2012

An Evaluation of the Consequences of Using Short Measures of the Big Five Personality Traits

Marcus Credé
University at Albany–State University of New York, mcrede@albany.edu

Peter D. Harms
University of Nebraska - Lincoln, pharms@gmail.com

Sarah Niehorster
University at Albany–State University of New York

Andrea Gaye-Valentine
University at Albany–State University of New York

Published in *Journal of Personality and Social Psychology* 102:4 (2012), pp. 874–888; doi: 10.1037/a0027403

Submitted February 15, 2011, revised October 10, 2011; accepted January 23, 2012; published online February 20, 2012.

An Evaluation of the Consequences of Using Short Measures of the Big Five Personality Traits

Marcus Credé
Department of Psychology, University at Albany–State University of New York

Peter Harms
Department of Management, University of Nebraska–Lincoln

Sarah Niehorster and Andrea Gaye-Valentine
Department of Psychology, University at Albany–State University of New York

Corresponding author – Marcus Credé, Department of Psychology, University at Albany–State University of New York, Albany, NY 12222; email mcrede@albany.edu

Abstract
Researchers often use very abbreviated (e.g., 1-item, 2-item) measures of personality traits due to their convenience and ease of use as well as the belief that such measures can adequately capture an individual's personality. Using data from 2 samples (N = 437 employees, N = 355 college students), we show that this practice, particularly the use of single-item measures, can lead researchers to substantially underestimate the role that personality traits play in influencing important behaviors and thereby overestimate the role played by new constructs. That is, the use of very short measures of personality may substantially increase both the Type 1 and Type 2 error rates. We argue that even slightly longer measures can substantially increase the validity of research findings without significant inconvenience to the researcher or research participants.

Keywords: short scales, personality measurement, validity, Type 1 error, Type 2 error

> Everything should be made as simple as possible,
> but no simpler. —Einstein's razor

Researchers are often faced with the need to assess complex psychological constructs in a very short amount of time. This might occur when researchers are given an extremely limited amount of time with respondents, as is sometimes the case in organizational settings where managers are concerned with minimizing interference with the completion of work-related tasks or when researchers realize that respondents might not enjoy extensive assessment sessions. Indeed, it has been argued (e.g., Gosling, Rentfrow, & Swann, 2003) that there are instances when researchers are faced with the choice of assessing constructs with very brief measures or not at all. These might include longitudinal studies requiring repeated assessments on numerous constructs, studies that require both self-reports and other-reports of a construct, and experience-sampling studies where respondents are asked to complete inventories numerous times per day for a number of consecutive days or even weeks. Asking respondents to complete a long survey with seemingly repetitive items can lead to boredom, fatigue, and annoyance (Burisch, 1984a; Robins, Hendin, & Trzesniewski, 2001) and, therefore, reduces the likelihood that respondents will attend to item-content with care or agree to participate in follow-up data collections. These practical imperatives and psychological reasons have led efforts to increase the efficiency with which psychological constructs can be assessed. Computer adaptive testing, for example, greatly reduces the amount of time necessary for the accurate assessment of a variety of constructs such as abilities, knowledge, and even attitudes (Koch & Dodd, 1990). A more common method for attempting to decrease the amount of time necessary for data collection is the development of shorter inventories of constructs.

Shorter inventories have subsequently been developed in a wide variety of psychological domains. One such domain is the measurement of job satisfaction where single-item measures such as the Faces Scale (Kunin, 1955) remain popular with many researchers, although concerns regarding the low construct validity and low reliability of such measures (e.g., Loo & Kells, 1998; Nagy, 2002; Wanous, Reichers, & Hudy, 1997) has resulted in a general decline in the use of such measures in favor of multi-item scales (e.g., Job Descriptive Index; P. C. Smith, Kendall, & Hulin, 1969). The measurement of personality, particularly Big Five traits, is another area in which the use of short scales in favor of longer scales has become widespread despite widely acknowledged psychometric problems that are similar to those identified for short measures of job satisfaction. Early measures of the Big Five personality traits (Extraversion, Emotional Stability, Agreeableness, Conscientiousness, and Openness to Experience) were relatively long (e.g., the 240-item NEO Personality Inventory—Revised; Costa & McCrae, 1992), but shorter scales have also become available and are widely used. Shorter inventories that claim to assess the Big Five traits include the following: Goldberg et al.'s (2006) 100-item trait-descriptive adjectives; the 60-item NEO Five-Factor Inventory (Costa & McCrae, 1992); the 50-item measure from the International Personality Item Pool (IPIP; Goldberg et al., 2006); the 44-item Big Five Inventory (John, Donahue, & Kentle, 1991); Saucier's (1994) 40-item

Big Five Mini-Markers; the 20-item mini-IPIP measure developed by Donnellan, Oswald, Baird, and Lucas (2006); a variety of 10-item inventories (e.g., the Ten Item Personality Inventory [TIPI]; Gosling et al., 2003); and a variety of five-item inventories (e.g., Aronson, Reilly, & Lynn, 2006; Bernard, Walsh, & Mills, 2005; Woods & Hampson, 2005).

Many of the shortest of these Big Five measures (i.e., one or two items per construct) are widely used by researchers. The 10-item measure developed by Gosling et al. (2003), for example, has been cited over 720 times, and the 20-item measure developed by Donnellan et al. (2006) has already been cited over 110 times despite its very recent publication and the publication delays that are typical among many psychology journals. The popularity of these short measures is based on a variety of factors that are worth briefly reviewing.

Reasons for the Use of Short Inventories of Personality

The reasons for the popularity of short inventories of personality fall into two broad categories. The first of these is largely practical in nature. Short inventories of personality take little time to complete and are therefore less likely to result in feelings of boredom or fatigue in respondents (Burisch, 1984a). As such they are less likely to result in negative participant reactions—as manifest by either a refusal to participate in research or a tendency to respond to items in a careless or effectively random fashion—than might be the case when respondents are asked to respond to a personality inventory comprised of hundreds of items (e.g., 240-item NEO Personality Inventory—Revised; Costa & McCrae, 1992). Participant reactions can, of course, be important influences on the validity of data because low response rates are often associated with concerns about the external validity of findings (volunteer bias; Rosnow & Rosenthal, 1974), whereas even very low base rates of careless or random responding can have significant effects on the validity of correlational research (Credé, 2010; Schmitt & Stults, 1985). The brevity of short inventories is also important in research settings in which researchers only have limited time with participants and where numerous other constructs must be assessed. Organizational researchers, for example, are often only given limited time to interact with employees due to the negative impact of research participation on employee productivity. In such settings, researchers are often faced with the choice of either assessing personality with very short inventories or not assessing personality at all. Short inventories may also have higher levels of face validity to the typical respondent who does not have a sophisticated understanding of measurement practices because short inventories do not contain numerous items that might appear to be redundant with each other (Wanous et al., 1997). Finally, it is also possible that an under-appreciation of the role and importance of personality constructs may result in some researchers treating personality constructs as "noise" that should be statistically controlled for, resulting in the use of personality measures that do not require much time for participants to complete.

The second category of reasons given for the use of short measures is psychometric in nature—specifically, evidence that the psychometric sacrifices involved in using very short inventories may not be as substantial as commonly assumed. For example, a review of personality scales by Burisch (Burisch, 1984a; see also Burisch, 1984b, 1997; cf. Paunonen & Jackson, 1985) concluded that short scales were no worse than longer scales of the same construct in terms of criterion validity. Similar evidence regarding the comparable criterion-validity of short versus longer scales has also been presented by other authors (e.g., Robins et al., 2001; Thalmeyer, Saucier, & Eigenhuis, 2011). Thalmeyer et al. (2011), for example, compared various medium length measures of personality such as the 60-item NEO Five-Factor Inventory (Costa & McCrae, 1992) to shorter measures such as Rammstedt and John's (2007) 10-item measure and found relatively small decrements in validity. There is also some evidence that the other psychometric properties of shortened scales, such as test–retest reliability (Gosling et al., 2003) and convergent validity (Robins et al., 2001; Wood, Nye, & Saucier, 2010), can also be highly satisfactory. Further, recent work by Yarkoni (2010) suggests that algorithmic approaches to scale shortening can result in scales that are reduced in length by more than 90% without substantial psychometric sacrifices. Despite the often sound practical reasons for the use of short measures of personality and the psychometric arguments mustered in defense of their use, there are also important reasons to be cautious about the size of dramatically shortened scales, particularly scales that have been shortened to single items.

Reasons for Caution in the Use of Short Inventories of Personality

Despite the frequently cited evidence that the psychometric weaknesses of short inventories or single-item measures of personality may be limited, there are two primary reasons[1] why scores on short measures of personality are likely to have less predictive validity than scores on well-constructed longer inventories. The first of these relates to random measurement error. Responses to individual items are typically characterized by a non-trivial amount of random measurement error such that an individual's response to a single item does not necessarily reflect that person's true standing on the construct being assessed. Averaging responses across multiple items (that assess the same construct) minimizes the amount of measurement error because random measurement errors are as likely to be positive as negative and therefore have a tendency to cancel each other out when averaged across multiple items. The subsequent reduction in random measurement error (i.e., higher reliability) associated with multi-item measures of scales is thought to result in higher validity for scores on multi-item inventories than is the case for scores on very short scales. From this perspective, even a simple shift from a single-item measure of a construct to a two-item measure would result in a dramatic reduction in measurement error and hence a dramatic improvement in criterion validity. Although this view that the criterion validity of scores on inventories is strongly related to the reliability of those scores as assessed by internal consistency estimates (e.g., Cronbach's alpha) is certainly the dominant paradigm in the measurement literature, recent work by McCrae, Kurtz, Yamagata, and Terracciano (2011) suggests that test–retest reliability may be more predictive of score validity than are estimates of internal consistency.

