Evaluations Of Online Personality Scales

A number of personality scales have been implemented as online instruments, mainly by amateurs but also by professional psychologists (Barak & English, 2002; Epstein & Klinkenberg, 2001). To date, however, relatively few evaluations of the psychometric properties of online personality tests have been published. Of those accounts that have appeared in the literature, there are several that permit evaluation of the online tests either by inclusion of a paper-and-pencil comparison condition, or by comparison with findings obtained from some other equivalent method. The majority of these reports focus on single personality constructs rather than broad inventories. In general, the findings from such studies have shown that online measures can be reliable and valid measures of the intended constructs (see Buchanan, 2001, 2002, for reviews).

However, the picture becomes a little more complex when multi-factorial inventories are considered. Robins, Tracy, and Trzesniewski (2001), in an examination of links between self-esteem and other personality traits, used data obtained from an online version of a Big Five personality inventory (John & Srivastava, 1999) in a very large sample. They reported internal-consistency reliability estimates (coefficient α) for each of the five dimensions that were as high as those obtained with the paper-and-pencil version of the inventory. Woolhouse and Myers (1999) evaluated a new measure (based on a Jungian personality typology) using both paper-and-pencil and internet modes of administration, and similarly found that the reliabilities were comparable. However, they also found some differences in the latent structures for the two versions of the instrument; the factors on which some items had their highest loadings varied across the two methods. Fouladi, McCarthy, and Moller (2002) compared findings from three questionnaires measuring, respectively, attachment to father and to mother; perceived ability to reduce negative mood; and awareness of mood and mood regulation strategies. These were administered either online or in paper-and-pencil format, with traditionally recruited participants randomly assigned to conditions. Although psychometric properties and latent structures were in general very similar, there were some small differences between the internet-mediated and traditional versions of the instruments (e.g., mean scores on two of the mood-related scales differed significantly, there were differences in means and variances on a number of items, and there were some distributional differences). Fouladi et al. argued that although their findings demonstrate that internet administration of psychological measures is clearly viable, they also indicate the need for further evaluation and refinement of such measures.

Johnson (2000) created a web-mediated version of an International Personality Item Pool (IPIP: Goldberg, 1999) representation of the constructs in Costa and McCrae's Five Factor Model as embodied in their NEO-PI-R (Costa & McCrae, 1992). Factor analyses of Johnson's data showed that, broadly speaking, the latent structure was as expected but that there were some anomalies, with a small minority of the facet-level constructs loading most highly on the "wrong" domain construct.

The findings from studies that permit an evaluation of the psychometric properties of online personality tests (both single and multiple construct measures) led Buchanan (2002) to conclude that internet-mediated tests could be reliable and valid, but that online and offline versions of the same test "can be equivalent but are not always identical. One cannot take the psychometric properties of online instruments for granted, even if they are direct translations of traditional instruments" (p. 150). For confidence that one is using a satisfactory measure – especially in these early days of online testing (Robins et al., 2001) – the psychometric properties of any online test need to be established before any real weight can be attached to data derived from it.

The purpose of the current exercise was to develop an online instrument with demonstrably acceptable psychometric properties, for use in internet-mediated psychological research. Given the current popularity of the Five Factor Model (Costa & McCrae, 1992) as a representation of some central constructs of personality, it is likely that an instrument measuring these five dimensions – Extraversion (E), Agreeableness (A), Conscientiousness (C), Neuroticism (N), and Openness to Experience (O) – could be extremely useful for online research.

Instrument Selected

The instrument chosen was the 50-item IPIP representation of the domain constructs of the Five Factor Model, as expressed in Costa and McCrae's (1992) revised NEO personality inventory (NEO-PI-R). There are a number of reasons why this particular instrument was selected.

First, the NEO-PI-R is a very widely used inventory. There is an extensive literature on the extent to which the constructs embodied in this inventory relate to various behavioral criteria and other phenomena of psychological interest, and the scales have proven to be useful tools in a number of applied fields. The scales in the IPIP implementation have been shown to correlate highly with the corresponding NEO-PI-R domain scores, with correlations that range from .85 to .92 when corrected for unreliability (International Personality Item Pool, 2001). The IPIP scales also outperformed the NEO-PI-R versions of the same constructs as predictors of a number of

clusters of self-reported behavioral acts (Goldberg, in press), although these findings come from the same sample as was used to construct the IPIP scales.

Second, the IPIP representation (like all IPIP measures) is freely available in the public domain (Goldberg, 1999). With proprietary instruments, there are potential copyright and test security issues that might prevent their use on the internet. With public-domain instruments, these problems do not arise.

Finally, the instrument is relatively short. Web experiments are subject to relatively high dropout rates (Musch & Reips, 2000; Reips, 2000), partly due the fact that it is easier to leave a web experiment than one conducted in a traditional environment. Although there are a number of factors likely to affect dropout, such as financial incentives (Frick, Bachtiger & Reips, 2001; Musch & Reips, 2000) it is likely that longer questionnaires (such as those that include a long personality inventory followed by some other questions addressing topics of interest) will lead to larger numbers of people abandoning the study (Knapp & Heidingsfelder, 2001). Such attrition is potentially a serious issue because of the possibility of selective dropout. It is likely that those participants who drop out early will differ from those who continue to the end of the survey in traits such as conscientiousness and patience. This may limit the generalizability of the findings and bias the results of studies where independent or outcome variables are related to those characteristics that might affect a person's likelihood to terminate participation early. As a consequence, other things being equal, short scales are desirable for use online.

Validation Strategy

A measure of personality is of little use if it does not predict any behavioral outcomes. Accordingly, much effort has been invested in discovering the correlates of major personality dimensions, and a number of behaviors associated with each of the domains in the Five Factor Model dimensions have been documented in the literature. The primary validation strategy adopted in the current study was to ask participants for self-reports of the frequency with which they had engaged in each of those behaviors that have been linked to scores on one or more of the Five Factor dimensions. Correlations between these self reports and the relevant dimensions may be regarded as support for the contention that the scales measure the intended constructs.

