upper-level composite scales compare in their cross-validity to the many more but less reliable lower-level ones? As has been argued elsewhere (Goldberg, 1993), the optimal number of variables to include in multiple regression analyses will be limited only by considerations of statistical power, and thus of sample size. This logical argument is based on the assumption that each of the lower-level measures includes some reliable (i.e., systematic) variance that is not included in the aggregate higher-level measure, and this unique variance should be differentially related to one or more criterion indices. If some criterion turned out to be optimally predicted by only the higher-level common variance, then the regression equation would weight the lower-level measures by the weights that maximize their common variance, and the predictions from each of the two levels would be the same. In all other cases, one would expect that predictions from lower-level measures will be more valid than those based on the aggregated higher-level measures.

Indeed, this logical argument has been supported by the findings from previous empirical comparisons, showing that large sets of lower-level constructs tend to be more valid than a few higher-level ones (e.g., Ashton, 1998; Ashton, Jackson, Paunonen, Helmes, & Rothstein, 1995; Judge, Bono, Erez, Locke, & Thoresen, 2002; Mershon & Gorsuch, 1988; Paunonen & Ashton, 2001a, 2001b; Paunonen & Nicol, 2001; Reynolds & Nichols, 1977; Stewart, 1999). However, the argument of Goldberg (1993) strictly applies only to predictions made in a population, or in samples large enough to approximate one. In any smaller sample, it is possible for predictions made using many lower-level measures to capitalize on chance vagaries of that sample, and such sample overfitting can lead to detrimental performance upon cross-validation in other samples; predictions made from the few higher-level measures will not be subject to so much capitalization on chance, and thus may cross-validate in a more optimal manner.

The present three studies are the first inventory-validity comparisons to use modern bootstrap resampling procedures to guard against sample overfitting. Based on 1,000 cross-validity analyses for each predictor-criterion pair, these present analyses will tend to penalize predictions at the lowest level as compared with those at the highest level. For example, a typical inventory (e.g., the NEO-PI-R) has only five highest-level measures, as compared with 30 lowest-level ones. Using a stepwise algorithm that includes the most valid five measures, all five NEO domains will be selected in every regression equation, and only the regression weights can vary from the derivation to the hold-out samples. In contrast, at the lower level there are 30 NEO facet scales involved in every analysis, and a specific subset of five of these then will be selected to be regression weighted. Thus, one would expect that sample overfitting, and thus increased shrinkage upon cross-validation, should be far more serious at the lower than at the higher level.

Seemingly, this is exactly what happened in each of our three studies, where there were virtually no differences in the overall cross-validity of the predictions as a function of the number of predictors available. What seems to have occurred is that in samples of this size, the advantages associated with the predictive usefulness of systematic variance at the lower level were almost exquisitely balanced by the increased shrinkage from the derivation to hold-out samples, and therefore there was no overall gain in cross-validity from predictions at any one of the three hierarchical levels that we employed. We expect that in studies using much larger samples, we would obtain higher cross-validities at the lowest level, even using bootstrap resampling, whereas in smaller samples one can expect some incremental cross-validities at the very highest level.

Statistical Significance of Differences Among the Validity Coefficients

As noted in our General Methods section, critical difference values were computed for each set of criteria. These numbers represent the minimum differences between the squared cross-validity coefficients for which the predictions of the inventories differ at $p < .05$, on average. We have provided an Appendix table for readers to quickly evaluate whether or not the difference between any pair of cross-validity coefficients is likely to be significant. For the six criteria of Study 1, the critical difference is .09 (see Notes after Table 3). Thus, inventories with cross-validity coefficients of .45 ($r^2 = .20$), for example, are significantly better at predicting a given criterion than inventories with coefficients of .34 ($r^2 = .11$) or lower. There were only a few differences of that magnitude for the prediction of individual criteria. For the mean validity coefficients across all of the criteria, the critical difference is considerably smaller (.04), but so is the range of cross-validity coefficients. Accordingly, the differences between the very highest and very lowest mean validity coefficients in Table 3 are of marginal significance.