1. We also refer readers to G. T. Smith, McCarthy, and Anderson (2000) for a discussion of nine common problems with the manner in which many short measures of psychological constructs are developed.

The second reason why scores on very short inventories of personality are typically thought to have less criterion validity than scores on longer inventories is related to the fact that short scales of personality are likely to be characterized by substantial content deficiency (G. T. Smith, McCarthy, & Anderson, 2000). Big Five traits are commonly thought to exhibit a hierarchical structure with each individual trait being comprised of several sub-facets (Costa & McCrae, 1992; Roberts, Bogg, Walton, Chernyshenko, & Stark, 2004), and no small set of items can adequately capture a single facet, let alone all the facets of a particular trait— especially when the items in short scales are often selected in a manner that maximize alpha reliability and hence maximize item redundancy (John & Soto, 2007; G. T. Smith et al., 2000). Consider extraversion as an example. Extraversion is thought to be comprised of six facets often referred to as warmth, gregariousness, assertiveness, activity level, excitement-seeking, and positive emotions—with the first three of these combined to form an overall sociability factor and the last three combined to form a dominance factor (e.g., Beck, Burnet, & Vosper, 2006). Very short measures are faced with one of two broad options. The first is try to assess the overall extraversion trait with specific Likert-scale items such as "I like to talk to a lot of different people at parties" (Goldberg, 1999), but these are likely to favor one facet (e.g., gregariousness) over other facets, resulting in construct deficiency if the total measure is only comprised of only one or two such items. As a consequence of this construct underrepresentation, substantial breadth in predictive validity is lost (Messick, 1995). The second option is to ask respondents to make overall judgments of their level of extraversion after providing respondents with a relatively detailed description of what it might mean to have high or low levels of extraversion. This approach is favored by a number of single-item measures of personality. Aronson et al. (2006), for example, presented respondents with a description of each Big Five trait, which is comprised of between eight and 10 adjective descriptors, and then asked for a single rating of standing on each trait. Bernard et al. (2005) provided respondents with descriptions (taken from the NEO Personality Inventory—Revised manual) of individuals with high, medium, and low levels of each Big Five trait and asked the respondents to indicate their own standing on the trait using these descriptors. Woods and Hampson (2005) took a similar approach, by providing relatively detailed descriptors of the end points of each Big Five personality continuum and asking respondents to indicate their standing along the continuum. Although these approaches may reduce the problem of content validity somewhat by providing respondents with descriptors of personality traits that are more complete than those found in typical single-item inventories, they do require the respondent to read and understand a very long (and often complex) item while performing a mental averaging of their standing on each of the individual adjectives or descriptors provided in the item. Whether respondents are able to understand such complex items and perform an appropriate averaging across all the item components is unclear. Similarly, it is unclear whether such very long and complex single items really take less time and effort than would be the case for multi-item inventories comprised of more traditional short items that are comprised of single adjectives or descriptive phrases.[2] Items with complex content have been identified as problematic by a variety of authors (e.g., Condon, Ferrando, & Demestre, 2006; Janes, 1999; Moreno, Martinez, & Muniz, 2006; Spector, 1992). Further, such single-item measures that ask respondents to indicate their overall standing on a Big Five trait are likely to be more susceptible to being contaminated by a lay understanding of what it means to have high levels of that trait. For example, in the case of extraversion, many non-psychologists may assume that the trait is primarily comprised of sociability such that self-ratings of overall extraversion (even when provided with a description of all the extraversion facets) are likely to not fully reflect the dominance components of the trait.

Together, these psychometric limitations relating to reliability and content validity result in scores on shortened measures that typically exhibit lower correlations with criteria than scores on longer measures of the same construct (e.g., Paunonen & Jackson, 1985). These concerns are widely acknowledged—even by those advocating the use of very short measures of personality and responsible for their construction (e.g., Donnellan et al., 2006; Gosling et al., 2003; Rammstedt & John, 2007; Woods & Hampson, 2005). In their description of the TIPI, Gosling et al. (2003) even go so far as to argue that

> . . . we hope that this instrument will not be used in place of established multi-item instruments. Instead, we urge that this instrument be used when time and space are in short supply and when only an extremely brief measure of the Big Five will do. (p. 525)

Impact on Type 1 and Type 2 Error Rates

Given these psychometric concerns, it is clear that the use of such scales for measuring the Big Five can increase the Type 2 error rate for tests of the null hypothesis that a given personality trait is not related to some other variable. Unfortunately, the use of very short Big Five measures has led researchers in a wide variety of domains to declare the influence of some Big Five traits on behaviors (e.g., political behavior; Mondak, Hibbing, Canache, Seligson, & Anderson, 2010) and important criteria (e.g., wellbeing; Sheldon & Hoon, 2007) to be non-significant or trivial. Similarly, single-item measures of Big Five traits have been used in studies that claim to establish the discriminant validity of new constructs based on low correlations with scores on very short measures of personality (e.g., Cognitive Styles; Cools & Van den Broeck, 2007).

Less well understood is that the use of short measure of personality can also dramatically increase the Type 1 error rate for instances in which researchers examine the incremental variance in a criterion explained by some new predictor over and above the variance explained by personality (typically Big Five traits). This increase in Type 1 error is due both to the decreased amount of initial variance in the criterion explained by the shortened Big Five measures but also by the lowered relationship between the new predictor and scores on the shortened Big Five measure. The incremental variance explained by the new predictor over and above the variance explained by scores on a measure of the Big Five is a function

2. The argument that longer tests may cause greater rater fatigue is often used as a justification for shorter scales. However, it has been pointed out that shorter scales may not only be more vulnerable to respondent carelessness but also that having few items makes tests more vulnerable to having responses to prior items, contaminating responses to other items (Podsakoff et al., 2003).

of (1) the relationship between the new predictor and the criterion, (2) the relationship of scores on the Big Five measure with the criterion, (3) the relationships between the new predictor and scores on the Big Five measure, and (4) the relationships among scores on the Big Five measure. Using a shortened (and hence less reliable and content deficient) measure of the Big Five will reduce the strength of both (2) and (3) and hence (artificially) increase the incremental variance provided by the new predictor.

This possibility of basing incremental validity claims of a new construct over and above Big Five personality traits assessed with very short measures of Big Five traits is not purely theoretical. For example, the two-item scales of Big Five traits developed by Gosling et al. (2003) have been used to support claims for the incremental validity of (among others) emotional intelligence (e.g., Chamorro-Premuzic, Bennett, & Furnham, 2007), free will (Stillman et al., 2010), self-realization (Miquelon & Vallerand, 2008), and psychological capital (Luthans, Avolio, Avey, & Norman, 2007).

Goals of the Study

The widespread use of dramatically shortened scales of Big Five personality traits suggests a need to examine the tradeoff between the increased convenience of using a shortened scale that can be completed in a matter of minutes and the loss of criterion-relevant variance and hence increased Type 1 and Type 2 error rates. This article aims to examine this issue by comparing the criterion-related variance captured by scores on eight publicly available shortened scales of the Big Five personality traits. This will not only illustrate the convergent validity of different measures of Big Five traits, as well as the degree to which criterion related variance is lost by using short measures of personality, but will also allow researchers to directly compare measures of equal length to each other because scales differ from each other not only in the number of items per construct but also in the format in which these items are presented (scaling, response options, instruction sets, etc.). The criterion chosen for this study are those considered important in both organizational and educational settings and that have previously been linked with Big Five traits, including task performance (Barrick, Mount, & Judge, 2001), organizational citizenship behaviors (OCBs; Ilies, Fulmer, Spitzmueller, & Johnson, 2009), counterproductive workplace behaviors (CWBs; Berry, Ones, & Sackett, 2007), job satisfaction (Ilies et al., 2009), stress (Kim, Shin, & Swagner, 2009), academic performance (O'Connor & Paunonen, 2007), and health behaviors (Bogg & Roberts, 2004).

Method

Samples

Data were drawn from two samples. The work sample was composed of 437 employed individuals who were recruited via the StudyResponse Project (Stanton & Weiss, 2002). The sample was 49% male and 79% Caucasian; the participants had an average age of 40.07 years (SD = 11.27). Seventy-five percent of respondents described their job as blue-collar, with 52% holding supervisory or managerial positions, and 83% reported at least some post high school education. The student sample was composed of 395 undergraduate students drawn from the participant pool of a large, public university in the northeastern United States. The sample was 58% female, 59% Caucasian, and largely comprised of freshman students (59%); the participants had an average age of 19.07 years (SD = 1.60).

Measures

Personality traits (work sample and student sample). The Big Five personality traits were assessed using a total of eight different scales with different numbers of items: three single-item measures (Aronson et al., 2006; Bernard et al., 2005; Woods & Hampson, 2005), two two-item measures (Gosling et al., 2003; Rammstedt & John, 2007), one four-item measure (Donnellan et al., 2006), one six-item measure (Shafer, 1999), and one eight-item measure (Saucier, 1994).

Task performance (work sample). Self-rated task performance was assessed using a seven-item scale described by Williams and Anderson (1991). Employees were asked to indicate their level of agreement with each statement (e.g., "I fulfill the responsibilities specified in my job description") using a 5-point response scale.

