Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Research in Personality 39 (2005) 103–129
JOURNAL OF RESEARCH IN PERSONALITY
www.elsevier.com/locate/jrp

Ascertaining the validity of individual protocols from Web-based personality inventories☆

John A. Johnson

Pennsylvania State University, Penn State DuBois, College Place, DuBois, PA 15801, USA

Available online 5 November 2004

Abstract

The research described in this article estimated the relative incidence of protocols invalidated by linguistic incompetence, inattentiveness, and intentional misrepresentation in Web-based versus paper-and-pencil personality measures. Estimates of protocol invalidity were derived from a sample of 23,994 protocols produced by individuals who completed an online version of the 300-item IPIP representation of the NEO-PI-R (Goldberg, 1999). Approximately 3.8% of the protocols were judged to be products of repeat participants, many of whom apparently resubmitted after changing some of their answers. Among non-duplicate protocols, about 3.5% came from individuals who apparently selected a response option repeatedly without reading the item, compared to .9% in a sample of paper-and-pencil protocols. The missing response rate was 1.2%, which is 2–10 times higher than the rate found in several samples of paper-and-pencil inventories of comparable length. Two measures of response consistency indicated that perhaps 1% of the protocols were invalid due to linguistic incompetence or inattentive responding, but that Web participants were as consistent as individuals responding to a paper-and-pencil inventory. Inconsistency did not affect factorial structure and was found to be related positively to neuroticism and negatively to openness to experience. Intentional misrepresentation was not studied directly, but arguments for a low incidence of misrepresentation are presented. Methods for preventing, detecting, and handling invalid response patterns are discussed. Suggested for future research are studies that assess the moderating effects of linguistic incompetence, inatten-

☆ Prepared for the special issue of the *Journal of Research in Personality 39* (1), February 2005, containing the proceedings of the 2004 meeting of the Association for Research in Personality.

E-mail address: j5j@psu.edu.

doi:10.1016/j.jrp.2004.09.009

tiveness, and intentional misrepresentation on agreement between self-report and acquaintance judgments about personality.

1. Introduction

World Wide Web-based personality measures have become increasingly popular in recent years due to the ease of administering, scoring, and providing feedback over the Internet. Web-based measures allow researchers to collect data, inexpensively, from large numbers of individuals around the world in a manner that is convenient to both researchers and participants. With this emerging technology, two important questions about Web-based measures have been raised. The first is the degree to which established paper-and-pencil personality measures retain their reliability and validity after porting them to the Web (Kraut et al., 2004). Although this question should be answered empirically for each personality measure in question, studies to date suggest that personality measures retain their psychometric properties on the Web (Buchanan, Johnson, & Goldberg, in press; Gosling, Vazire, Srivastava, & John, 2004).

This article addresses a second kind of validity concern for Web-based measures, *protocol validity* (Kurtz & Parrish, 2001). The term *protocol validity* refers to whether an individual protocol is interpretable via the standard algorithms for scoring and assigning meaning. For decades psychologists have realized that even a well-validated personality measure can generate uninterpretable data in individual cases. The introduction of this article first reviews what we know about the impact of three major influences on the protocol validity of paper-and-pencil measures: linguistic incompetence, careless inattentiveness, and deliberate misrepresentation. Next, the introduction discusses why these threats to protocol validity might be more likely to affect Web-based measures than paper-and-pencil measures. The empirical portion of this article provides estimates of the incidence of protocol invalidity for one particular Web-based personality inventory, and compares these estimates to similar data for paper-and-pencil inventories. Finally, the discussion reflects on the significance of protocol invalidity for Web-based measures and suggests strategies for preventing, detecting, and handling invalid protocols.

2. Three major threats to protocol validity

Researchers have identified three major threats to the validity of individual protocols. These threats can affect protocol validity, regardless of the mode of presentation (paper-and-pencil or Web). The first is linguistic incompetence. A research participant who has a limited vocabulary, poor verbal comprehension, an idiosyncratic way of interpreting item meaning, and/or an inability to appreciate the impact of language on an audience will be unable to produce a valid protocol, even for a well-validated test (Johnson, 1997a, 2002). A second threat is carelessness and inattentiveness that leads to random responding, leaving many answers blank, misreading items, answering in the

wrong areas of the answer sheet, and/or using the same response category repeatedly without reading the item (Kurtz & Parrish, 2001). A third threat is any conscious, deliberate attempt to portray one's self uncharacteristically, for example, as better-adjusted or worse-adjusted than the way one is characteristically regarded by others (Paulhus, 1991). The significance of each of these threats to protocol validity is analyzed below.