For the criteria in Study 2, the critical difference is .11 (see Notes after Table 7). Here, the range of cross-validity coefficients for some of the criteria was somewhat larger than in Study 1. For example, in predicting peer-rated Conscientiousness, the Appendix table shows that cross-validity coefficients over .50 ($r^2 = .25$) are significantly greater than those under .37 ($r^2 = .14$). The HEXACO Domain scores had a coefficient of .50, and several other inventories had coefficients of .49 and .48. In contrast, there were a number of inventories with coefficients well below .37, indicating significant differences for the prediction of this and other peer-rated criteria. For the mean cross-validity coefficients, the critical difference is .05. Examination of the Appendix table will show that there were a number of differences between mean cross-validity coefficients that were well above the critical value.

Likewise for Study 3, the range of cross-validity coefficients for most criteria was quite broad, and there were some highly significant differences between validity coefficients. A number of factors may contribute to the significant differences in Studies 2 and 3. Given that the criteria used in Study 2 were derived from Big-Five factors, those inventories most directly related to the Big-Five could have an advantage. For Study 3, The TCI and the MPQ include scales that seem to be uniquely proficient in the prediction of two of the criteria. At the other end of the spectrum in Study 3, the relatively short 6FPQ was something of an outlier at both the higher and lower hierarchical levels. Removal of these three inventories would substantially narrow the range of cross-validity coefficients listed in Table 10.

The Comparative-Validity of the 11 Personality Inventories

Our comparative-validity results are summarized in Table 13, which lists the average cross-validity coefficient for each inventory in each study and also provides the mean cross-study coefficient for every inventory across all three studies. These across-study means range from a low of .37 for the 6FPQ higher-level scales to a high of .45 for the NEO and TCI domains at the higher level, as well as the MPQ scales at the middle level, and the AB5C facets at the lower level. Yet the distribution is not as broad as it appears to be. If one omits as outliers both sets of 6FPQ predictors and the three higher-level CPI vectors, all of the remaining across-study mean validities fall in the remarkably narrow range of .42 to .45.[4]

In the scientific literature in psychological assessment, there are relatively few genuine comparative-validity studies, and none focusing on a sizeable set of broad-bandwidth personality inventories. Some of those closest in spirit to the present article include Buss and Craik (1984), Johnson (2000), and Mikulay and Goffin (1998). One reason for the dearth of comparative-validity studies is purely practical: Because it requires a substantial effort, and considerable cost, to administer many inventories and many criterion measures to the same sample of research participants, it is likely that only long-continuing studies will permit the range of analyses we were able to conduct with the ESCS. Clearly, some test publishers might have the resources to carry out research of this kind, but so far none of them has elected to do so.

One reason for such reluctance could stem from the realization that their own test products may not fare so well in this sort of competitive arena. Judging from their relative performance across our three studies, the publishers of the 6FPQ, for example, might be disappointed if not displeased.

[4] These remarkably small differences in average cross-validity among the personality inventories are reminiscent of the findings from a related body of literature, namely, the average cross-validities of inventories developed from different scale-construction *strategies*, using the same item pool (e.g., Goldberg, 1972; Hase & Goldberg, 1967). Burisch (1984a), who provided an overview of 15 such comparative-validity studies, concluded, “A review of more than a dozen comparative studies revealed no consistent superiority of any strategy in terms of validity or predictive effectiveness” (p. 214). We can now say much the same thing for the 11 different inventories here under comparison.

Proponents of the 6FPQ might argue that this mini-inventory might look more cost effective if one considers its length (only 108 items), although in Study 2 it was less valid than Goldberg's Big-Five markers, which include only 100 items. On the other hand, inventory publishers may have little to fear if large numbers of criteria are employed for comparative-validity analysis; our findings show that the average cross-validities across a broad array of criteria differ only slightly.

Another major finding from our three studies is that, although most inventories differ little in terms of their average validity across many criteria, the inventories exhibit clear differences in their validities for individual criteria. Moreover, inventories that do not do particularly well against one set of criteria can be superior to all other inventories in analyses of some other criterion set. Given the similarities in overall validity for most inventories, it is in the best interests of both test publishers and inventory users to identify the regions of the criterion spectrum in which each particular inventory exhibits exceptionally good or poor validity. Another implication of this finding is the apparent utility of having a broad array of personality inventories to choose from. Critics of personality measurement, and even some proponents of particular models of human personality traits, may insist that only a single widely agreed upon approach to personality measurement will indicate that significant progress has been made in the field of personality assessment. Our results, however, clearly demonstrate the pragmatic advantages of having a variety of broad-bandwidth measures available so as to be able to select the one best suited for any given application.