Academic performance (student sample). Academic performance was assessed using a self-report of grade-point average. Previous meta-analytic research (Kuncel, Credé, & Thomas, 2005) has shown self-reported grades to be very strongly correlated with actual grades.

Health behaviors (student sample). Health behaviors were assessed with the 40-item Health Behavior Checklist (Vickers, Conway, & Hervig, 1990), which is comprised of four subscales: Preventative Health Behaviors, Accident Control, Traffic Risk Behaviors, and Substance Risk. Respondents are asked to indicate their level of agreement with each statement using a 5-point response format.

Daily behaviors (student sample). Daily student behaviors were assessed with the 54-item self-report Daily Behavior Survey (Wu & Clark, 2003), which assesses four types of daily behaviors: Exhibitionism (7 items), Aggression (18 items), Failure to Plan (9 items), and Spontaneity (7 items). For each of the 54 behaviors, respondents were asked to indicate whether they had engaged in the described behavior in the previous 24 hr using a Yes–No response format.

Contextual performance (work sample). OCBs and CWBs were, respectively, assessed using a 17-item self-report scale (e.g., "volunteered to orient or train others") and a 19-item self-report scale (e.g., "Attempted to pass on own work to others") described by Credé, Chernyshenko, Stark, Dalal, and Bashshur (2007). Employees were asked to indicate their level of agreement with each statement using a 5-point response scale.

Withdrawal cognitions (work sample). Withdrawal behaviors were assessed using a three-item scale comprised of statements reflecting the desire to change jobs and behavior related to finding a new job (e.g., "Made plans to leave the organization"). Employees were asked to indicate their level of agreement with each statement using a 5-point response scale.

Job satisfaction (work sample). Job satisfaction was assessed using the eight-item Abridged Job-In-General Scale described by Russell et al. (2004). Each item is an adjective or short phrase describing a job (e.g., "excellent"), and em-

Table 1. Comparison of Means and Standard Deviations for the Work Sample and Student Sample

	Work sample			Student sample				
Constructs and measures	*M*	*SD*	α	*M*	*SD*	α	*t*	*d*
Extraversion								
Single item: Aronson et al. (2006)	3.38	1.18		3.70	1.04		–4.07	–0.28
Single item: Woods & Hampson (2005)	4.83	2.14		5.32	2.06		–3.36	–0.23
Single item: Bernard et al. (2005)	4.50	1.37		4.67	1.16		–1.93	–0.13
Two items: Gosling et al. (2003)	4.11	1.49	.65	4.72	1.37	.67	–6.11	–0.42
Two items: Rammstedt & John (2007)	3.08	1.01	.61	3.37	0.98	.65	–4.21	–0.29
Four items: Donnellan et al. (2006)	4.08	1.41	.78	4.57	1.50	.85	–4.85	–0.34
Six items: Shafer (1999)	4.48	1.26	.86	4.84	1.20	.88	–4.20	–0.29
Eight items: Saucier (1994)	5.59	1.44	.83	5.89	1.38	.83	–3.08	–0.21
Agreeableness								
Single item: Aronson et al. (2006)	4.13	0.84		4.18	0.87		–0.84	–0.06
Single item: Woods & Hampson (2005)	5.76	1.96		6.16	1.86		–2.98	–0.21
Single item: Bernard et al. (2005)	4.37	1.61		4.90	1.30		–5.16	–0.36
Two items: Gosling et al. (2003)	5.16	1.25	.45	5.04	1.12	.47	1.49	0.10
Two items: Rammstedt & John (2007)	3.66	0.87	.46	3.76	0.84	.37	–1.73	–0.12
Four items: Donnellan et al. (2006)	5.27	1.09	.74	5.34	1.13	.81	–0.88	–0.06
Six items: Shafer (1999)	4.67	1.00	.75	4.50	0.95	.72	2.50	0.17
Eight items: Saucier (1994)	6.45	1.08	.73	7.06	1.19	.82	–7.68	–0.53
Conscientiousness								
Single item: Aronson et al. (2006)	4.43	0.73		4.06	0.93		6.35	0.44
Single item: Woods & Hampson (2005)	5.82	1.95		5.48	2.00		2.48	0.17
Single item: Bernard et al. (2005)	5.03	1.28		4.47	1.28		6.36	0.44
Two items: Gosling et al. (2003)	5.67	1.18	.55	5.44	1.17	.45	2.72	0.19
Two items: Rammstedt & John (2007)	4.04	0.84	.45	3.53	0.85	.45	8.73	0.61
Four items: Donnellan et al. (2006)	5.27	1.12	.65	4.74	1.28	.77	6.37	0.44
Six items: Shafer (1999)	5.73	0.89	.75	5.13	1.00	.80	9.14	0.71
Eight items: Saucier (1994)	6.96	1.29	.85	6.44	1.26	.82	5.78	0.40
Emotional Stability								
Single item: Aronson et al. (2006)	4.09	0.89		3.82	1.08		3.83	0.27
Single item: Woods & Hampson (2005)	5.30	2.01		4.62	1.95		4.93	0.34
Single item: Bernard et al. (2005)	4.82	1.51		4.50	1.45		3.10	0.22
Two items: Gosling et al. (2003)	5.01	1.43	.69	4.68	1.39	.64	3.30	0.23
Two items: Rammstedt & John (2007)	3.48	1.03	.60	3.08	1.00	.55	5.70	0.40
Four items: Donnellan et al. (2006)	4.62	1.37	.78	4.41	1.21	.70	2.29	0.16
Six items: Shafer (1999)	4.76	1.22	.87	4.36	1.13	.85	4.88	0.17
Eight items: Saucier (1994)	6.11	1.68	.85	5.49	1.39	.81	5.72	0.40
Openness								
Single item: Aronson et al. (2006)	4.24	0.76		4.25	0.81		–0.09	–0.01
Single item: Woods & Hampson (2005)	5.36	1.93		5.64	2.05		–2.01	–0.14
Single item: Bernard et al. (2005)	4.91	1.26		4.84	1.35		0.73	0.05
Two items: Gosling et al. (2003)	5.00	1.21	.53	5.35	1.13	.48	–4.22	–0.29
Two items: Rammstedt & John (2007)	3.49	0.90	.40	3.64	0.98	.54	–2.27	–0.16
Four items: Donnellan et al. (2006)	4.98	1.25	.77	5.12	1.21	.80	–1.63	–0.11
Six items: Shafer (1999)	4.51	0.95	.66	4.47	0.92	.66	0.61	0.04
Eight items: Saucier (1994)	6.42	1.21	.79	6.52	1.27	.82	–1.13	–0.08

ployees were asked to indicate the degree to which the adjective or phrase accurately describes their job using a 3-point response scale.

Stress (work sample). Work stress was assessed using the 15-item Stress in General Scale (Stanton, Balzer, Smith, Parra, & Ironson, 2001). Each item is an adjective or short phrase (e.g., "demanding") describing a job, and employees were asked to indicate the degree to which the adjective or phrase accurately describes their job using a 3-point response scale.

College performance (student sample). College performance was assessed using 19 items that assess students self-rated academic competence (five items), problems due to drug and alcohol use (four items), students' ability to deal with stress effectively (five items), presence of study skills (two items), and their development of social skills (three items).

Self-ratings (student sample). Self-ratings of physical attractiveness, intelligence, popularity, and integrity were assessed with single items from the Behavior Report Form (Paunonen, 2003).

Results

Means, standard deviations, and internal consistency estimates for scores on all personality measures for both samples are presented in Table 1. The student sample rated itself as, on average, more extraverted, less conscientious, less emotionally stable, and slightly more agreeable.

Table 2. Correlations of Scores on Eight Versions of Big Five Traits With Criteria (Work Sample)

Constructs and measures	Task performance	OCBs	CWBs	Withdrawal cognitions	Job satisfaction	Stress
Extraversion						
Single item: Aronson et al. (2006)	–.12	.06	.06	.02	.17	–.02
Single item: Woods & Hampson (2005)	.03	.09	–.07	.00	.08	.07
Single item: Bernard et al. (2005)	–.09	.15	–.03	–.08	.29	–.05
Two items: Gosling et al. (2003)	.06	.20	–.11	–.03	.16	.02
Two items: Rammstedt & John (2007)	.03	.16	–.06	–.01	.10	.03
Four items: Donnellan et al. (2006)	.02	.18	–.05	.00	.14	.01
Six items: Shafer (1999)	.15	.25	–.16	–.13	.24	–.05
Eight items: Saucier (1994)	.15	.22	–.14	–.08	.20	.00
Agreeableness						
Single item: Aronson et al. (2006)	.17	.26	–.14	–.10	.24	–.11
Single item: Woods & Hampson (2005)	.13	.18	–.16	–.16	.18	–.07
Single item: Bernard et al. (2005)	.35	.28	–.28	–.19	.05	–.07
Two items: Gosling et al. (2003)	.41	.42	–.42	–.30	.24	–.10
Two items: Rammstedt & John (2007)	.33	.36	–.40	–.27	.34	–.16
Four items: Donnellan et al. (2006)	.38	.47	–.36	–.24	.28	–.07
Six items: Shafer (1999)	.31	.33	–.37	–.25	.34	–.19
Eight items: Saucier (1994)	.50	.49	–.51	–.41	.32	–.13
Conscientiousness						
Single item: Aronson et al. (2006)	.31	.27	–.25	–.21	.15	–.06
Single item: Woods & Hampson (2005)	.22	.06	–.27	–.14	–.02	.00
Single item: Bernard et al. (2005)	.20	.15	–.20	–.11	.08	–.02
Two items: Gosling et al. (2003)	.54	.35	–.47	–.34	.18	–.11
Two items: Rammstedt & John (2007)	.51	.37	–.48	–.36	.23	–.06
Four items: Donnellan et al. (2006)	.50	.34	–.50	–.28	.17	–.09
Six items: Shafer (1999)	.52	.44	–.45	–.30	.27	–.07
Eight items: Saucier (1994)	.60	.43	–.49	–.33	.22	–.11
Emotional Stability						
Single item: Aronson et al. (2006)	.25	.33	–.22	–.13	.20	–.09
Single Item: Woods & Hampson (2005)	–.02	.14	–.02	.12	–.02	.03
Single item: Bernard et al. (2005)	.02	.21	.00	.04	.09	–.08
Two items: Gosling et al. (2003)	.34	.36	–.31	–.13	.11	–.06
Two items: Rammstedt & John (2007)	.21	.23	–.15	–.06	.05	–.01
Four items: Donnellan et al. (2006)	.36	.28	–.33	–.11	.09	–.02
Six items: Shafer (1999)	.05	.16	–.07	.00	.06	–.05
Eight items: Saucier (1994)	.35	.34	–.20	–.05	.07	–.01
Openness						
Single item: Aronson et al. (2006)	.18	.19	–.24	–.17	.31	–.20
Single item: Woods & Hampson (2005)	.15	.14	–.15	–.08	.14	–.18
Single item: Bernard et al. (2005)	.13	.20	–.23	–.17	.34	–.21
Two items: Gosling et al. (2003)	.33	.33	–.38	–.30	.31	–.21
Two items: Rammstedt & John (2007)	.28	.26	–.31	–.16	.32	–.25
Four item: Donnellan et al. (2006)	.36	.33	–.36	–.27	.32	–.21
Six items: Shafer (1999)	.32	.30	–.31	–.26	.35	–.27
Eight items: Saucier (1994)	.43	.36	–.49	–.36	.35	–.23