2.1. Linguistic incompetence

Most adult personality inventories are written between a 5th grade and 8th grade reading level. Obviously, persons reading below that level will be unable to provide valid responses. But basic verbal comprehension is not enough to insure valid responding. To respond validly, a person must also understand the constitutive rules (Johnson, 2004; Wiggins, 1974/1997) that determine how linguistic acts are interpreted (e.g., that agreeing with "I like parties" constitutes evidence of extraversion). Those who understand these rules will provide response patterns leading to scores that correspond to the way others see them (Johnson, 2002; Mills & Hogan, 1978). In contrast, some individuals will construe items idiosyncratically. Too many idiosyncratic interpretations can invalidate a protocol. Validity scales such as the Communality (*Cm*) scale of the California Psychological Inventory (CPI; Gough & Bradley, 1996) will identify some individuals who do not share the communal constitutive rules underlying the scoring of personality measures. Language difficulties will also show up as inconsistency in responding to items that are expected to be answered in either the same or opposite direction (Goldberg & Kilkowski, 1985).

2.2. Carelessness and inattentiveness

The impact of carelessness and inattentiveness on protocol validity needs little explanation. Obviously, frequently skipping items, misreading items, or responding without reading items will invalidate a protocol. Less obvious is the fact that inattentive responding has effects comparable to linguistic incompetence and therefore can be detected with similar techniques. The CPI *Cm* scale identifies not only individuals with poor comprehension but also individuals who respond without attending to item content. The scale accomplishes this because it consists of items that virtually everyone answers the same way, leading to a scale mean that is near the maximum possible score. Linguistically incompetent or inattentive respondents will fail to provide enough common answers to produce an expected high score. Carelessness can also be detected by checking response consistency (Kurtz & Parrish, 2001).

2.3. Misrepresentation

Once again, it seems obvious why misrepresenting one's self on a personality measure would invalidate a protocol, but some issues in misrepresentation seem to be underappreciated. For example, research shows that describing one's behavior with literal accuracy can be insufficient for creating a valid protocol. What matters more than literal descriptive accuracy is whether people respond in such a way that their

scores reflect the way that they are perceived in everyday life. Sometimes telling the truth produces invalid scores and lying produces valid scores (Johnson, 1990). Paper-and-pencil measures have focused on two broad forms of misrepresentation, "faking good" and "faking bad." To "fake good," means to claim to be more competent, well-adjusted, or attractive than one actually appears to be in everyday life. The CPI Good Impression (*Gi*) scale (Gough & Bradley, 1996) was designed to detect this kind of misrepresentation. To "fake bad" means to seem more incompetent or maladjusted than one normally appears to be in everyday life. The CPI Well-Being (*Wb*) scale is an example of such a "fake bad" protocol validity scale (Gough & Bradley, 1996).

The kinds of misrepresentation that may occur on Web-based inventories may transcend the simple "faking good" and "faking bad" that have been assumed to occur on paper-and-pencil inventories. Some research (Turkle, 1995, 1997) indicates that many people who communicate with others on the Internet construct entirely fictional identities that bear little resemblance to the way they are known in everyday life. This is a large step beyond exaggerating positive or negative qualities. This kind of misrepresentation cannot be detected (on either paper-and-pencil or Web-based measures) without comparing protocols to informant ratings or other external criteria. Because the current research collected information only from the test-takers, the study of Internet misrepresentation is left to future research. The current research focuses on the first two threats to validity, linguistic incompetence and carelessness/inattentiveness.