SOME CAVEATS AND CONCLUSIONS

Most readers will realize that the present collection of criteria is hardly a definitive set. Industrial/organizational psychologists should demand a replication of these analyses in the workplace against some reasonably objective indices of job performance. Clinical psychologists should demand replications against alternative indices of psychopathology. Counseling psychologists might demand replications using measures of educational choices and outcomes. Certainly, it would be fair to demand that the inventories be compared with additional assessments by knowledgeable informants, only now based on a more differentiated set of attributes than the broad five-factor measures used in our Study 2. And, one could use the existing data set to predict the frequency with which our participants carry out specific acts, rather than the broad act clusters used in our Study 1.

Ultimately, one would hope for comparative-validity studies that use a set of criteria of wide scope and import. A recent theoretically derived taxonomy of "consequential outcomes" has been proposed by Ozer and Benet-Martinez (2006), and these constructs might be used as guides for criterion selection in future studies. Among the potential outcomes suggested by this analysis are the following: (a) Aspects of psychopathology and criminality; (b) indices of physical health and eventual longevity; (c) measures of occupational choice, work performance, and job effectiveness; (d) estimates of the quality of peer, family, neighborhood, and romantic relationships; (e) indices of volunteerism and community involvement; (f) aspects of political and social attitudes and values, including spirituality and other virtues; and, more generally, (g) other measures of life success and the quality of one's life, including happiness and subjective well-being.

Comparative-validity studies using a variety of such outcomes should be welcomed and supported. What should no longer be accepted so facilely are single-instrument studies, regardless of the criteria that they employ. Clearly, this report is not the last word on the comparative merits of modern personality inventories. But proponents of any such inventories now have a target at which to aim. And, before they purchase competitive products, discerning scientists and practitioners should require *evidence* that other inventories perform more validly than the best of those reviewed here. As Sternberg and Williams (1998) have indicated, if the testing industry is to change, it will come about because consumers demand that their products be tested and compared.

ACKNOWLEDGMENTS

Support for this project was provided by Grant MH-49227 (LRG) and by K01-DA-16618 (RAG) from the National Institutes of Health. A portion of our report on Study 1 in this article was based on a chapter that originally was prepared by Goldberg for a *Handbook of Adult Personality Inventories*, to have been edited by Briggs, Cheek, and Donahue, but which is no longer under development; some findings from that original chapter were later summarized by Wiggins (2003; pp. 146–148).

With the exception of the NEO-PI-R (for whose use we were charged), access to each of the other inventories was provided to the project free-of-charge by the test developer or the test publisher (along with scoring services providing item responses and scale scores) or both, in exchange for access to project data. The authors are extremely grateful to Harrison Gough, Heather Cattell, Robert Hogan, Robert Cloninger, Auke Tellegen, and Douglas N. Jackson for providing us with the CPI, 16PF, HPI, TCI, MPQ, JPI-R, and 6FPQ inventories.

In addition, the authors are grateful to each of the following colleagues for their thoughtful criticisms and suggestions: Jack Block, Shawn Boles, Michael H. Bond, Matthias Burisch, David Buss, David Campbell, Heather Cattell, William F. Chaplin, C. Robert Cloninger, Robyn M. Dawes, Eileen Donahue, Vincent Donnelly, Herbert Eber, Donald W. Fiske, David Funder, Robert Guion, Sarah E. Hampson, Paul J. Hoffman, Willem K. B. Hofstee, Robert Hogan, John A. Johnson, Edward Lichtenstein, Clarence McCormick, William McConochie, Alan Mead, Gregory J. Meyer, Dean Peabody, Robert Perloff, Steven Reise, William Revelle, Brent Roberts, Leonard G. Rorer, James A. Russell, Gerard Saucier, Oya Somer, Dennis Sweeney, Auke Tellegen,

Thomas M. Vogt, Niels Waller, Erika Westling, Jerry S. Wiggins, and Richard Zinbarg.

REFERENCES

Ashton, M. C. (1998). Personality and job performance: The importance of narrow traits. *Journal of Organizational Behavior, 19,* 289–303.

Ashton, M. C., Jackson, D. N., Paunonen, S. V., Helmes, E., & Rothstein, M. G. (1995). The criterion validity of broad factor scales versus specific facet scales. *Journal of Research in Personality, 29,* 432–442.