OCBs – organizational citizenship behaviors; CWBs – counterproductive workplace behaviors.

Correlations With Criteria (Work Sample)

The correlations of scores on the eight different measures of Big Five personality traits with the examined criteria for the work sample are provided in Table 2. These tables do not show the correlations among scores on each the Big Five measures (due to space constraints), but correlations among the different measures for a single traits exhibited moderate average intercorrelations of r = .68 (for Extraversion), r = .48 (Agreeableness), r = .46 (Conscientiousness), r = .61 (Emotional Stability), and r = .48 (Openness). However, it should be noted that the highest degree of convergence was found for the longer scales. For example, single-item measures of Conscientiousness correlated between .20 and .49 with other measures of Conscientiousness. For the scales with four or more items, intercorrelations on Conscientiousness ranged from .58 to .74.

The results also illustrate not only that internal consistency estimates tend to increase as the number of items increase (not surprising given that alpha is a function of scale length) but also that the criterion validities of scores on Big Five measures are often dramatically higher for scores based on longer scales. For example, scores on single-item measures of Conscientiousness correlated between r = .20 and r = .31 with self-ratings of task performance, whereas scores on the eight-item Conscientiousness scale correlated r = .60 with task performance.

To explore this issue further, we conducted a series of regression analyses in which the various criteria assessed for the work sample were regressed onto scores on each of the Big Five

Table 3. Comparison of the Proportion of Variance in Criteria Explained by Different Versions of the Big Five Measures

	Task performance		OCBs		CWBs		Withdrawal cognitions		Average across criteria	
Big Five measures	Big Five *R*	Δ*R* from JS	Big Five *R*	Δ*R* from JS	Big Five *R*	Δ*R* from JS	Big Five *R*	Δ*R* from JS	Big Five *R*	Δ*R* from JS
Single item: Aronson et al. (2006)	.390	.013–	.391	.033–	.322	.050–	.224	.309–	.332	.101
Single item: Woods & Hampson (2005)	.293	.032–	.253	.089–	.365	.056–	.247	.291–	.289	.117
Single item: Bernard et al. (2005)	.429	.031–	.389	.048–	.394	.042–	.263	.263–	.368	.096
Two items: Gosling et al. (2003)	.608	.000	.525	.014–	.566	.009–	.410	.167–	.527	.048
Two items: Rammstedt & John (2007)	.559	.000	.470	.011–	.555	.003	.392	.172–	.494	.047
Four items: Donnellan et al. (2006)	.612	.000	.529	.008	.582	.009–	.363	.185–	.522	.051
Six items: Shafer (1999)	.548	.000	.483	.009	.510	.005	.339	.185–	.470	.050
Eight items: Saucier (1994)	.649	.000	.554	.011-	.590	.002	.448	.127-	.560	.035

Δ*R* from JS = incremental *R* provided by job satisfaction after controlling for scores on Big Five measure. All *R* values are based on shrunken (i.e., adjusted) values. OCBs = organizational citizenship behaviors; CWBs = counterproductive workplace behaviors.
* $p < .05$

trait measures. As noted earlier, Big Five trait measures are often used as controls to illustrate the incremental validity of other constructs over scores on measures of personality, and the amount of incremental validity that a construct can exhibit is, in part, a function of the amount of variance already explained by other predictors (in this case scores on a Big Five measure). Results of these regression analyses (see Table 3) illustrate dramatic differences in the proportion of variance accounted for by different versions of Big Five inventories, with the eight-item per trait measure by Saucier (1994) explaining, on average, more than twice as much variance as explained by any of the three single-item-per-trait measures. This effect was even larger for the three job performance criteria (task performance, OCBs, and CWBs). Table 3 also provides estimates of the amount of incremental variance explained by job satisfaction for each criterion over scores on each of the examined Big Five trait measures. Again, dramatic differences in incremental validity findings are evident. For example, job satisfaction scores explain statistically significant and substantial amounts of incremental variance in counterproductive behavior (average Δ*R* = .049) when using single-item measures of Big Five traits but non-significant incremental variance when using a six- or eight-item measure of Big Five traits (average Δ*R* = .0035).

Correlations with Criteria (Student Sample)

Correlations of scores on the eight measures of Big Five traits with the criteria examined for the student sample are summarized in Table 4 and Table 5. Given that not all Big Five traits should be expected to correlate significantly with all of the examined criteria, we further summarized the relationships in two different ways. First, Table 6 provides a summary of the adjusted *R* values obtained when each of the criteria are regressed onto scores on each of the different Big Five trait measure. Second, we focused our attention on those trait– criterion relationships for which at least one of the eight trait measures correlated at $r > .20$ (or $r < -.20$). We then averaged, for each of the different trait measures, the absolute correlations for the selected criteria. For Extraversion, the selected criteria were exhibitionism, traffic risk, substance risk, academic competence, alcohol and drug use, stress management, social skills, attractiveness, popularity, and integrity. For Agreeableness, the selected criteria were aggression, academic competence, stress management, social skills, and integrity. For Conscientiousness, the selected criteria were exhibitionism, planfulness, wellness maintenance, accident control, traffic risk, substance risk study skills, academic competence, alcohol and drug use, stress management, and integrity. For Emotional Stability, the selected criteria were exhibitionism, aggression, wellness maintenance, study skills, alcohol and drug use, stress management, social skills, attractiveness, popularity, and integrity. For Openness, the selected criteria were academic competence, social skills, self-rated physical attractiveness, self-rated intelligence, and self-rated popularity. The resultant summary of the criterion related validity of each of the trait measures is summarized in Table 7. Also presented in Table 7 is an overall average criterion-related validity coefficient for Big Five measures of different lengths. Although far less dramatic than our findings from the work sample, these findings highlighted in Table 6 and Table 7 replicate the general increase in validity coefficients as the length of the scale increases. Notable exceptions to this general pattern observed in the student sample is that Aronson et al.'s (2006) measure performed significantly better than the other single-item measures, and in many instances, it explained almost as much variance in criteria as longer measures (a finding we explore in more detail in our discussion).

Discriminant Validity With Short Measures of Personality

Although the concern of many test designers is to ensure that new measures correlate with criteria as highly as possible, both Campbell (1960) and Messick (1995) have reminded us that discriminant validity is a critical feature of construct validity. To establish whether scores on the very-short measures of personality traits correlated with criteria in the manner that would be anticipated given prior meta-analytic estimates, we correlated the observed validity coefficients for scores on each Big Five trait measure with those from prior meta-analyses that utilized a Big Five framework. If these measures are valid assessments of the Big Five, we would expect to find high-profile correlations between their results and those of prior research. Given the variables in the current study, we

Table 4. Correlations of Scores on Eight Versions of Big Five Traits and Criteria (Student Sample)