Self-Reports of Criterion Behaviors

Booth-Kewley and Vickers (1994) found that traffic risk-taking was negatively related to Agreeableness (A) and Conscientiousness (C), and positively to Neuroticism (N). There is also evidence that N is related positively to traffic accidents (Schenke & Rausche, 1979) and C is negatively related to involvement in driving accidents (Arthur & Graziano, 1996). Two behavioral items related to these findings were constructed: "Had a speeding ticket or fine?" (Behavior 1) and "Been involved in a traffic accident which was at least partly your fault?" (Behavior 2). It was hypothesized that, among those respondents who have access to cars, N will be positively related to both of these behaviors, while A and C will be negatively related to them.

Heaven (1996) found that Agreeableness was negatively associated with interpersonal violence for both men and women. Therefore, a negative correlation was predicted between A and the behavioral item "Been involved in a fight?" (Behavior 3). Heaven (1996) also reported that Conscientiousness was negatively associated with vandalism for both men and women. We may thus expect a negative correlation between C and the behavioral item "Vandalized something or damaged the property of another?" (Behavior 4).

In introductory personality textbooks, a liking for parties and social situations is often presented as one of the defining characteristics of the extravert. Goldberg (in press) found a positive correlation ($r = .31$) between Extraversion (E) and the item "planned a party" in a large community sample. We, therefore, expected that E would correlate positively with the behavioral item "Planned a party?" (Behavior 5). Eysenck (1976) reported that extraverts tended to have more sexual partners and were more likely than introverts to report more than one sexual partner in the past year. It was, thus, hypothesized that in the current sample a positive correlation would be found between E and the behavioral item "Had sexual intercourse with a person other than your current (or most recent) partner?" (Behavior 6).

A number of studies have explored the associations between personality traits and smoking. Helgason, Fredrikson, Dyba, and Steineck (1995) found that extraverts smoked more and found it harder to quit. We, therefore, predicted a positive correlation between E and the behavioral item "Smoked tobacco (cigarettes, cigars, pipe)?" (Behavior 7).

McCrae and Costa (1997) consider artists to be "prime examples of individual high in Openness to Experience" (p. 825). Although the number of professional artists is limited, many people – as McCrae and Costa note – have artistic dispositions or interests, and may engage in production of artworks at an amateur level. Goldberg (in press) found that the item "produced a work of art" correlated positively ($r = .38$) with Openness (O) in a community sample. Accordingly, we predicted a positive correlation between O and the behavioral item "Created a

work of art (e.g., painting or sculpture)?" (Behavior 8) in the internet sample.

McCrae and Costa (1997) reported positive correlations between O and the needs for Sentience and Understanding assessed in the Personality Research Form, and suggested that "the intellectual interests of open men and women may lead them to seek higher levels of education" (p. 831). It was thus predicted that O would correlate positively with the behavioral item "Taken a course or class you did not have to, purely out of interest?" (Behavior 9).

In addition, Goldberg (in press) has found relationships between scores on the paper-and-pencil version of the scale to be used in this study and various behavioral acts in a large community sample. Some of the highest, which are clearly conceptually related to the dimensions with which they correlate, were selected for use in this study. We predicted that N would correlate positively with the behavioral item "Taken medication for depression?" (Behavior 10), that E would correlate positively with the behavioral item "Started a conversation with strangers?" (Behavior 11), that O would correlate positively with "Attended an art exhibition?" (Behavior 12), that A would correlate negatively with "Made fun of someone?" (Behavior 13), and that C would correlate negatively with "Let work pile up until just before a deadline?" (Behavior 14). The correlations from Goldberg (in press) between these dimensions and behaviors were, respectively, .38, .34, .48, –.32, and –.29.

Other Predictions

In addition to these behavioral acts, a number of demographic and other variables accessible through self-reports are known to be linked to one or more of the five factors. If the same associations are found in the current sample, this may also be interpreted as evidence for construct validity.

A number of sex differences in personality traits have been reported (Budaev, 1999). Normative data on the domain constructs of the NEO-PI-R by Costa and McCrae (1992) show some such differences. The most marked are with N and A, with values of Glass's *d* around the magnitude corresponding to a medium effect size in Cohen's (1992) terms; men score lower on both N (d = .38) and A (d = .55) than do women[1]. One would expect to find the same pattern here.

As already noted, McCrae and Costa (1997) suggested that people high on O will be more likely to seek out educational experiences. In addition to the education-seeking behavior (Behavior 9), one might expect high O scorers to have higher educational levels. Similarly, there is evidence (De Raad & Schouwenburg, 1996) that people high in C are likely to have higher levels of education. Therefore, one might hypothesize that among participants who do not report their occupations as "student" (thereby implying that their formal education has finished at least temporarily), both O and C will be positively associated with the highest level of formal education completed.

Negative associations have been reported between N and job satisfaction, with higher scorers tending to be less happy with their jobs, whatever those jobs may be (Perone, DeWaard, & Baron, 1979). Although job satisfaction may be best characterized as multidimensional in nature (e.g., satisfaction with elements such as salary, relationships with workmates, or working environment), it seems reasonable to predict that people scoring high on N will report less satisfaction with their jobs in general. Associations between N and health were also expected. Costa and McCrae (1992) suggested that "N is a potent predictor of somatic complaints" (p. 35), and Vollrath, Knoch, and Cassano (1999) reported that people high on N tend to worry more about their health. Finally, links between N and mental health are also well documented (e.g., Miller, 1991). It was, therefore, predicted that N would be negatively correlated with self-reports of both physical and mental health.

In summary, then, the purpose of Study 1 was to establish whether an internet-mediated version of an IPIP Five Factor Inventory was psychometrically acceptable in terms of factor structure and reliability. In addition, a preliminary assessment of its validity was conducted by testing 25 hypothesized relations between the five factors and self-reports of behaviors, demographic, and other variables.