Ashton, S. G., & Goldberg, L. R. (1973). In response to Jackson's challenge: The comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices, and laymen. *Journal of Research in Personality, 7,* 1–20.

Benet-Martinez, V., & John, O. P. (1998). Los Cinco Grandes across cultures and ethnic groups: Multitrait multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology, 75,* 729–750.

Bernstein, E. M., & Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *Journal of Nervous and Mental Disorders, 174,* 727–735.

Bremner, J. D., Southwick, S., Brett, E., Fontana, A., Rosenheck, R., & Charney, D. S. (1992). Dissociation and posttraumatic stress disorder in Vietnam combat veterans. *American Journal of Psychiatry, 149,* 328–332.

Breslau, N. (1985). Depressive symptoms, major depression, and generalized anxiety: A comparison of self-reports on CES-D and results from diagnostic interviews. *Psychiatry Research, 15,* 219–229.

Brinkley, C. A., Schmitt, W. A., Smith, S. S., & Newman, J. P. (2001). Construct validation of a self-report psychopathy scale: Does Levenson's self-report psychopathy scale measure the same constructs as Hare's pychopathy checklist-revised. *Personality and Individual Differences, 31,* 1021–1038.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24,* 445–455.

Burisch, M (1984a). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39,* 214–227.

Burisch, M. (1984b). You don't always get what you pay for: Measuring depression with short and simple versus long and sophisticated scales. *Journal of Research in Personality, 18,* 81–98.

Buss, D. M., & Craik, K. H. (1983). The act frequency approach to personality. *Psychological Review, 90,* 105–126.

Buss, D. M., & Craik, K. H. (1984). Acts, dispositions, and personality. In B. A. Maher & W. B. Maher (Eds.), *Progress in experimental personality research* (vol. 13, pp. 241–301). Orlando, FL: Academic Press.

Buss, D. M., & Craik, K. H. (1985). Why *not* measure that trait? Alternative criteria for identifying important dispositions. *Journal of Personality and Social Psychology, 48,* 934–946.

Buss, D. M., & Craik, K. H. (1987). Act criteria for the diagnosis of personality disorders. *Journal of Personality Disorders, 1,* 73–81.

Chapman, L. J., Chapman, J. P., Kwapil, T. R., Eckblad, M., & Zinser, M. C. (1994). Putatively psychosis-prone subjects 10 years later. *Journal of Abnormal Psychology , 103,* 171–183.

Chapman, L. J., Chapman, J. P., & Raulin, M. L. (1978). Body-image aberration in schizophrenia. *Journal of Abnormal Psychology, 87,* 399–407.

Cloninger, C. R., Przybeck, T. R., Svrakic, D. M., & Wetzel, R. D. (1994). *The Temperament and Character Inventory (TCI): A guide to its development and use*. St. Louis, MO: Center for Psychobiology of Personality, Washington University.

Cloninger, C. R., Sigvardsson, S., & Bohman, M. (1988). Childhood personality predicts alcohol abuse in young adults. *Alcoholism-Clinical and Experimental Research, 12,* 494–505.

Cloninger, C. R., Svrakic, D. M., & Przybeck, T. R. (1993). A psychobiological model of temperament and character. *Archives of General Psychiatry, 50,* 975–990.

Conn, S. R., & Rieke, M. L. (1994). *The 16PF fifth edition technical manual*. Champaign, IL: Institute for Personality and Ability Testing.

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.

Eckblad, M., & Chapman, L. J. (1983). Magical ideation as an indicator of schizotypy. *Journal of Consulting and Clinical Psychology, 51,* 215–225.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association, 78,* 316–330.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1,* 54–77.

Fanous, A., Gardner, C. O., Prescott, C. A., Cancro, R., & Kendler, K. S. (2002). Neuroticism, major depression and gender: A population-based twin study. *Psychological Medicine, 32,* 719–728.

Farmer, A., Mahmood, A., Redman, K., Harris, T., Sadler, S., & McGuffin, P. (2003). A sib-pair study of the Temperament and Character Inventory scales in major depression. *Archives of General Psychiatry, 60,* 490–496.

Foa, E. B., Kozak, M. J., Salkovskis, P. M., Coles, M. E., & Amir, N. (1998). The validation of a new obsessive-compulsive disorder scale: The Obsessive-Compulsive Inventory. *Psychological Assessment, 10,* 206–214.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monograph, 7,* No. 72-2.