Constructs and measures	Exhibitionism	Aggression	Planfulness	Spontaneity	Wellness maintenance	Accident control	Traffic risk	Substance risk
Extraversion								
Single item: Aronson et al. (2006)	.30	.06	−.11	.11	.15	.03	.30	.29
Single item: Woods & Hampson (2005)	.40	.16	−.08	.17	.11	.05	.33	.28
Single item: Bernard et al. (2005)	.37	.14	−.06	.17	.07	.12	.31	.27
Two items: Gosling et al. (2003)	.35	.09	−.09	.15	.14	.06	.33	.30
Two items: Rammstedt & John (2007)	.31	.06	−.15	.16	.08	−.03	.36	.35
Four items: Donnellan et al. (2006)	.41	.09	−.09	.17	.16	.07	.39	.44
Six items: Shafer (1999)	.33	.04	−.05	.14	.13	.06	.33	.35
Eight items: Saucier (1994)	.33	.07	−.05	.15	.14	.08	.36	.29
Agreeableness								
Single item: Aronson et al. (2006)	−.05	−.21	−.01	.01	.10	.01	−.03	−.04
Single item: Woods & Hampson (2005)	−.10	−.18	.01	−.07	.04	.00	−.09	−.11
Single item: Bernard et al. (2005)	−.08	−.22	−.11	−.02	.04	−.02	.01	−.01
Two items: Gosling et al. (2003)	−.17	−.29	.04	.01	.07	.03	−.10	−.10
Two items: Rammstedt & John (2007)	−.19	−.31	.05	−.07	.09	.07	−.03	−.12
Four items: Donnellan et al. (2006)	.04	−.23	.03	−.09	.07	.07	−.03	.07
Six items: Shafer (1999)	−.12	−.29	.00	.03	.07	.08	−.06	−.05
Eight items: Saucier (1994)	−.10	−.27	.05	.02	.15	.05	−.04	−.04
Conscientiousness								
Single item: Aronson et al. (2006)	−.22	−.15	.18	−.16	.31	.19	−.19	−.25
Single item: Woods & Hampson (2005)	−.09	−.03	.23	−.12	.21	.17	−.19	−.17
Single item: Bernard et al. (2005)	−.16	−.02	.14	−.06	.20	.16	−.19	−.28
Two items: Gosling et al. (2003)	−.17	−.11	.19	−.19	.28	.15	−.11	−.21
Two items: Rammstedt & John (2007)	−.20	−.12	.20	−.17	.35	.24	−.09	−.22
Four items: Donnellan et al. (2006)	−.17	−.15	.25	−.17	.31	.21	−.20	−.21
Six items: Shafer (1999)	−.20	−.14	.19	−.15	.33	.25	−.11	−.23
Eight items: Saucier (1994)	−.18	−.11	.23	−.16	.35	.24	−.11	−.20
Emotional Stability								
Single item: Aronson et al. (2006)	−.11	−.21	.03	.02	.23	.12	−.04	−.01
Single item: Woods & Hampson (2005)	−.04	−.06	−.02	.16	.00	−.06	.11	.01
Single item: Bernard et al. (2005)	−.09	−.16	−.07	.12	.10	.03	.11	−.06
Two items: Gosling et al. (2003)	−.14	−.19	−.02	.04	.13	.05	.10	−.06
Two items: Rammstedt & John (2007)	−.01	−.07	−.11	.12	.07	.07	.17	.06
Four items: Donnellan et al. (2006)	−.14	−.18	−.05	.08	.11	.05	.08	−.01
Six items: Shafer (1999)	−.08	−.16	−.13	.09	.11	.06	.18	.01
Eight items: Saucier (1994)	−.22	−.21	−.07	.01	.07	.05	.00	−.12
Openness								
Single item: Aronson et al. (2006)	.10	.00	−.05	.03	.04	.05	.11	.15
Single item: Woods & Hampson (2005)	.10	−.02	.02	−.03	−.06	.05	.05	.08
Single item: Bernard et al. (2005)	.04	−.06	−.01	.02	−.04	.03	.06	.09
Two items: Gosling et al. (2003)	.07	−.05	.03	.03	.01	.09	.16	.15
Two items: Rammstedt & John (2007)	.01	−.06	.01	−.08	.06	.11	.02	.07
Four items: Donnellan et al. (2006)	.03	−.07	−.02	−.03	.00	.14	.05	.12
Six items: Shafer (1999)	.07	−.04	.04	−.02	−.02	.16	.08	.10
Eight items: Saucier (1994)	.01	−.08	.04	−.06	.09	.15	.08	.07

used meta-analytic results from studies linking personality traits with task performance (Barrick & Mount, 1991), OCBs (Borman, Penner, Allen, & Motowidlo, 2001), CWBs (Salgado, 2002), and job satisfaction (Judge, Heller, & Mount, 2002). For example, Barrick and Mount (1991) reported mean corrected correlations of the Big Five traits with job performance of .13 (Extraversion), .08 (Emotional Stability), .07 (Agreeableness), .22 (Conscientiousness), and .04 (Openness). These five estimates were transformed into *z* scores, the correlations observed in this study were similarly transformed into *z* scores, and a correlation was calculated between these two sets of *z* scores. This was repeated for each of the eight measures of Big Five traits examined in this study to calculate a profile correlation for each measure—a correlation that reflects the degree to which the pattern of correlations observed in this study matches those from meta-analytic review of the literature. Results (see Table 8) are supportive of earlier results; that is, in general, longer measures produce scores that correlate with the criteria in a pattern that is in line with previous meta-analytic estimates of the relationships of Big Five personality traits with the examined criteria. This pattern was found for task performance, OCBs, and CWBs. Curiously though, all of the measures produced dramatically different patterns to those established in Judge et al.'s (2002) earlier work on job satisfaction. Judge et al. found no relationship between openness and job satisfaction; the very-short measures consistently ranked openness as the best or second best predictor of job satisfaction. Moreover, in nearly every case, the very-short measures estimated neuroticism, which Judge et al. showed was the best predictor of job satisfaction, as the worst predictor. It seems

Table 5. Correlations of Scores on Eight Versions of Big Five Traits and Criteria (Student Sample)

Constructs and measures	Study skills	Academic competence	Alcohol and drug use	Stress management	Social skills	Attractiveness	Intelligence	Popularity	Integrity
Extraversion									
Single item: Aronson et al. (2006)	.08	.17	.22	.27	.47	.39	.06	.50	.15
Single item: Woods & Hampson (2005)	.05	.16	.26	.17	.35	.26	.02	.42	.10
Single item: Bernard et al. (2005)	.06	.14	.22	.22	.42	.32	.07	.46	.11
Two items: Gosling et al. (2003)	.09	.21	.24	.29	.43	.33	.06	.51	.15
Two items: Rammstedt & John (2007)	.11	.20	.33	.25	.46	.38	.10	.50	.11
Four items: Donnellan et al. (2006)	.12	.23	.38	.27	.53	.41	.09	.59	.14
Six items: Shafer (1999)	.08	.21	.27	.25	.53	.33	.03	.52	.10
Eight items: Saucier (1994)	.12	.30	.23	.29	.46	.39	.15	.55	.23
Agreeableness									
Single item: Aronson et al. (2006)	−.01	.03	−.07	.16	.24	.07	−.05	.06	.16
Single item: Woods & Hampson (2005)	.03	−.01	−.06	.03	.18	−.08	−.17	−.01	.11
Single item: Bernard et al. (2005)	−.03	.04	.00	.09	.18	.02	−.01	−.02	.20
Two items: Gosling et al. (2003)	.07	.12	−.06	.17	.24	−.01	−.04	.01	.21
Two items: Rammstedt & John (2007)	.13	.02	−.10	.20	.29	.02	−.10	.06	.18
Four items: Donnellan et al. (2006)	.06	.19	.04	.12	.31	.08	.02	.10	.26
Six items: Shafer (1999)	.04	.05	−.04	.15	.30	−.08	−.11	.02	.22
Eight items: Saucier (1994)	.09	.21	−.06	.23	.34	.10	.06	.12	.23
Conscientiousness									
Single item: Aronson et al. (2006)	.35	.20	−.23	.30	.03	.02	.01	−.05	.16
Single item: Woods & Hampson (2005)	.11	.10	−.17	.10	−.10	.02	.06	−.07	.07
Single item: Bernard et al. (2005)	.21	.02	−.28	.12	−.07	.00	.14	−.07	.09
Two items: Gosling et al. (2003)	.26	.21	−.26	.35	.10	.13	.18	.06	.20
Two items: Rammstedt & John (2007)	.36	.33	−.22	.39	.17	.12	.19	.09	.27
Four items: Donnellan et al. (2006)	.26	.19	−.21	.26	.07	.06	.05	−.02	.18
Six items: Shafer (1999)	.31	.28	−.29	.36	.13	.16	.19	.09	.30
Eight items: Saucier (1994)	.31	.23	−.24	.34	.09	.14	.15	.09	.22
Emotional Stability									
Single item: Aronson et al. (2006)	.20	.18	−.11	.54	.31	.24	.14	.25	.15
Single item: Woods & Hampson (2005)	−.04	−.02	.00	.14	.12	.15	.08	.20	.07
Single item: Bernard et al. (2005)	.05	−.01	−.06	.33	.28	.22	.09	.23	.07
Two items: Gosling et al. (2003)	.07	.10	−.12	.36	.31	.20	.11	.23	.21
Two items: Rammstedt & John (2007)	.09	.15	−.01	.37	.25	.24	.19	.29	.17
Four items: Donnellan et al. (2006)	.08	.11	−.11	.38	.29	.23	.14	.25	.17
Six items: Shafer (1999)	.11	.11	−.04	.38	.32	.26	.17	.33	.27
Eight items: Saucier (1994)	.14	.15	−.21	.35	.28	.19	.15	.20	.22
Openness									
Single item: Aronson et al. (2006)	.10	.11	.09	.09	.17	.15	.11	.18	.08
Single item: Woods & Hampson (2005)	−.02	.12	.09	.00	.07	.06	.09	.05	.05
Single item: Bernard et al. (2005)	.03	.13	.05	−.01	.10	.15	.05	.05	.05
Two items: Gosling et al. (2003)	.09	.19	.07	.18	.28	.23	.18	.22	.09
Two items: Rammstedt & John (2007)	.01	.16	.09	.08	.09	.09	.15	.05	.03
Four items: Donnellan et al. (2006)	.04	.25	.07	.09	.13	.08	.21	.12	.04
Six items: Shafer (1999)	.04	.22	.05	.03	.09	.09	.16	.12	.03
Eight items: Saucier (1994)	.07	.32	.02	.12	.06	.20	.35	.14	.08

likely that the social desirability of the items retained in the item-reduction process may have played a role in determining these results. If only the three types of performance outcomes were taken into account, it appeared that the Rammstedt and John (2007) 10-item scale produced the most accurate pattern of results. This would seem to illustrate that simply making scales longer or shorter does not necessarily increase or decrease validities and that careful scale design and item selection can make up for a lack of several different items.