Study 1: Materials and Methods

Materials

An internet version of the short IPIP instrument measuring the domain constructs of the Five Factor model was created as a series of web pages. All pages were dynamically generated in the scripting language Perl. The scripts generated the code to display the pages in the participant's browser, and to log their responses (and other contextual information: date, time, browser type, and internet address of the computer they were using) to our data-files.

Participants first saw a page with a brief description of the inventory, explaining that it was part of a research

1 These calculations are based on the data presented by Costa and McCrae (1992) in Appendix B, Table B–1 "Means and Standard Deviations for Form S for Adults" (p. 75).

project. On clicking a button to indicate that they consented to participate, they then saw a second page which included these instructions: "On the following pages, there are phrases describing people's behaviors. Please use the rating scale below to describe how accurately each statement describes you. Describe yourself as you generally are now, not as you wish to be in the future. Describe yourself as you honestly see yourself, in relation to other people you know of the same sex as you are, and roughly your same age."

The participants responded to each of the 50 items in the inventory by clicking on one of five radio buttons, labeled *Very Inaccurate*, *Moderately Inaccurate*, *Neither Accurate nor Inaccurate*, *Moderately Accurate*, or *Very Accurate*. In addition, they answered demographic questions about their age, gender, and amount of education, as well as questions about their occupations. Finally, they responded to items asking for ratings of job satisfaction, mental and physical health, and each of the 14 behavioral acts described above. The behavioral items were answered using a drop-down menu next to each item, with response options reading *Never in my life*, *Not in the past year*, *ONCE or TWICE in the past year*, *THREE OR MORE TIMES in the past year*, *MORE THAN 15 TIMES IN THE PAST YEAR*, and *Prefer not to answer*.

After responding to the items, respondents were then asked to click on another button to submit their data. Any missing responses were detected at this point, and participants who had not answered all the questions were requested to go back and complete the missing items. Those who had completed all the items then saw a debriefing page thanking them for their participation, and providing their scores on each of the scales. In addition, they were given information to help interpret the scores, including a brief description of the meaning of each of the scales, and normative information about their scores relative (top third, middle, bottom third) to those of others who had completed the inventory to date. They were told that because this instrument was still under development, they should not attach too much weight to the scores. Links were provided to contact the experimenters, and to information about personality research elsewhere on the internet.

Participants

Between April and August of 1999, 2,875 data submissions were recorded. After the deletion of possible multiple submissions (described below), 2,448 response protocols were analyzed in this study. Of these, 991 (40.5%) were from men and 1,457 (59.5%) were from women. The median age group was 21–25 years, with about 70% between 16 and 30 years in reported age. Approximately two-thirds of the participants had obtained at least some post-compulsory education (e.g., college). More than half of the respondents (58%) indicated that they lived in the United States, 20% in Europe, and the remainder in countries throughout the world. In response to the question about their occupations, 46% indicated that they were presently employed, and 40% were currently students.

Procedures

No attempts were made to actively solicit participants for the study. Instead, the address of the survey's web page was submitted to a number of search engines (including Altavista, Yahoo, and Lycos), and our research participants entered requests like "personality test" in one of these search engines.

Two of the main threats to the integrity of internet-mediated data collection are multiple submissions from the same respondent and spurious or mischievous data entries (Buchanan, 2000; Buchanan & Smith, 1999b; Johnson, 2001; Schmidt, 1997). The former can be dealt with by inspecting the data-file for multiple submissions from the same internet address and discarding all but the first submission from each. This conservative[2] strategy is bound to result in the loss of some genuine data (e.g., individuals sharing a computer at home or in a classroom, or using an internet Service provider that dynamically assigns IP numbers), but it increases one's confidence in the independence of the observations (Buchanan, 2000). The latter can be minimized by looking for unrealistic patterns of demographic data (e.g., persons under the age of 16 claiming to have doctoral degrees; those claiming to live in Antarctica). Experience and empirical reports (e.g., Stieger & Voracek, 2001) suggest that the percentage of persons who provide such misrepresentations is quite low.

Based on their internet addresses, 425 records were identified as possible multiple submissions, and all of these were excluded from our analyses. Examination of the data-file for implausible patterns of demographic information led to the exclusion of two more protocols, in

2 This is not a maximally conservative technique, as it does not detect participants submitting data from different computers or IP addresses. Estimates of the extent to which multiple submissions occur varies across studies (e.g, Buchanan, 2000). Ways of estimating the degree of potential data contamination have been developed (e.g., the subsampling technique described by Reips, 2000), but they are difficult to implement with large datasets and require participants to provide identifying information such as e-mail addresses. This is a problem for which more research is required to develop better control procedures.

both of which a person claimed to be between 11 and 15 years of age and to have obtained a postgraduate degree.

Results

To establish whether the expected five-factor structure of the 50 IPIP items was present in the current data-set, five principal components were extracted and rotated using the Varimax procedure. The factor loadings are presented in Table 1.

Of the 50 items, 48 had their highest loadings on their expected factors, and only five items had loadings of .30 or higher on any of the other factors. The robustness of the original solution seems quite surprising, given that these scales were developed from the responses of a sample in one community of one state in the U.S. and were here being used by persons from many cultures and nations throughout the world.

However, given that the purpose of this project was to develop a robust online instrument that measures the domain constructs of the Five-Factor Model with a high degree of fidelity, it seemed appropriate to exclude problematic items and, thus, create a new set of more highly factor-univocal scales. For this purpose, we adopted the procedure described by Saucier (1994). Saucier's most conservative definition of a factor-pure item was one that not only had its highest loading on the expected factor but also had a loading on that factor that was at least double the loading on any other factor. The factor loadings were examined to identify those items that met these criteria for factor-purity in the present internet sample.

As noted, two items (one from A+ and one from O+) had their highest loadings on the "wrong" factors. In addition, seven others had a secondary loading that was greater than half the highest loading (two from A–, two from N–, two from O–, and one from E–). In the analyses that follow, we present findings using both the original scales and those excluding the nine problematic items. Table 2 provides internal consistency reliability estimates (coefficient α) for both sets of scales in the present sample, along with the corresponding reliabilities of the original scales obtained by Goldberg (International Personality Item Pool, 2001) in the scale-development sample.