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

Goldberg, L. R. (1993). The structure of personality traits: Vertical and horizontal aspects. In D. C. Funder, R. D. Parke, C. Tomlinson-Keasey, & K. Widaman (Eds.), *Studying lives through time: Personality and development* (pp. 169–188). Washington, DC: American Psychological Association.

Goldberg, L. R. (1999a). 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.

Goldberg, L. R. (1999b). The Curious Experiences Survey, a revised version of the Dissociative Experiences Scale: Factor structure, reliability, and relations to demographic and personality variables. *Psychological Assessment, 11,* 134–145.

Goldberg, L. R., Grenier, J. R., Guion, R. M., Sechrest, L. B., & Wing, H. (1991). *Questionnaires used in the prediction of trustworthiness in pre-employment selection decisions: An A.P.A. Task Force report*. Washington, DC: American Psychological Association.

Gough, H. G., & Bradley, P. (2002). *CPI Manual: Third Edition*. Mountain View, CA: Consulting Psychologists Press.

Hajcak, G., Huppert, J. D., Simons, R. F., & Foa, E. B. (2004). Psychometric properties of the OCI-R in a college sample. *Behavioral Research and Therapy, 42,* 115–123.

Hase, H. D., & Goldberg, L. R. (1967). Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67,* 231–248.

Hofstee, W. K. B., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big-Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology, 63,* 146–163.

Hogan, R., & Hogan, J. (1995). *Hogan Personality Inventory Manual: Second Edition*. Tulsa, OK: Hogan Assessment Systems.

Jackson, D. N. (1994). *Jackson Personality Inventory—Revised manual*. Port Huron, MI: Sigma Assessment Systems.

Jackson, D. N., Paunonen, S. V., & Tremblay, P. F. (2000). *Six Factor Personality Questionnaire*. Port Huron, MI: Sigma Assesment Systems.

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* (2nd ed., pp. 102–138). New York: Guilford.

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.

Judge, T. A., Bono, J. E., Erez, A., Locke, E. A., & Thoresen, C. J. (2002). The scientific merit of valid measures of general concepts: Personality research and core self-evaluations. In J. M. Brett & F. Drasgow (Eds.), *The psychology of work: Theoretically based empirical research* (pp. 55–77). Mahwah, NJ: Erlbaum.

Krueger, R. F. (1999). Personality traits in late adolescence predict mental disorders in early adulthood: A prospective-epidemiological study. *Journal of Personality, 67,* 39–65.

Krueger, R. F., Hicks, B. M., Patrick, C. J., Carlson, S. R., Iacono, W. G., & McGue, M. (2002). Etiologic connections among substance dependence, antisocial behavior, and personality: Modeling the externalizing spectrum. *Journal of Abnormal Psychology, 111,* 411–424.

Lee, K., & Ashton, M. C. (2004). Psychometric properties of the HEXACO Personality Inventory. *Multivariate Behavioral Research, 39,* 329–358.

Leichsenring, F. (1999). Development and first results of the Borderline Personality Inventory: A self-report instrument for assessing borderline personality organization. *Journal of Personality Assessment, 73,* 45–63.

Levenson, M. R., Kiehl, K. A., & Fitzpatrick, C. M. (1995). Assessing psychopathic attributes in a noninstitutionalized population. *Journal of Personality and Social Psychology, 68,* 151–158.

Loehlin, J. C., & Nichols, R. C. (1976). *Heredity, environment, and personality: A study of 850 sets of twins*. Austin: University of Texas.

Lynam, D. R., Whiteside, S., & Jones, S. (1999). Self-reported psychopathy: A validation study. *Journal of Personality Assessment, 73,* 110–132.

McGrath, R. (2005) Conceptual complexity and construct validity. *Journal of Personality Assessment, 85,* 112–124.

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology, 55,* 675–680.

Mikulay, S. M., & Goffin, R. D. (1998). Measuring and predicting counter-productivity in the laboratory using integrity and personality testing. *Educational and Psychological Measurement, 58,* 768–790.

Moskowitz, A. K., Barker-Collo, S., & Ellson, L. (2005). Replication of dissociation-psychosis link in New Zealand students and inmates. *Journal of Nervous and Mental Disease, 193,* 722–727.

Myers, J. K., & Weissman, M. M. (1980). Use of a self-report symptom scale to detect depression in a community sample. *American Journal of Psychiatry, 137,* 1081–1084.