3. Incidence and detection of invalid protocols for paper-and-pencil inventories

Many of the major personality inventories, e.g., the California Psychological Inventory (CPI; Gough & Bradley, 1996), Hogan Personality Inventory (HPI; Hogan & Hogan, 1992), Multidimensional Personality Questionnaire (MPQ, Tellegen, in press), and Minnesota Multiphasic Personality Inventory (MMPI; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), have built-in protocol validity scales to detect cases in which individuals are not attending to or failing to understand item meanings or are presenting themselves in an uncharacteristically positive or negative way. There is a cost to researchers in developing validity scales and a cost to administrators and respondents in the extra time required to complete these scales. That inventories contain validity scales implies that the inventory authors presume that enough respondents will provide invalid protocols to make the inclusion of such scales cost-effective. Validity scales must themselves be validated, and this is normally accomplished by instructing experimental participants (or computers) to simulate a particular kind of invalid responding (e.g., responding randomly; faking good or bad). A successful validity scale correctly identifies most simulated protocols as invalid while misidentifying very few unsimulated protocols as invalid. Experimental evidence shows that validity scales on the major, established personality inventories can in fact distinguish simulated from unsimulated protocols with a degree of accuracy. This has led some to conclude that "personality assessment instruments, to be effective, must have validity scales that can appraise the subject's level of

cooperation, willingness to share personality information, and degree of response exaggeration" (Butcher & Rouse, 1996, p. 94).

In contrast to authors who include validity scales in their inventories, Costa and McCrae (1992) deliberately omitted such scales from their Revised NEO Personality Inventory (NEO-PI-R). Their view is that under naturalistic—as opposed to experimental—conditions, the incidence of invalid protocols is extremely low. Gough and Bradley (1996) themselves report that, under non-evaluative testing conditions, the incidence of faking good on the CPI to be about. 6%, faking bad, .45%, and random responding, .7%. Even under plainly evaluative conditions such as actual personnel selection where people might be motivated to present an inordinately positive view of themselves, inappropriate responding is far more rare than one might expect (Dunnette, McCartney, Carlson, & Kirchner, 1962; Orpen, 1971). Accurately identifying relatively rare events, even with a well-validated scale, is in principle a very difficult psychometric problem (Meehl & Rosen, 1955). Furthermore, research indicates that "correcting" scores with validity scales can actually *decrease* the validity of the measure (Piedmont, McCrae, Riemann, & Angleitner, 2000). Piedmont et al. conclude that researchers are better off improving the quality of personality assessment than trying to identify relatively infrequent invalid protocols.

If Gough and Bradley's (1996) research findings generalize to well-validated paper-and-pencil inventories, administered to literate respondents under non-evaluative conditions, the incidence of protocol invalidity on paper-and-pencil inventories should be less than 2%. The question is whether the incidence of protocol invalidity for Web-based personality measures differs from this figure and how one might detect invalid protocols on the Web.

4. Vulnerability of Web-based personality measures to protocol invalidity

4.1. Linguistic incompetence as a special problem for Web-based measures

Because unregulated Web-based personality measures are readily accessible to non-native speakers from all backgrounds around the world, linguistic competency may be a greater concern for Web-based measures than for paper-and-pencil measures administered to the native-speaking college students often used in research. Non-native speakers may have difficulty with both the literal meanings of items and the more subtle sociolinguistic trait implications of items (Johnson, 1997a). At the time Web-based data were being collected for the research reported here, information on the country of the participant was not being recorded. Later samples with the same Web-based measure show the participation rates to be about 75% from the United States, 8% Canada, 4% United Kingdom and Ireland, 3% Australia and New Zealand, and the remaining 10% from non-English speaking countries. Although the present research cannot attempt to directly compare protocol validity from different cultures, one can assume that linguistic difficulties will create response patterns similar to random responding (see earlier section on carelessness and inattentiveness). Random or incoherent responding can be assessed with measures of infrequency, such as the CPI *Cm* scale, or

measures of response consistency, such as the scales used for the current study. If linguistic difficulties pose special problems for Web-based inventories, consistency scores should be lower than what has been reported for paper-and-pencil inventories.