For three of the original scales (E, N, and A), the reliability estimates in the internet sample were virtually identical to those in the offline U.S. sample. For the O scale, reliability was reduced in the internet sample (from .82 to .76), whereas for C it was somewhat augmented (from .81 to .84).

Table 3 presents the intercorrelations among the scales in each of the two versions, as well as the correlations between the two sets of scales. For the original scales, the intercorrelations ranged from .02 (C vs. O) to –.41 (A vs. N), and averaged .22 (with the signs for N reflected). These values are virtually identical to those from the original sample in which the scales were initially developed; in that sample, the intercorrelations ranged from .01 (C vs. O) to –.43 (A vs. N), and averaged .23. For the revised scales, the correlations ranged from .00 (N vs. O) to –.36 (N vs. A), and averaged .19. Because the dimensions of the Five-Factor Model are assumed to be orthogonal in theory, the revised scales are slightly more in accord with this expectation.

To assess the extent to which the scales measure their intended constructs, the predicted associations with self-reported behavioral acts were analyzed. Although the Pearson product-moment correlation (r) is the index of association that is most commonly used for this purpose, we also calculated Spearman rank-difference (nonparametric) correlations (r–s), given that the frequencies for many of the behavioral acts are unlikely to have interval measurement properties. For comparison purposes, both types of correlation coefficients are provided. The correlations between the scale scores and the criteria with which they should be associated are summarized in Table 4. Because of missing data, the sample sizes in Table 4 vary depending on the criterion.[3] All the correlations are statistically significant ($p < .05$) and in the expected direction.

We predicted that men would score lower than women on both N and A. Using the revised scales, the mean scores on N were 21.6 ($SD = 6.8$) for men and 23.0 ($SD = 6.7$) for women ($t_{2446} = -5.01$, $p < .0005$); the mean scores on A were 25.2 ($SD = 64.8$) for men and 26.8 ($SD = 4.6$) for women ($t_{2446} = -8.37$, $p < .0005$). Using the full scales, the mean scores on N were 27.4 ($SD = 8.1$) for men and 29.1 ($SD = 8.0$) for women ($t_{2446} = -5.11$, $p < .0005$); the mean scores on A were 34.3 ($SD = 6.3$) for men and 36.2 ($SD = 6.2$) for women ($t_{2446} = -7.60$, $p < .0005$). As expected, men did indeed score significantly lower than women on both N and A.

Included in Table 4 are two types of association indices, one parametric (Pearson r) and one nonparametric (Spearman r–s). The correlations are all virtually identical in size regardless of the index used. Also included in the table are comparisons between the original scales and

3 For the driving-related criteria, the analyses were restricted to those individuals who indicated having at least one car accident or traffic violation, thereby excluding those who did not drive; because those persons with perfect driving records were also thus excluded, we can assume that the present correlations are at least somewhat attenuated when compared to those from a sample including the full range of driving behaviors.

Table 1. Factor loadings of all 50 items on each of the five varimax-rotated components.

Item		Expected Factor	Extracted Factor 1	2	3	4	5
8	Am the life of the party.	E+	**.76***	−.04	−.03	.02	.01
9	Am skilled in handling social situations.	E+	**.75***	−.13	.18	.14	−.05
21	Make friends easily.	E+	**.71***	−.12	.04	.25	.03
23	Know how to captivate people.	E+	**.70***	−.02	.10	−.02	−.15
47	Feel comfortable around people.	E+	**.67***	−.27	.14	.26	.02
39	Don't talk a lot.	E−	**−.70***	.06	.04	.08	.07
34	Keep in the background.	E−	**−.66***	.15	−.02	.08	.11
50	Have little to say.	E−	**−.61***	.11	−.05	.07	.26
17	Don't like to draw attention to myself.	E−	**−.57***	.08	.11	.21	.10
15	Would describe my experiences as somewhat dull.	E−	**−.37***	.28	−.16	−.07	.23
43	Often feel blue.	N+	−.22	**.78***	−.15	−.08	−.02
40	Am often down in the dumps.	N+	−.22	**.77***	−.16	−.10	−.02
12	Dislike myself.	N+	−.24	**.66***	−.19	−.05	.01
2	Have frequent mood swings.	N+	.07	**.65***	−.08	−.16	−.05
29	Panic easily.	N+	.02	**.55***	−.11	−.07	.07
16	Seldom feel blue.	N−	.15	**−.63***	.09	.12	.08
35	Feel comfortable with myself.	N−	.29	**−.62***	.19	.09	−.03
25	Am very pleased with myself.	N−	**.35**	−.58*	.17	.07	.04
49	Rarely get irritated.	N−	−.07	**−.47***	.00	**.40**	−.01
3	Am not easily bothered by things.	N−	−.01	**−.42***	−.04	.21	−.02
10	Am always prepared.	C+	.03	−.04	**.70***	.05	.10
11	Make plans and stick to them.	C+	.02	−.03	**.70***	.04	.12
18	Carry out my plans.	C+	.07	−.11	**.70***	.05	.04
38	Get chores done right away.	C+	.03	−.01	**.61***	.10	.12
33	Pay attention to details.	C+	.07	.04	**.41***	.05	−.11
27	Find it difficult to get down to work.	C−	−.04	.15	**−.67***	−.08	.00
26	Do just enough work to get by.	C−	.04	.10	**−.64***	−.07	.06
41	Shirk my duties.	C−	−.03	.20	**−.62***	−.15	.05
46	Don't see things through.	C−	−.04	.18	**−.62***	−.05	.07
36	Waste my time.	C−	−.09	.21	**−.61***	−.09	−.02
45	Have a good word for everyone.	A+	.19	−.04	.05	**.63***	−.07
13	Respect others.	A+	.14	−.05	.24	**.60***	−.06
24	Believe that others have good intentions.	A+	.22	−.15	.06	**.58***	−.03
31	Accept people as they are.	A+	.11	−.06	.05	**.55***	−.10
48	Make people feel at ease.	A+	**.58***	−.07	.12	**.42**	.01
44	Cut others to pieces.	A−	.09	.21	−.12	**−.65***	.09
14	Insult people.	A−	.12	.15	−.20	**−.62***	.05
20	Have a sharp tongue.	A−	**.34**	.08	.03	**−.52***	−.09
37	Get back at others.	A−	.17	.16	−.14	**−.50***	.13
4	Suspect hidden motives in others.	A−	−.03	.27	.00	**−.36***	.01
6	Believe in the importance of art.	O+	.08	.08	.04	.12	−.76*
7	Have a vivid imagination.	O+	.24	.05	.01	−.01	−.48*
5	Enjoy hearing new ideas.	O+	.14	−.04	.07	.23	**−.43***
22	Tend to vote for liberal political candidates.	O+	.05	.05	−.08	.04	**−.39***
28	Carry the conversation to a higher level.	O+	**.39***	−.06	.05	−.04	**−.37**
42	Do not like art.	O−	−.04	.02	−.07	−.11	**.73***
32	Do not enjoy going to art museums.	O−	−.03	.00	−.09	−.12	**.67***
19	Am not interested in abstract ideas.	O−	.01	.04	.04	.00	**.63***
30	Avoid philosophical discussions.	O−	−.04	.08	.00	.06	**.55***
1	Tend to vote for conservative political candidates.	O−	−.01	−.04	.13	.01	**.39***