Discussion

The motivation for the construction and use of ever shorter measures of psychological constructs is clear and understandable. Researchers are under great pressure to obtain data from individuals on multiple constructs—preferably from multiple sources and across multiple time points. The development of methods that make the collection of such data easier—or feasible at all—is highly desirable, especially as researchers attempt to better understand within-person variation in constructs across time and situations and study populations with which extended contact time is difficult to ensure. Very short measures of personality have largely been developed to allow the assessment of individuals' standing on personality traits in such research settings and have become increasingly popular—often even in settings where more extensive contact time with research participants is available. The use of highly abbreviated measures is based upon the belief that such measures can be used to measure personality traits without significant sacrifices to the quality of obtained data and without

Table 6. *A*djusted R Values for Scores on Different Big Five Trait Measures External variables

	Big Five trait measure							
	Aronson et al. (2006) 1 item	Woods & Hampson (2005) 1 item	Bernard et al. (2005) 1 item	Gosling et al. (2003) 2 items	Rammstedt & John (2007) 2 items	Donnellan et al. (2006) 4 items	Shafer (1999) 6 items	Saucier (1994) 8 items
Exhibitionism	.411	.420	.425	.436	.414	.475	.443	.459
Aggression	.341	.243	.311	.324	.326	.355	.329	.333
Planfulness	.148	.228	.130	.192	.274	.241	.239	.235
Spontaneity	.138	.249	.182	.214	.266	.253	.232	.214
Wellness maintenance	.330	.232	.212	.313	.351	.354	.335	.358
Accident control	.161	.164	.170	.145	.245	.259	.283	.251
Traffic risk	.341	.381	.332	.371	.386	.434	.409	.390
Substance risk	.385	.333	.369	.373	.436	.474	.447	.382
Study skills	.373	.071	.205	.272	.371	.281	.305	.321
Academic competence	.281	.212	.164	.326	.392	.385	.377	.438
Alcohol and drug use	.346	.300	.329	.352	.415	.444	.417	.391
Stress management	.587	.245	.392	.535	.549	.511	.497	.504
Social skills	.506	.394	.480	.530	.539	.600	.575	.539
Self-rated attractiveness	.390	.310	.363	.407	.418	.448	.432	.425
Self-rated intelligence	.095	.197	.148	.265	.318	.243	.330	.390
Self-rated popularity	.503	.458	.490	.543	.531	.614	.569	.557
Self-rated integrity	.207	.176	.247	.319	.324	.330	.392	.344
Average across criteria	.326	.271	.291	.348	.386	.394	.389	.384

All values are adjusted *R* values for the criterion variables listed in the first column.

Table 7. Average of Correlations Between Scores on Different Big Five Trait Measures and Selected Criteria (Student Sample)

Measure	Extraversion	Agreeableness	Conscientiousness	Emotional Stability	Openness	Average
Single item: Aronson et al. (2006)	.30	.16	.24	.23	.14	.216
Single item: Woods & Hampson (2005)	.28	.10	.15	.07	.08	.135
Single item: Bernard et al. (2005)	.28	.15	.17	.15	.10	.170
Two items: Gosling et al. (2003)	.31	.20	.22	.20	.22	.231
Two items: Rammstedt & John (2007)	.33	.20	.27	.15	.11	.210
Four items: Donnellan et al. (2006)	.38	.22	.23	.19	.16	.235
Six items: Shafer (1999)	.33	.20	.27	.20	.14	.225
Eight items: Saucier (1994)	.33	.26	.25	.21	.22	.253

significantly affecting the validity of conclusions drawn from such data. The results presented in this article suggest that these beliefs are generally incorrect and that the use of very short measures of personality traits, particularly single-item measures, can be associated with significant decrements in the validity of research findings. The pursuit of ever shorter measures of personality traits may have reached the point where the cost in terms of a loss in criterion-related validity, in some circumstances, outweighs the benefits associated with such convenient low-burden measures.

A reliance on very short measures of Big Five traits appears to result in a substantial increase in both Type 1 and Type 2 errors. The impact on Type 2 error rates is easily (and widely) understood. The poor content validity and low reliability of short measures of traits results in a general underestimation of the strength of relationships between traits and criteria. The absolute strength of relationships is therefore underestimated, and studies with low to moderate power are more likely to conclude that no statistically significant relationship exists between a trait and a criterion. Our results confirm this general pattern, with relationships between criteria and scores on very short measures of Big Five traits being, on average, lower than the relationships observed for slightly longer measures of Big Five traits.

Our results also illustrate the likelihood of an increase in Type 1 errors. Scores on very short measures of Big Five traits account for significantly less variance in most examined criteria than scores on slightly longer measures of these traits. As a result, the incremental variance explained by a variable above and beyond the variance accounted for by scores on Big Five trait measures is likely to be artificially inflated when Big Five traits are assessed using very short measures. The results from Study 1 clearly illustrate this issue. Job satisfaction appears to explain substantial incremental variance in task performance over Big Five traits when personality is assessed using single items but explains no incremental variance when longer measures of personality are used. Similarly, the incremental *R* provided by job satisfaction for withdrawal behaviors declines from an average of $\Delta R = .287$ for single items of personality to $\Delta R = .127$ when personality is assessed with eight items per trait.

Table 8. Profile Correlations Between Meta-Analytic Relationships and Relationships With Different Big Five Trait Measures

Measure	Task performance	OCBs	CWBs	Job satisfaction
Single item: Aronson et al. (2006)	.24	.70	.16	–.92
Single item: Woods & Hampson (2005)	.34	–.21	.44	–.73
Single item: Bernard et al. (2005)	.05	.03	.24	–.58
Two items: Gosling et al. (2003)	.42	.66	.50	–.97
Two items: Rammstedt & John (2007)	.50	.87	.59	–.76
Four items: Donnellan et al. (2006)	.27	.56	.50	–.93
Six items: Shafer (1999)	.50	.71	.49	–.74
Eight items: Saucier (1994)	.38	.75	.29	–.86

All relationships are from the work sample. OCBs = organizational citizenship behaviors; CWBs = counterproductive workplace behaviors.

Longer scales are likely to have both greater content validity and are also likely to result in more reliable scores (because alpha reliability is function of scale length). This general confounding of scale length and content validity makes it difficult to determine whether the improved criterion validity of longer scales is the result of the improved reliability of measurement or the result of greater content validity. Results from the work sample suggest that both play a role inasmuch as they illustrated a substantial difference in criterion validities between single-item and two-item measures; the average total adjusted R for the criteria reported in Table 3 improved from an average of .33 for single-item measures to an average of .51 for two-item measures—a 54% increase. Not all of this increase can be explained by an improvement in the reliability of measurement. The average internal consistency estimate of scores on the two-item measures was $\alpha = .53$, and application of the Spearman–Brown prophecy formula would suggest a reliability of $\alpha = .36$ for a single-item measure—an approximate 30% reduction in reliability that cannot account for the observed criterion validity differences. Similarly, content validity differences alone do not appear to account adequately for the substantial differences in criterion validities between single-item and two-item measures—particularly because the very substantial length and complexity of many single-item measures appear to be designed to ensure adequate content validity.

The increase in validities for scores on longer scales observed for the work sample was only moderately replicated in the student sample. This finding is similar to the findings based on a student sample presented by Thalmeyer et al. (2011). Thalmayer et al. did not examine single-item measures and only examined one two-item measure of Big Five traits but found that scores on this 10-item measure of the Big Five traits together explained approximately as much variance in various criteria as longer scales (and even outperformed them for some criteria). We speculate that this phenomenon of lower decrements in validity for our student sample is primarily due to differences between the two samples in both the examined criteria and participant characteristics. In general, the criteria examined for the student sample exhibited much weaker relationships with Big Five traits than was the case for the criteria examined for the work sample, making it more difficult for substantial differences in validities across different trait measures to be observed. Further, the student sample was comprised of undergraduate students who received class credit for participation, whereas the work sample was comprised of employees who were paid for participation. Given the well-known difficulties of conducting research with undergraduate populations (e.g., Gallen & Berry, 1996, 1997), and the substantial differences in conscientiousness (see Table 1), we suspect that a non-trivial proportion of our student sample may have responded carelessly to our longer measures because these contained multiple items of similar (and hence seemingly redundant) content.

In addition to the general finding that criterion validities increased as the number of items increased, two more specific findings are also worth noting. First, non-trivial differences in validities were observed among the three different single-item measures of Big Five traits, with scores on the measure described by Woods and Hampson (2005) performing significantly worse than scores on the measures by Aronson et al. (2006) or Bernard et al. (2005). Formatting differences, particularly the lack of a verbal descriptor for the middle response option in Woods and Hampson's measure, may have resulted in lower differentiation among participants and may have lowered the observed validities. Providing verbal descriptions for each response-option has also been shown to increase the reliability of scores (e.g., Krosnick, 1999; Weng, 2004). Second, scores on the six-item measure by Shafer (1999) exhibited average validities that were lower or equal to those observed for scores on the Donnellan et al. (2006) four-item scale. We attribute this primarily to the structural characteristics of Shafer's scales: bipolar response options and items that require a fairly high vocabulary and reading level (e.g., unagitated, persevering, uninquisitive, antagonistic). Item analysis suggests that a number of the items of Shafer's scale correlated relatively weakly with other items in the same subscale, thereby reducing their reliability.