4.2. Inappropriate levels of attentiveness as a special problem for Web-based measures

Personality measures administered on the Web have two distinctive features that might lead to inappropriate levels of conscious attentiveness (too little or too much) during responding. One such feature of Web-based measures is the psychological distance between the administrator and the participants that arises from the lack of personal, face-to-face contact, especially when respondents can participate anonymously. This distance may give participants a sense of reduced accountability for their actions (although anonymity may sometimes encourage participants to be more open and genuine; see Gosling et al., 2004). A second feature of Web-based measures is the ease of responding, submitting one's protocol, and receiving feedback. Because the process is relatively effortless, participants might rush more carelessly than they would on a paper-and-pencil measure to get their results.

4.3. Inappropriate attentiveness and repeat participation

The ease of responding to Web-based measures increases the probability of a problem that rarely affects paper-and-pencil measures: repeat participation. Because a delay sometimes occurs between the moment a participant clicks a button to submit his or her answers for scoring and the time feedback is returned, impatient and inattentive participants will ignore instructions to click the button only once and consequently submit several copies of the same protocol. Other, more thoughtful, individuals may deliberately choose to complete an unregulated Web-based measure as many times as they like because they are curious about the stability of their results. A variation on this problem arises when participants finish a questionnaire, immediately return to it by clicking the back button on their browser, and then change a few answers to see how the results will be affected. Such playful experimentation results in multiple nearly duplicate protocols with slight mutations.[1] Just as researchers in a

[1] As mentioned earlier, some participants may go beyond mild experimentation and strive to create a wholly artificial profile, answering as they imagine a particular person or type of person (e.g., a celebrity, a fictitious character, a totally well-adjusted or maladjusted person, a self-contradicting individual) might respond. Those who are motivated to do so might participate numerous times, striving to create different personas. In everyday life, repeated encounters with acquaintances sets limits on who you can claim to be without contradicting yourself (Hogan, 1987). Also, when individuals consciously and deliberately attempt to act in uncharacteristic ways in everyday life (e.g., introverts try to act like extraverts), their characteristic patterns "leak through" and are readily detected by observers (Lippa, 1976), although some characteristics are easier to suppress and some individuals are better at making uncharacteristic behaviors seem characteristic (Lippa, 1978). But on the Internet, identity claims cannot always be double-checked, and this has resulted in a proliferation of artificial self-presentations (Turkle, 1995, 1997). One might predict that most people who complete Web-based measures are motivated more often toward increasing self-insight than designing false personas, but only research beyond the work reported here can answer this question.

traditional study would not want a portion of their participants to participate in a study multiple times, researchers using the Web to collect data want to avoid having some participants contributing data more than once.

Protocols that appear consecutively in time or share the same nickname and contain identical responses to every item can be confidently classified as duplicates. More uncertain would be whether two protocols sharing, say, 80% identical responses were generated by the same individual or by two individuals with highly similar personalities. The strategy of the present study was to examine the frequency curve of identical responses between adjacent protocols (sorted by time and nickname) after the almost certain duplicate protocols were eliminated. Setting a cutting point for the number of duplicate responses allowable before protocols are judged to be from the same participant is an arbitrary decision. The lower the cut-point is set, the greater the probability will be that protocols from the same person will be identified, but this also increases the probability of false positives. The hope was that the dispersion of scores would suggest an appropriate cutting point.

The on-line inventory used in the present study presented 60 items on a screen, which allowed participants to return to the previous 60 items by hitting the back button on their browser. On a hunch, I thought it might be useful to examine the number of duplicate responses to only the first 120 items, as well as duplicate responses to all items. This procedure would identify participants who went back several screens, leaving the first 120 responses unchanged, but answering the remaining items differently enough such that the overall number of duplicate responses between protocols did not seem excessive. Therefore, duplicate responses to only the first 120 items, both time-sorted and nickname-sorted, were also computed, and examination of the frequency curve led to judgments about the likelihood of two protocols coming from the same participant.

4.4. Inappropriate attentiveness and careless responding

Hurrying on Web-based inventories, combined with a sense of reduced accountability, increases the probability of response styles associated with too little attention reading items carelessly or not at all, random responding, skipping items, marking answers next to the wrong item, using the response scale in the wrong direction (marking "agree" when "disagree" was intended), and/or using the same response category (e.g., "3" on a 5-point scale) repeatedly to get through the inventory as quickly as possible to see the results. Two of these careless response styles can be measured directly: using the same response category repeatedly and leaving items unanswered. Misreading, misplacing responses, and responding randomly can only be estimated by measuring internal consistency.