Note. Loadings of .30 or higher are listed in **bold** type. * Highest loading for each item.

Table 2. Internal consistency reliability (Coefficient α) for the revised and original scales in the internet sample, and the original scales in Goldberg's (International Personality Item Pool, 2001) U.S. Sample.

		Revised scales	Original scales
Scale	Items in Revised Scale	Internet sample	U.S. sample
E	8+, 9+, 21+, 23+, 47+; 17–, 34–, 39–, 50–	.88	.85	.86
N	2+, 12+, 29+, 40+, 43+; 3–, 16–, 35–	.83	.86	.86
C	10+, 11+, 18+, 33+, 38+; 26–, 27–, 36–, 41–, 46–	.84	.84	.81
A	13+, 24+, 31+, 45+; 14–, 37–, 44–	.76	.77	.77
O	6+, 22+; 1–, 19–, 30–, 32–, 42–	.74	.76	.82

Note. The original and revised scales for C are identical, as no items were dropped from that scale.

Table 3. Intercorrelations among and between the original and revised scales.

	Revised Scales					Original Scales				
	E'	N'	C'	A'	O'	E	N	C	A	O
E'	–	**–.32**	.14	.12	.16	**.99***	**–.33**	.14	.14	.25
N'	**–.32**	–	**–.34**	**–.36**	.00	**–.35**	**.99***	**–.34**	**–.38**	–.04
C'	.14	**–.34**	–	**.31**	–.01	.16	**–.34**	**1.00***	**.30**	.02
A'	.12	**–.36**	**.31**	–	.16	.14	**–.38**	**.31**	**.96***	.18
O'	.16	.00	–.01	.16	–	.17	–.01	–.01	.14	**.96***
E	**.99***	**–.35**	.16	.14	.17	–	**–.35**	.16	.16	.26
N	**–.33**	**.99***	**–.34**	**–.38**	–.01	**–.35**	–	**–.34**	**–.41**	–.05
C	.14	**–.34**	**1.00***	**.31**	–.01	.16	**–.34**	–	**.30**	.02
A	.14	**–.38**	**.30**	**.96***	.14	.16	**–.41**	**.30**	–	.15
O	.25	–.04	.02	.18	**.96***	.26	–.05	.02	.15	–

Note. $N = 2448$. No items were omitted from C, so the two versions are the same. Part-whole correlations between the same scales in the two versions are indicated by an asterisk. Correlations of .30 or more are indicated in bold. Abbreviations for the revised scales are indicated by an apostrophe.

shortened versions that are more homogeneous in content. Again, differences in their criterion correlations were typically minuscule, and when any differences appeared, they tended to favor the longer original scales.

Some of the correlations observed, although statistically significant and psychologically meaningful, are low in magnitude and some of the effect sizes are very low. There is a danger that the large sample sizes, and consequent high statistical power, obtainable in online research may lead to trivial associations emerging as significant relations in correlational studies such as the present one. However, there are some reasons to suggest that the correlations we observe here are not trivial. Instead, the low magnitude of some of the correlations can be attributed to some methodological features of our research design.

One such reason is that many of the behavioral acts are of low base-rate, thus providing frequency estimates that suffer from restriction of range, which will attenuate any correlations with other variables. For example, being involved in a traffic accident (which has low correlations with relevant personality variables) is a much rarer event than starting a conversation with a stranger (which has higher correlations with the relevant personality variable). An additional complication is the inherent unreliability of single-item measures; all correlations with single-act criteria are likely to be attenuated for this reason. When personality scales are correlated with behavioral information that has been aggregated across classes of trait-relevant behaviors rather than single instances, stronger links are observed (e.g., Mischel & Shoda, 1999). As would be expected, this is the case in the present study.

Aggregated behavioral measures were constructed by averaging the frequencies of behavioral acts relevant to each of the personality domains (first reverse-scoring where necessary so that all predicted relations were in the same direction, and converting the score on each act to a percentage so that all were on the same scale and carried equal weights in the analyses). Correlations between the resulting aggregated values and the participants' personality scale scores are presented in the top half of Table 5.

The aggregated behavioral indices theoretically relevant to each of the five domains were all strongly associated with the scale scores ($p < .001$), and these relations were much the same for the original and the revised scales and for each of the two indices of association. Of the five domains, the correlations were highest for Neuroticism: The aggregate index relevant to N (speeding,

Table 4. Correlations (Pearson *r* and Spearman *r*-s) between the scales and the criteria.