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57,* 401–421.

Paunonen, S. V., & Ashton, M. S. (2001a). Big Five predictors of academic achievement. *Journal of Research in Personality, 35,* 78–90.

Paunonen, S. V., & Ashton, M. S. (2001b). Big Five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology, 81,* 524–539.

Paunonen, S. V., & Nicol, A. A. M. (2001). The personality hierarchy and the prediction of work behaviors. In B. W. Roberts & R. Hogan (Eds.), *Personality psychology in the workplace*. Washington, DC: American Psychological Assocation.

Putnam, F. W., Carlson, E. B., Ross, C. A., Anderson, G., Clark, P., Torem, M., et al. (1996). Patterns of dissociation in clinical and nonclinical samples. *Journal of Nervous and Mental Disease, 184,* 673–679.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for use in the general population. *Applied Psychological Measurement, 1,* 385–401.

Reynolds, C. H., & Nichols, R. C. (1977). Factor scales of the CPI: Do they capture the valid variance? *Educational and Psychological Measurement, 37,* 907–915.

Roberts, R. E., & Vernon, S. W. (1983). The Center for Epidemiologic Studies Depression Scale: Its use in a community sample. *American Journal of Psychiatry, 140,* 41–46.

Russell, M. T., & Karol, D. L. (1994). *The 16PF Fifth Edition Administrator's Manual*. Champaign, IL: Institute for Personality and Ability Testing.

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63,* 506–516.

Sternberg, R. J., & Williams, W. M. (1998). You proved our point better than we did: A reply to our critics. *American Psychologist, 53,* 576–577.

Stewart, G. L. (1999). Trait bandwidth and stages of job performance: Assessing differential effects for conscientiousness and its subtraits. *Journal of Applied Psychology, 84,* 959–968.

Tellegen, A. (in press). *MPQ (Multidimensional Personality Questionnaire): Manual for administration, scoring, and interpretation*. Minneapolis: University of Minnesota Press.

Tellegen, A., & Waller, N. G. (in press). Exploring personality through test construction: Development of the Multidimensional Personality Questionnaire. In G. J. Boyle, G. Matthews, & D. H. Saklofske (Eds.), *Handbook of personality theory and testing: Vol. II. Personality measurement and assessment*. London: Sage.

Tremblay, R. E., Pihl, R. O., Vitaro, F., & Dobkin, P. L. (1994). Predicting early onset of male antisocial behavior from preschool behavior. *Archives of General Psychiatry, 51,* 732–739.

Wiggins, J. S. (2003). *Paradigms of personality assessment*. New York: Guilford.

Woodworth, R. S. (1919). Examination of emotional fitness for warfare. *Psychological Bulletin, 16,* 59–60.

APPENDIX
Guide for Determining the Statistical Significance of Differences in the Cross-Validity Coefficients Between Any Two Inventories ($p < .05$)

	Maximum Cross-Validity r Value That is Significantly Lower ($p < .05$) *Than the Coefficient in Column 1 for the Critical* Δ *in* r^2			
	Single Criterion		*Mean Across Criteria*	
Higher Cross-Validity r Coefficient	Δ *in* $r^2 = .11$ *(Study 2)*	Δ*in* $r^2 = .09$ *(Studies 1 & 3)*	Δ*in* $r^2 = .05$ *(Study 2)*	Δ*in* $r^2 = .04$ *(Studies 1 & 3)*
.30	.00	.00	.20	.22
.35	.11	.18	.27	.29
.40	.22	.26	.33	.35
.45	.30	.34	.39	.40
.50	.37	.40	.45	.46
.55	.44	.46	.50	.51
.60	.50	.52	.56	.57
.65	.56	.58	.61	.62
.70	.62	.63	.66	.67

Note. The values in the table are the maximum cross-validities that are significantly lower than the coefficients listed in the first column. Δ in r^2 = the critical values for significance for the findings listed in Table 3 (Study 1), Table 7 (Study 2), and Table 10 (Study 3). The first two columns (.11 and .09) are used for tests of differences for any single criterion. The last two columns are used for tests of the mean cross-validity coefficients across all of the criteria.

Lewis R. Goldberg
Oregon Research Institute
1715 Franklin Boulevard
Eugene, OR 97403-1983
Email: lewg@ori.org

Received September 1, 2006
Revised December 29, 2006