The longest string of each response category and the number of missing responses in a protocol are easily calculated. The decision about how long a string must be or how many items can be left blank before a protocol is considered to be invalid is a problem similar to determining whether protocols are from the same participant based on the number of identical responses.

Frequency curves can help identify potential cut points, although it is impossible to know the optimal point for maximizing correct positives and negatives and minimizing false positives and negatives. It would have been useful to have normative data on repeat responses and missing items from a paper-and-pencil version of the Web-based personality measure used in the current research, but such data were not available. However, normative data from paper-and-pencil inventories of comparable length, described below, are available.

4.5. Normative data on using the same response category for consecutive items

In a sample of 983 volunteers whom Costa and McCrae (in press) believed to be cooperative and attentive, no participant used the "strongly disagree" response for more than six consecutive items, "disagree" for more than nine consecutive items, "neutral" for more than 10, "agree" for more than 14, or "strongly agree" for more than nine consecutive items on their 240-item NEO-PI-R. They suggest that NEO-PI-R protocols containing response strings greater than any of these values be viewed as possibly invalid due to inattentive responding. The likelihood of any string of identical responses resulting from valid or inattentive responding will depend on the category endorsement frequencies of the consecutive items, and these endorsement frequencies will vary for different items on different inventories. Nonetheless, the Costa and McCrae data at least provide reference points for consecutive identical Likert responses in a relatively long personality inventory.

4.6. Normative data on missing responses

The author has on file archival protocols from three long paper-and-pencil inventories used in previous studies. These inventories were completed by college students on their own time and then returned to the author. The inventories include the CPI (251 cases of the 462-item version and 237 cases of the 480-item version), HPI (135 cases of the 310-item version and 276 cases of an augmented 380-item version containing unlikely virtues—see Johnson, 1990), and NEO-PI-R (450 cases of the 240-item version). The CPI and HPI employ a True–False rather than 5-point Likert response format, and all three inventories differ in length and item content from the inventory used in the current study. Nonetheless, they can provide reference points for the average number of missing responses in a long paper-and-pencil personality inventory. The average percentage of missing responses in these data sets are, respectively, .49, .42, .39, .11, and .23%. These values, along with the frequency curve for missing responses to the Web-based inventory used in the current research, guided the decision as to whether a protocol had too many missing responses to be included in analyses.

4.7. Measures of internal consistency

The final sets of analyses used participants' internal response consistency to assess inattentiveness. Item response theory models of consistency (e.g., Reise, 1999) were

considered, but judged to be overly stringent and computationally intensive for data screening. A semantic antonym approach (Goldberg & Kilkowski, 1985) in which items judged to be semantic opposites (and therefore should be answered in opposite directions) was also considered, but seemed more appropriate for single-adjective items than for the phrases that comprise items in the current study. Instead, I used two alternative methods, one suggested by Douglas Jackson (1976) and one suggested by Lewis R. Goldberg (personal communication, June 20, 2000).

In Jackson's method, items within each of the standard scales are numbered sequentially in the order in which they appear in the inventory and then divided into odd-numbered and even-numbered subsets. Scores are computed for the half-scale subsets, a product moment correlation is computed between the odd- and even-numbered half-scale scores across all scales, and corrected for decreased length by the Spearman–Brown formula. Jackson refers to this value as an "individual reliability" coefficient. In Goldberg's method, all item responses on the inventory are inter-correlated to identify the 30 unique pairs of items with the highest negative correlations. Such pairs (e.g., in the current study, #31, "Fear for the worst" and #154 "Think that all will be well") are called "psychometric antonyms." Psychometric antonyms are not necessarily semantic antonyms (cf. Goldberg & Kilkowski, 1985; Kurtz & Parrish, 2001; Schinka, Kinder, & Kremer, 1997) and do not necessarily represent forward-scored and reversed-scored items from the same scale. Consistent responders should tend to answer the psychometric antonyms in opposite directions, which means that a correlation across the antonyms within one protocol should be negative. The sign on these correlations was reversed so that a higher number indicated more consistency.