Scale	Criterion	Expected Direction	Revised Scales		Original Scales		*N*
			r	*r*–s	*r*	*r*–s	
E	Beh. 5: Planned party	+	.39	.38	.40	.39	2426
E	Beh. 6: Sex With Another Partner	+	.13	.12	.14	.14	2293
E	Beh. 8: Smoked Tobacco	+	.12	.12	.13	.13	2411
E	Beh. 11: Started Conversation	+	.45	.45	.46	.46	2421
N	Beh. 1: Speeding	+	.08	.09	.07	.08	1070
N	Beh. 2: Traffic Accident	+	.06	.07	.06	.07	1041
N	Beh. 10: Depression Medication	+	.28	.28	.28	.28	2411
N	Job Satisfaction	–	–.27	–.27	–.27	–.27	1137[b]
N	Physical Health	–	–.29	–.29	–.29	–.29	2448
N	Mental Health	–	–.65	–.65	–.65	–.65	2448
C	Beh. 1: Speeding	–	–.14	–.14	–.14	–.14	1070
C	Beh. 2: Traffic Accident	–	–.08	–.08	–.08	–.08	1041
C	Beh. 4: Vandalism	–	–.26	–.26	–.26	–.26	2409
C	Beh. 14: Let Work Pile Up	–	–.50	–.50	–.50	–.50	2400
C	Education	+	.12	.11	.12	.11	1475[a]
A	Beh. 1: Speeding	–	–.15	–.14	–.15	–.14	1070
A	Beh. 2: Traffic Accident	–	–.14	–.13	–.12	–.12	1041
A	Beh. 3: Fighting	–	–.28	–.25	–.29	–.26	2394
A	Beh. 13: Made Fun	–	–.38	–.39	–.39	–.39	2389
O	Beh. 7: Created Artwork	+	.36	.36	.37	.38	2420
O	Beh. 9: Took Extra Class	+	.22	.22	.23	.23	2411
O	Beh. 12: Attended Exhibition	+	.48	.48	.47	.47	2427
O	Education	+	.16	.17	.15	.16	1475[a]

[a] Analysis restricted to nonstudent participants. [b] Analysis restricted to those participants who were employed for wages. All correlations are significantly different from zero ($p < .05$).

traffic accident, medication for depression, physical health, mental health, job satisfaction) correlated above .55 with the N scale scores. For the other four domains, the correlations averaged ranged from .41 to .48 and averaged around .45. These substantial correlations with the aggregated behavioral acts provide some important preliminary evidence on the construct validity of these scales in an internet context.

Discussion

Both the original and the revised scales appear to have some degree of validity as measures of their intended constructs; of the 25 hypothesized links with various criteria (the 23 listed in Table 4 plus those related to gender), results consistent with all 25 were found.

One question that remains is whether the instrument will also prove useful with other kinds of samples recruited via the internet. Buchanan and Smith (1999b) noted that a possible problem for internet-mediated personality assessments (especially in the context of research) was the volunteer status of participants; their voluntary participation may be indicative of a higher degree of self-selection than is the case in other assessment contexts. This is especially true of the sample reported here, given that these individuals were actively seeking out personality tests that they could complete online. Reips (2000) discussed some of the problems that may arise from such self-selection and suggested a technique whereby any bias arising from self-selection may be identified, thus informing subsequent analyses. This "multiple site entry technique" involves comparing samples recruited either from different internet sites or in different ways that are associated with different degrees of, or motivations for, self-selection. If similar findings are obtained with each, one has grounds to argue that no systematic biases have been introduced as a function of self-selection.

In any study – especially one conducted via the internet – there is bound to be some degree of self-selection. However, using an active, rather than a passive, recruiting technique is likely to modify the degree of self-selection involved. It is likely that the motivation to participate will be different for people recruited by experimenters compared to those who recruit themselves (as in

Study 1). There is evidence that individuals who enter experiments through such different routes to participation may produce different patterns of behavior; for example, Oakes (1972) demonstrated that findings obtained with traditional student samples could differ from those obtained with true volunteers recruited from the wider community (see also Rosenthal & Rosnow, 1975).

Accordingly, a second study was conducted with participants recruited through an active advertising campaign. Although these will still be self-selected participants showing an interest in testing, the degree of self-selection differs from Study 1. If the psychometric properties of the inventory remain similar across the two samples, there are grounds to suggest that any biases arising through self-selection have not impacted upon the current study in an important manner.

Study 2: Materials and Methods

Materials

The materials used were identical to those in Study 1, with the exception of a banner advertisement that was used to recruit participants. This was a simple white banner, 468 pixels wide by 60 high, bearing the text: "Want to take a free personality test? Explore your personality. Help psychological research. Click here now." Clicking on the advertisement took participants directly to the informed-consent page as described in Study 1.

Participants

During April 2000, 264 submissions of data were recorded as responses to the recruiting advertisement. After the deletion of possible multiple submissions 249 remained. Of these, 114 (45.8%) were male and 135 (54.2%) were female. The median age was 21–25 years, with 54% reporting their age as being between 16 and 30 years. The majority of participants (61%) had at least some post-compulsory education (e.g., college). Two-thirds of the respondents indicated that they lived in the U.S., 12% lived in Europe, and the remainder in other locations. In response to the question about their main occupation, approximately half of the sample indicated that they were employed for wages, and 28.5% were currently students.

Procedures

Participants were recruited from the Link Exchange Banner Network, in which banners are displayed on the pages of randomly selected websites belonging to the advertising network (over 400,000 sites at the time the campaign was conducted). During the period of the campaign our banner was displayed some 140,000 times.

The data were examined and problematic submissions identified using the same criteria as in Study 1. Fourteen possible multiple submissions by the same individuals were identified, and one possible misrepresentation (a person claiming to be in the 16–20 age group and to have "some postgraduate education"). Although this is possible, this participant was excluded for reasons of consistency and caution. In all, fifteen problematic records were thus excluded from analysis.

Results and Discussion

The internal-consistency reliability estimates of the revised scales were calculated for the new sample and were found to be very similar to those obtained in Study 1. The coefficient α values for E, N, C, A, and O were respectively .89, .83, .84, .74, and .71 (corresponding values for the full scales were .88, .85, .84, .78, and .74).