In settings in which personality assessment with long inventories with well-illustrated content validity may not be possible or feasible, researchers should base their choice of shortened measure on four primary considerations. First, single-item measures should be avoided whenever possible. Burisch (1984a) may have concluded that short inventories often have satisfactory psychometric properties, but this conclusion should not be taken as justification for shortening measures of complex psychological phenomena to the point of a single-item; our results clearly illustrate the significant

sacrifices to criterion validity associated with such a practice. Second, the nature of the research question must be considered. We urge researchers to not rely on very short scales when aiming to make strong claims about whether a particular construct is distinct from personality traits or whether it explains variance in a criterion above and beyond the variance explained by personality. The inflated Type 1 and Type 2 error rates are likely to result in unnecessary construct profusion. The only possible exception is when the criterion being examined is narrow in scope. Broad criteria, like job performance, are widely thought to require broad predictors (Cronbach, 1960; Hogan & Roberts, 1996), and the content deficiency associated with very short measures of broad personality traits are thus likely to result in reduced criterion validity for scores on such short measures. For more specific and narrow criteria, the impact of using very short measures of personality may be less severe. For example, for the criterion of risky behaviors (e.g., drug use, speeding), scores on a very short measure of excitement-seeking (a facet of extraversion) may actually exhibit higher criterion validity than a longer measure of overall extraversion that includes facets of the broad trait (e.g., warmth) that are not strongly related to the risky behaviors.

The third consideration is related to the nature of the research setting and characteristics of the participants. Short measures may be appropriate in research settings where participants' propensity for boredom, fatigue, or disinterest may be relatively high because longer measures may increase the rate of careless or random responding to individual items, thereby artificially decreasing or increasing observed criterion validities (Credé, 2010). This consideration is particularly important when considering that the majority of research published in many psychology journals is based on data gathered from psychology students who participate in research in return for course credit—a population that is often not particularly interested in responding to long surveys in a highly attentive manner. This phenomenon may explain why the criterion validity sacrifice of using very short measures was substantially lower for our student sample than for our employee sample.

The final consideration relates to the time taken to complete an inventory—a feature of measures that is likely to be relatively distinct from the number of items in the inventory. Our results suggest that two-item measures represent a very substantial improvement over single-item measures in criterion validity with further moderate gains evident for scales that are slightly longer. Adding an additional five items to a measure (i.e., using two-item rather than single-item measure) would likely only add a minute or two to the time taken for completion of the survey but appears to substantially decrease both Type 1 and Type 2 errors. The time taken to complete slightly longer measures may even be lower than the time taken to complete single-item measures when considering that single-item measures often have extremely lengthy and relatively complex item content. For example, the single-item measure of extraversion proposed by Aronson et al. (2006) is comprised of a total of 19 words, whereas the two extraversion items proposed by Gosling et al. (2003) have a combined total of 12 words. Similarly, the single-item measure of extraversion from Woods and Hampson (2005) requires the respondent to read 41 words before formulating a response, whereas Saucier's (1994) eight-item measure of extraversion requires respondents to only read a combined total of eight words.

Limitations and Future Research

Findings based on data from single-method sources are typically affected by common-method variance that can both artificially increase and artificially decrease the correlations observed between variables (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003). The data presented in this article are all self-report and hence prone to common method influence, but our examination of the influence of scale length requires a common-method approach for the measurement of personality to ensure that scale length (and format) differences are not confounded by method differences. Future research should nevertheless attempt to examine the questions examined in this article using sources for both trait ratings and criterion information that are not self-ratings particularly because short measures of personality traits are likely to be most useful in settings in which other-reports of personality (e.g., parent-report, spouse-report, coworker-report) are being gathered. Similarly, we encourage future researchers to examine the different predictive validities observed for scores on short measures of personality for more objective criteria—particularly for criteria that are used to make high stakes decisions (e.g., employment, health, or college admissions settings). It may also be of interest to attempt to measure the content breadth of different short measures of personality in some fashion (perhaps using expert ratings), as this would allow a more detailed examination of the degree to which the increased validities that tend to be observed for longer measures can be attributed to better content validity as opposed to a simple reduction in measurement error due to a greater number of items.

Conclusions

The use of short measures of personality for research and practice (particularly single-item measures) is questionable at best. The present results indicate that shortened inventories are associated with often substantially reduced criterion validity. This is not an unanticipated effect because the loss of both content validity and reliability associated with short measures has been noted in the past (e.g., Schmitt, 1996; G. T. Smith et al., 2000).

Given what seems to be an unrelenting push toward the use of shorter measures, it seems that the focus should shift toward conducting better validity studies of the new measures being introduced. It should no longer be acceptable to simply correlate a new measure of personality with a slightly longer one and check the convergent correlations, especially because any common items artificially inflate the correlation. Instead, researchers should be encouraged to conduct more rigorous construct validation efforts across multiple independent samples to establish acceptable discriminant and predictive validity, to ensure that the theoretical factor structure has been reproduced and that the content coverage and classification rates of the shortened scale are not significantly worse than is the case for the longer scale (see G. T. Smith et al., 2000). New methods for the development of shortened scales may ultimately hold more promise than the current practice of relying on exploratory factor analysis and reliability analysis as means of reducing scale length and illustrating the ade-

quacy of a dramatically shortened measure. Two excellent examples of such new methods that appear to result in scales that are both short and retain satisfactory breadth of measurement are the development of the Index of Individual Differences in the Lexicon (Wood et al., 2010), which was based on cluster analysis, and the Analog to Multiple Broadband Inventories (Yarkoni, 2010), which relied on genetic algorithm-based item selection. For example, the method described by Yarkoni (2010) allows researchers to examine many different personality constructs (high breadth of measurement) with relatively few items (high brevity), thereby greatly enhancing the efficiency and quality with which personality information can be gathered. Until measures using these new approaches are made available and have been properly validated across multiple independent samples, the results of our study indicate that researchers and practitioners alike should resist the usage of very short measures of personality or at least acknowledge the inevitable reductions in construct validity associated with them.

References

Aronson, Z. H., Reilly, R. R., & Lynn, G. S. (2006). The impact of leader personality on new product development teamwork and performance: The moderating role of uncertainty. *Journal of Engineering and Technology Management, 23,* 221–247. doi: 10.1016/j.jengtecman.2006 .06.003

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1–26. doi: 10.1111/j.1744-6570.1991.tb00688.x

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9,* 9–30. doi: 10.1111/1468-2389.00160

Beck, E., Burnet, K. L., & Vosper, J. (2006). Birth-order effects on facets of extraversion. *Personality and Individual Differences, 40,* 953–959. doi: 10.1016/j.paid.2005.09.012

Bernard, L. C., Walsh, R. P., & Mills, M. (2005). Ask once, may tell: Comparative validity of single and multiple item measurement of the Big-Five personality factors. *Counseling and Clinical Psychology Journal, 2,* 40–57.

Berry, C. M., Ones, D. S., & Sackett, P. R. (2007). Interpersonal deviance, organizational deviance, and their common correlates: A review and meta-analysis. *Journal of Applied Psychology, 92,* 410 – 424. doi: 10.1037/0021-9010.92.2.410

Bogg, T., & Roberts, B. W. (2004). Conscientiousness and health behaviors: A meta-analysis. *Psychological Bulletin, 130,* 887–919. doi: 10.1037/0033-2909.130.6.887

Borman, W., Penner, L., Allen, T., & Motowidlo, S. (2001). Personality predictors of citizenship performance. *International Journal of Selection and Assessment, 9,* 52–69. doi: 10.1111/1468-2389.00163

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

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. doi: 10.1016/0092- 6566(84)90040-0

Burisch, M. (1997). Test length and validity revisited. *European Journal of Personality, 11,* 303–315. doi: 10.1002/(SICI)1099-0984(199711)11: 4–303::AID-PER292–3.0.CO;2-#

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, and discriminant validity. *American Psychologist, 15,* 546–553. doi: 10.1037/h0048255

Chamorro-Premuzic, T., Bennett, E., & Furnham, A. (2007). The happy personality: Mediational role of trait emotional intelligence. *Personality and Individual Differences, 42,* 1633–1639. doi: 10.1016/j.paid .2006.10.029

Condon, L., Ferrando, P. J., & Demestre, J. (2006). A note on the item characteristics related to acquiescent responding. *Personality and Individual Differences, 40,* 403–407. doi: 10.1016/j.paid.2005.07.019

Cools, E., & Van den Broeck, H. (2007). Development and validation of the Cognitive Style Indicator. *The Journal of Psychology: Interdisciplinary and Applied, 141,* 359–387. doi: 10.3200/JRLP.141.4.359-388

Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO-PI–R professional manual*. Odessa, FL: Psychological Assessment Resources.

Credé, M. (2010). Random responding as a threat to the validity of effect size estimates in correlational research. *Educational and Psychological Measurement, 70,* 596–612. doi: 10.1177/0013164410366686

Credé, M., Chernyshenko, O. S., Stark, S., Dalal, R. S., & Bashshur, M. (2007). Job satisfaction as mediator: An assessment of job satisfaction's position within the nomological network. *Journal of Occupational and Organizational Psychology, 80,* 515–538. doi: 10.1348/096317906X136180

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York, NY: Harper & Row.