Once again, determining cut points that identified protocols as too inconsistent was based on the frequency curves for the individual reliability coefficients and psychometric antonym scores. Statistics for the individual reliability coefficients were also compared to values reported by Jackson (1977) for his Jackson Vocational Interest Inventory (JVIS). A distribution of actual JVIS individual reliability coefficients from 1706 respondents shows a sharp peak around .80, with virtually all individual reliability coefficients falling between .50 and 1.00. Monte Carlo studies produce an average individual reliability coefficient of zero ($SD = .18$), which is what would be expected from purely random responding. Jackson (1977) suggests that respondents who obtain an individual reliability coefficient of less than .30 "can be categorized as probably primarily attributable to careless, non-purposeful, and/or inarticulated responding" (p. 41).

One concern voiced about consistency as an index of protocol validity is misidentifying inconsistent, but valid, protocols (Kurtz & Parrish, 2001). Costa and McCrae (1997) presented data indicating that convergent validity coefficients (based on correlations between NEO PI-R domain scores and self-ratings on Goldberg's (1992) Big Five adjective markers) from inconsistent respondents are not appreciably different from consistent respondents. Their findings were replicated by Kurtz and Parrish (2001). These researchers conclude that consistency versus inconsistency may be an individual-differences variable, but one that does not impact upon protocol validity. However, the inconsistency scales used in these studies employ semantic judgments

about items that seem opposite in meaning (Schinka et al., 1997) rather than psychometric antonyms that are actually endorsed in opposite directions by most respondents, or half-scales containing items that tend to be answered in the same direction. To see whether consistency as measured by the Jackson and Goldberg indices impacted on factor structure, item-level principle component factor analyses were compared for the upper quartile and lower quartiles on each measure of consistency. To see if the Jackson and Goldberg indices might be regarded as individual differences variables within the normal range of personality, the two consistency indices were entered into a principal components factor analysis with the standard scales of the personality inventory used in the study.

5. Summary of the present research plan

The most direct way of assessing protocol validity would be to compare the results of testing (trait level scores, narrative descriptions) with another source of information about personality in which we have confidence (e.g., averaged ratings or the consensus of descriptions from knowledgeable acquaintances—see Hofstee, 1994). Gathering such non-self-report criteria validly over the Internet while protecting anonymity is logistically complex, and ongoing research toward that end is still in the early stages. In lieu of external criteria, the present study used internal indices to assess protocol validity. The rules of thumb developed to assess protocol validity with these internal criteria should be considered estimates to be tested against external criteria in future research. Likewise, the analyses reporting the incidence of protocol invalidity should be regarded as estimates, pending further study.

The following is a brief overview of the plan and goals for the research. The initial data set contained nearly 24,000 protocols collected via the Web. The plan was to derive cutoff rules for excluding cases in a stepwise fashion. The first goal was to note the incidence of protocols judged to be duplicates or near-duplicates of other protocols. These cases would then be excluded. These duplicate protocols were not necessarily invalid, but were removed because in non-repeated-measures research each participant is expected to participate only once. The next goal was to estimate the number of protocols judged to contain too many consecutive responses with the same response category and compare this estimate to similar data for the paper-and-pencil NEO-PI-R. Then these cases would be removed. The third goal was to record the number of missing responses for each protocol and to compare these figures with the incidence of missing responses on paper-and-pencil versions of the CPI, HPI, and NEO-PI-R. Cases judged to have too many responses would then be removed. The final major goal was to note the incidence of protocols judged to be too inconsistent by Jackson's individual reliability coefficient and Goldberg's psychometric antonyms, and to compare the Jackson data to findings for the paper-and-pencil JVIS. It was predicted that the susceptibility of Web-based inventories to linguistic incompetence and inattentiveness would result in longer strings of the same response category, more missing responses, and less internal consistency.