As a check on whether the scales are related in the expected way to the behavioral criteria in the current sample, the personality relevant behaviors were again aggregated and correlations with the relevant personality scales computed. (Only the aggregated criteria were used for reasons of parsimony and statistical power.) The correlations, shown in the lower half of Table 5, indicate that once again each of the scales correlated significantly with the aggregated criteria it should theoretically predict. The pattern of associations – and indeed their magnitude – was very close to that observed in Study 1.

Table 5. Correlations of the personality scales with the aggregated criteria.

	Revised Scales		Original Scales		
Scale	*r*	*r*–s	*r*	*r*–s	*N*
Study 1					
E	.41	.42	.42	.43	2277
N	.56	.56	.57	.57	637
C	.41	.41	.41	.41	583
A	.42	.42	.43	.44	696
O	.45	.47	.45	.48	1445
Study 2					
E	.47	.44	.49	.47	231
N	.53	.51	.54	.48	79
C	.39	.40	.39	.40	74
A	.45	.44	.44	.44	85
O	.50	.48	.51	.48	176

Note. All correlations are significantly different from zero ($p < .001$).

Consequently, it appears that the psychometric properties of the inventory, in terms of the reliabilities of the scales and their observed associations with classes of expected correlates, are relatively unchanged across the two samples. There is, thus, no evidence that the findings of Study 1 are biased by the higher degree of self-selection associated with the recruiting technique used, or that the inventory would not "work" when other recruiting techniques were used.

General Discussion

Across the two studies, the power of the revised and full scales as predictors of behavioral acts seem comparable. In Study 1, the mean correlations (Pearson's *r*) between the various scales and criteria presented in Table 4 were the same to two decimal places (.26) for the full and revised scales. Table 5 further demonstrates that the correlations between the scales and relevant aggregated criteria are very similar for the revised and full versions of the scales. Also, as shown in Table 2, the internal consistencies of the full and revised versions of the scales are very similar. So why bother creating or using factor-univocal revisions of the scales at all?

One reason is that the revised versions of the scales are very slightly closer to orthogonality than the originals. This is clearly desirable in measures that are supposed to represent orthogonal constructs, as the domains of the Five-Factor Model are held to be (Saucier, 2002).

Another advantage conferred by using the revised versions of the scales is length: the revised inventory is almost 20% shorter than the original. Given that shorter instruments are desirable for web-mediated research – as longer inventories are more prone to participant dropout – the revised inventory is preferable for that reason. Although the two versions of the inventory seem largely equivalent in terms of their ability to measure the desired constructs, use of the shorter revised scales might confer a functional advantage in the context of online research.

The modified IPIP inventory evaluated here appears to have satisfactory psychometric properties as a brief online measure of the domain constructs of the Five-Factor Model. Across two studies using different recruiting techniques, acceptable levels of internal reliability and significant correlations with relevant criterion variables were observed. We, therefore, consider that it is appropriate for use in online research projects where measures of these variables are desired. Given the appropriate technical expertise, the information presented in Tables 1 and 2 should be sufficient to permit implementation of either the original or modified version. However, given the fact that not all internet samples will be the same – and indeed are likely to be quite heterogeneous, depending on the recruiting techniques used – researchers using this set of scales should consider examining its psychometric properties in their own samples.

The findings from this project may be interpreted as adding to the evidence that valid internet-mediated assessments of personality attributes are quite feasible. They also add to the evidence that online and offline versions of the same tests may not be entirely equivalent, and especially that the latent structures of multi-factorial inventories may change subtly when administered via the internet. Although in this case the changes did not seem to adversely affect the online implementation of the original inventory's power as a predictor of relevant criteria in comparison to the revised version, it is clear that one cannot simply mount an existing questionnaire on the world wide web and assume that it will be exactly the same instrument.

The reasons for these differences may include different interpretations of the item content by respondents from different countries or cultures, and some measure of increased self-disclosure associated with computer-mediated communication. Understanding more about the basis for these differences will inform both theory and practice, and, thus, this is an extremely useful topic for internet-oriented personality researchers to address. Additional issues that might be investigated in further research include the feasibility of using longer inventories to measure the lower-level facets of the Five-Factor Model, as well as studies of the practical importance of detecting invalid protocols.

Acknowledgments

Support for the third author was provided by Grant MH49227 from the National Institute of Mental Health, U.S. Public Health Service. Some of the findings from Study 1 were presented at the Meeting of the Society for Computers in Psychology, Los Angeles, CA, November 1999. We gratefully acknowledge Herbert W. Eber, Sarah E. Hampson, Ulf-Dietrich Reips, and Gerard Saucier for their thoughtful suggestions and comments on an earlier draft of this article.

References

Arthur, W., & Graziano, W.G. (1996) The five-factor model, conscientiousness, and driving accident involvement. *Journal of Personality, 64*, 593–618.

Barak. A., & English, N. (2002). Prospects and limitations of psychological testing on the internet. *Journal of Technology in Human Services, 19*, 65–89.

Bartram, D. (1998, January). *Distance assessment: Psychological*

assessment through the internet. Paper presented at the 1998 British Psychological Society Division of Occupational Psychology Conference, Eastbourne, UK.

Birnbaum, M.H. (Ed.). (2000). *Psychological experiments on the internet.* San Diego, CA: Academic Press.

Booth-Kewley, S., & Vickers, R.R. (1994). Associations between major domains of personality and health behavior. *Journal of Personality, 62,* 281–298.

Buchanan, T. (2000). Potential of the internet for personality research. In M.H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 121–140). San Diego, CA: Academic Press.

Buchanan, T. (2001). Online personality assessment. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 57–74). Lengerich, Germany: Pabst Science Publishers.

Buchanan, T. (2002). Online assessment: Desirable or dangerous? *Professional Psychology: Research and Practice, 33,* 148–154.

Buchanan, T., & Smith, J.L. (1999a). Research on the internet: Validation of a world-wide web mediated personality scale. *Behavior Research Methods, Instruments, & Computers, 31,* 565–571.