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

Gallen, R. T., & Berry, D. T. (1996). Detection of random-responding in MMPI-2 protocols. *Assessment, 3,* 171–178.

Gallen, R. T., & Berry, D. T. (1997). Partially random MMPI-2 protocols: When are they interpretable? *Assessment, 4,* 61–68.

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

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. C. (2006). The International Personality Item Pool and the future of public-domain personality measures. *Journal of Research in Personality, 40,* 84–96. doi: 10.1016/j.jrp.2005.08.007

Gosling, S. D., Rentfrow, P. J., & Swann, W. B. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37,* 504–528. doi: 10.1016/S0092-6566(03)00046-1

Hogan, J., & Roberts, B. W. (1996). Issues and non-issues in the fidelity–bandwidth trade-off. *Journal of Organizational Behavior, 17,* 627–637. doi: 10.1002/(SICI)1099-1379(199611)17:6–627::AID-JOB2828–3.0.CO; 2-F

Ilies, R., Fulmer, I. S., Spitzmuller, M., & Johnson, M. D. (2009). Personality and citizenship behavior: The mediating role of job satisfaction. *Journal of Applied Psychology, 94,* 945–959. doi: 10.1037/a0013329

Janes, J. (1999). Survey construction. *Library Hi Tech, 17,* 321–325. doi: 10.1108/07378839910289376

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

John, O. P., & Soto, C. J. (2007). The importance of being valid: Reliability and the process of construct validation. In R. W. Robins, R. C. Fraley, & R. F. Krueger (Eds.), *Handbook of research methods in personality psychology* (pp. 461–494). New York, NY: Cambridge University Press.

Judge, T. A., Heller, D., & Mount, M. K. (2002). Five-factor model of personality and job satisfaction: A meta-analysis. *Journal of Applied Psychology, 87,* 530–541. doi: 10.1037/0021-9010.87.3.530

Kim, H. J., Shin, K. H., & Swanger, N. (2009). Burnout and engagement: A comparative analysis using the Big Five personality dimensions. *International Journal of Hospitality Management, 28,* 96–104. doi: 10.1016/j.ijhm.2008.06.001

Koch, W. R., & Dodd, B. G. (1990). Computer adaptive measurements of attitudes. *Measurement and Evaluation in Counseling and Development, 23,* 20–30.

Krosnick, J. A. (1999). Survey research. *Annual Review of Psychology, 50,* 537–567. doi: 10.1146/annurev.psych.50.1.537

Kuncel, N. R., Credé, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class rank, and test scores: A meta-analysis and review of the literature. *Review of Educational Research, 75,* 63–82. doi: 10.3102/00346543075001063

Kunin, T. (1955). The construction of a new type of attitude measure. *Personnel Psychology, 8,* 65–77. doi: 10.1111/j.1744-6570.1955 .tb01189.x

Loo, R., & Kells, P. (1998). A caveat on using single-item measures. *Employee Assistance Quarterly, 14,* 75–80. doi: 10.1300/J022v14n02–06

Luthans, F., Avolio, B., Avey, J., & Norman, S. M. (2007). Positive psychological capital: Measurement and relationship with performance and satisfaction. *Personnel Psychology, 60,* 541–572. doi: 10.1111/ j.1744-6570.2007.00083.x

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. doi: 10.1037/0003-066X.50.9.741

Miquelon, P., & Vallerand, R. J. (2008). Goal motives, well-being, and physical health: An integrative model. *Canadian Psychology/ Psychologie canadienne, 49,* 241–249. doi: 10.1037/a0012759

Mondak, J. J., Hibbing, M. V., Canache, D., Seligson, M. A., & Anderson, M. R. (2010). Personality and civic engagement: An integrative framework for the study of trait effects on political behavior. *American Political Science Review, 104,* 85–110. doi: 10.1017/S0003055409990359

Moreno, R., Martinez, R. J., & Muniz, J. (2006). New guidelines for developing multiple-choice items. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 2,* 65–72. doi: 10.1027/1614-2241.2.2.65

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology, 75,* 77–86. doi: 10.1348/096317902167658

O'Connor, M. C., & Paunonen, S. V. (2007). Big Five personality predictors of post-secondary academic performance. *Personality and Individual Differences, 43,* 971–990. doi: 10.1016/j.paid.2007.03.017

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology, 84,* 411–422. doi: 10.1037/0022-3514.84.2.411

Paunonen, S. V., & Jackson, D. N. (1985). The validity of formal and informal personality assessment. *Journal of Research in Personality, 19,* 331–342. doi: 10.1016/0092-6566(85)90001-7

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y., & Podsakoff, N. P. (2003). Common-method bias in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88,* 879–903. doi: 10.1037/0021-9010.88.5.879

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41,* 203–212. doi: 10.1016/ j.jrp.2006.02.001

Roberts, B. W., Bogg, T., Walton, K., Chernyshenko, O., & Stark, S. (2004). A lexical approach too identifying the lower-order structure of conscientiousness. *Journal of Research in Personality, 38,* 164–178. doi: 10.1016/S0092-6566(03)00065-5

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27,* 151–161. doi: 10.1177/0146167201272002

Rosnow, R. L., & Rosenthal, R. (1974). Taming of the volunteer problem: On coping with artifacts by benign neglect. *Journal of Personality and Social Psychology, 30,* 188–190. doi: 10.1037/h0036535

Russell, S. S., Spitzmueller, C., Fin, L. F., Stanton, J. M., Smith, P. C., & Ironson, G. H. (2004). Shorter can also be better: The abridged Job in General Scale. *Educational and Psychological Measurement, 64,* 878– 893. doi: 10.1177/0013164404264841

Salgado, J. (2002). The Big Five personality dimensions and counterproductive behaviors. *International Journal of Selection and Assessment, 10,* 117–125. doi: 10.1111/1468-2389.00198

Saucier, G. (1994). Mini-markers: A brief version of Goldberg's unipolar Big-Five markers. *Journal of Personality Assessment, 63,* 506–516. doi: 10.1207/s15327752jpa6303–8

Schmitt, N. (1996). The uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353. doi: 10.1037/1040-3590.8.4.350

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement, 9,* 367–373. doi: 10.1177/014662168500900405

Shafer, A. B. (1999). Brief bipolar markers for the five factor model of personality. *Psychological Reports, 84,* 1173–1179.

Sheldon, K. M., & Hoon, T. H. (2007). The multiple determination of well-being: Independent effects of positive traits, needs, goals, selves, social supports, and cultural context. *Journal of Happiness Studies, 8,* 565–592. doi: 10.1007/s10902-006-9031-4

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12,* 102–111. doi: 10.1037/1040-3590.12.1.102

Smith, P. C., Kendall, L. M., & Hulin, C. L. (1969). *The measurement of satisfaction in work and retirement.* Chicago, IL: Rand McNally.

Spector, P. (1992). *Summated rating scale construction: An introduction.* Newbury Park, CA: Sage

Stanton, J. M., Balzer, W. K., Smith, P. C., Parra, L. F., & Ironson, G. (2001). A general measure of work stress: The Stress in General Scale. *Educational and Psychological Measurement, 61,* 866 – 888. doi: 10.1177/00131640121971455

Stanton, J. M., & Weiss, E. M. (2002). *Online panels for social science research: An introduction to the StudyResponse Project* (Tech. Rep. No. 13001). Syracuse, NY: Syracuse University, School of Information Studies.

Stillman, T. F., Baumeister, R. F., Vohs, K. D., Lambert, N. M., Fincham, F. D., & Brewer, L. E. (2010). Personal philosophy and personnel achievement: Belief in free will predicts better job performance. *Social Psychological and Personality Science, 1,* 43–50. doi: 10.1177/ 1948550609351600

Thalmeyer, A. G., Saucier, G., & Eigenhuis, A. (2011). Comparative validity of brief and medium-length Big Five and Big Six personality questionnaires. *Psychological Assessment, 23,* 995–1009.

Vickers, R. R., Conway, T. L., & Hervig, L. K. (1990). Demonstration of replicable dimensions of health behaviors. *Preventive Medicine, 19,* 377–401. doi: 10.1016/0091-7435(90)90037-K

Wanous, J. P., Reichers, A. E., & Hudy, M. J. (1997). Overall job satisfaction: How good are single item measures? *Journal of Applied Psychology, 82,* 247–252. doi: 10.1037/0021-9010.82.2.247

Weng, L. (2004). Impact of the number of response categories and anchor labels on coefficient alpha and test–retest reliability. *Educational and Psychological Measurement, 64,* 956–972. doi: 10.1177/ 0013164404268674

Williams, L. J., & Anderson, S. E. (1991). Job satisfaction and organizational commitment as predictors of organizational citizenship and in-role behaviors. *Journal of Management, 17,* 601–617. doi: 10.1177/ 014920639101700305

Wood, D., Nye, C., & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive markers in the English lexicon. *Journal of Research in Personality, 44,* 258–272. doi: 10.1016/j.jrp.2010.02.003

Woods, S. A., & Hampson, S. E. (2005). Measuring the Big Five with single items using a bipolar response scale. *European Journal of Personality, 19,* 373–390. doi: 10.1002/per.542

Wu, K. D., & Clark, L. A. (2003). Relations between personality traits and self-reports of daily behavior. *Journal of Research in Personality, 37,* 231–256.

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality, 44,* 180–198. doi: 10.1016/j.jrp.2010.01.002