Buchanan, T., & Smith, J.L. (1999b). Using the internet for psychological research: Personality testing on the world-wide web. *British Journal of Psychology, 90,* 125–144.

Buckley, N., & Williams, R. (2002). Testing on the web – Response patterns and image management. *Selection & Development Review, 18,* 3–8.

Budaev, S.V. (1999). Sex differences in the Big Five personality factors: Testing an evolutionary hypothesis. *Personality and Individual Differences, 26,* 801–813.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159.

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.

De Raad, B., & Schouwenburg, H.C. (1996). Personality in learning and education: A review. *European Journal of Personality, 10,* 303–336.

Epstein, J., & Klinkenberg, W.D. (2001). From Eliza to internet: A brief history of computerized assessment. *Computers in Human Behavior, 17,* 295–314.

Eysenck, H. (1976). *Sex and personality.* London: Open Books Publishing Ltd.

Fouladi, R.T., McCarthy, C.J., & Moller, N.P. (2002). Paper-and-pencil or online? Evaluating mode effects on measures of emotional functioning and attachment. *Assessment, 9,* 204–215.

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.J. 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. (in press). The comparative validity of adult personality inventories: Applications of a consumer testing framework. In S.R. Briggs, J.M. Cheek, & E.M. Donahue (Eds.), *Handbook of adult personality inventories.* New York: Plenum.

Gosling, S.D., Vazire, S., Srivastava, S., & John, O.P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59,* 93–104.

Heaven, P. (1996). Personality and self-reported delinquency: Analysis of the "Big Five" personality dimensions. *Personality and Individual differences, 20,* 47–54.

Helgason, A.R., Fredrikson, M., Dyba, T., & Steineck, G. (1995). Introverts give up smoking more often than extraverts. *Personality and Individual Differences, 18,* 559–560.

International Personality Item Pool. (2001). A scientific collaboratory for the development of advanced measures of personality traits and other individual differences (http://ipip.ori.org/). Internet web site.

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, March). *Web-based personality assessment.* Poster session presented at the 71st Annual Meeting of the Eastern Psychological Association, Baltimore, MD.

Johnson, J.A. (2001, May). *Screening massively large data sets for nonresponsiveness in web-based personality inventories.* Invited talk to the joint Bielefeld-Groningen Personality Research Group, University of Groningen, The Netherlands. Retrieved from http://www.personal.psu.edu/~j5j/papers/screening.html.

Joinson, A.N. (1999). Anonymity, disinhibition, and social desirability on the internet. *Behavior Research Methods, Instruments, and Computers, 31,* 433–438.

Joinson, A.N., & Buchanan, T. (2001). Doing educational psychology research on the web. In C. Wolfe (Ed.), *Teaching and learning on the world wide web* (pp. 221–242). San Diego, CA: Academic Press.

Knapp, F., & Heidingsfelder, M. (2001). Drop-out analysis: Effects of the survey design. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of internet science* (pp. 221–230). Lengerich, Germany: Pabst Science Publishers.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of the Board of Scientific Affairs' Advisory Group on the conduct of research on the internet. *American Psychologist, 59,* 105–117.

McCrae, R.R., & Costa, P.T. (1997). Conceptions and correlates of openness to experience. In R. Hogan, S.R. Briggs, & J.A. Johnson (Eds.), *Handbook of personality psychology* (pp. 825–847). San Diego, CA: Academic Press.

Miller, T. (1991). The psychotherapeutic utility of the five-factor model of personality: A clinician's experience. *Journal of Personality Assessment, 57,* 449–464.

Mischel, W., & Shoda, Y. (1999). Integrating dispositions and processing dynamics within a unified theory of personality: The cognitive-affective personality system. In L.A. Pervin & O.P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 197–218). New York: Guilford.

Musch, J., & Reips, U.-D. (2000). A brief history of web experimenting. In M.H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 61–88). San Diego, CA: Academic Press.

Oakes, W. (1972). External validity and the use of real people as subjects. *American Psychologist, 27,* 959–962.

Perone, M., DeWaard, R.J., & Baron, A. (1979). Satisfaction with real and simulated jobs in relation to personality variables and drug use. *Journal of Applied Psychology, 64,* 660–668.

Reips, U.-D. (2000). The web experiment method: Advantages, disadvantages, and solutions. In M.H. Birnbaum (Ed.), *Psychological experiments on the internet* (pp. 89–117). San Diego, CA: Academic Press.

Reips, U.-D., & Bosnjak, M. (Eds.). (2001). *Dimensions of internet science.* Lengerich, Germany: Pabst Science Publishers.

Robins, R.W., Tracy, J.L., & Trzesniewski, K. (2001). Personality correlates of self-esteem. *Journal of Research in Personality, 5,* 463–482.

Rosenthal, R., & Rosnow, R.L. (1975). *The volunteer subject.* New York: Wiley.

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

Saucier, G. (2002). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal of Research in Personality, 36,* 1–31.

Schenke, J., & Rausche, A. (1979). The personality of accident-prone drivers. *Psychologie und Praxis, 23,* 241–247.

Schmidt, W.C. (1997). World-wide web survey research: Benefits, potential problems, and solutions. *Behavior Research Methods, Instruments, & Computers, 29,* 274–279.

Stieger, S., & Voracek, M. (2001, May). *Exploring sexual behavior online: Male-female differences in gender switching and attrition rate.* Poster session presented at German Online Research '01, Göttingen, Germany.

Vollrath, M., Knoch, D., & Cassano, L. (1999). Personality, risky health behavior, and perceived susceptibility to health risks. *European Journal of Personality, 13,* 39–50.

Woolhouse, L., & Myers, S. (1999, September). *Factors affecting sample make-up: Results from an internet-based personality questionnaire.* Paper presented at the 1999 British Psychological Society Social Psychology Section Conference, Lancaster, UK.

Address for correspondence

Tom Buchanan
Department of Psychology
University of Westminster
309 Regent Street
London
W1B 2UW
UK
Tel. +44 207 911-5000 ext 2017
Fax +44 207 9115174
E-mail buchant@wmin.ac.uk