The two consistency indices, designed to assess invalidity due to linguistic difficulties, carelessness, and inattention, were of particular interest. No previous study had examined the relation between the measures, their relation to structural validity, or their relation to the five-factor model (FFM; John & Srivastava, 1999). Therefore, the research plan also included looking at the correlation between the two measures, differences in factor structure for individuals at the low and high ends of the two consistency scales, and loadings of the two measures within five-factor space.

6. Method

6.1. Participants

Before screening for repeat participation, the sample consisted of 23,994 protocols (8764 male, 15,229 female, 1 unknown) from individuals who completed, anonymously, a Web-based version of the IPIP-NEO (Goldberg, 1999; described below). Reported ages ranged from 10 to 99, with a mean age of 26.2 and *SD* of 10.8 years. Participants were not actively recruited; they discovered the Web site on their own or by word-of-mouth. Protocols used in the present analyses were collected between August 6, 1999 and March 18, 2000.

6.2. Personality measure

To work around various problems associated with commercial personality tests, Goldberg (1999) developed, in collaboration with researchers in The Netherlands and in Germany, a set of 1252 items they dubbed the International Personality Item Pool (IPIP). By administering the IPIP with a variety of commercial personality inventories to an adult community sample, Goldberg's research team has been able to identify, empirically, sets of IPIP items that measure constructs similar to those assessed by commercial inventories. Scales formed from these item sets have demonstrated reliability equal to or greater than the original scales on which they are based and have been found to outperform them in head-to-head predictions of the same real-world criteria (Goldberg, in press). Because the scales are in the public domain on the World-Wide Web at http://ipip.ori.org/, they can be downloaded and ported to the Web without violating copyright restrictions.

I chose from among the various personality inventories at Goldberg's IPIP Web site his 300-item proxy for the revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992), which I call the IPIP-NEO. I chose to work with the IPIP-NEO because the NEO PI-R is one of the most widely used and well-validated commercial inventories in the world (Johnson, 2000a). Furthermore, the NEO PI-R is based on today's most significant paradigm for personality research, the five-factor model (FFM; John & Srivastava, 1999). The average correlation between corresponding scales of the extensively validated NEO PI-R and the IPIP-NEO is .73 (.94 when corrected for attenuation due to unreliability), which suggests promising validity for the IPIP-NEO scales (Goldberg, 1999). A description of the way in which the IPIP-NEO

was formatted for administering, scoring, and providing feedback on the Web can be found in Johnson (2000b).

6.3. Analyses

6.3.1. Multiple participation

Repeat participators were identified by using the LAG command in SPSS Version 10.0, which counts the number of responses in each protocol that are identical to responses in the previous protocol in the data file. This procedure can detect duplicate participation if little or no time elapses between submissions. To identify multiple participation when other participants' protocols enter the data file in the intervening time, protocols were sorted by the participant-supplied nickname and the number of duplicate responses with the previous protocol was recomputed. Frequencies of duplicate responses for both the time-sorted and nickname-sorted data set were computed, and judgments were made about the likelihood of two protocols coming from the same participant. For reasons discussed in the Introduction, duplicate responses to only the first 120 items (sorted by time and then by nickname) were also computed, and frequencies of duplicate responses were examined to determine whether protocols were came from the same participant.

6.3.2. Inattentive responding

SPSS scripts were written to compute the longest string of each of the five response categories (Very Inaccurate, Moderately Inaccurate, Neither Inaccurate nor Accurate, Moderately Accurate, Very Accurate). Frequency curves, the mean, range, and *SD* for these longest strings were computed. These statistics were examined, and potential cutoffs for excluding cases with excessively long strings of a response category were compared to cutoffs suggested for the NEO-PI-R by Costa and McCrae (in press).

6.3.3. Missing responses

The frequencies and means for the number of blank responses were computed and compared to the average percentage of missing responses of the archived paper-and-pencil protocols from the CPI, HPI, and NEO-PI-R. Based on these statistics, a judgment was made about the maximum number of total missing responses that would be allowed before a protocol was judged to be uninterpretable. For cases containing an acceptably low number of missing responses, the midpoint of the scale (3 on the 5-point Likert scale) was substituted for each missing response.

6.3.4. Protocol consistency

Jackson's (1976) individual reliability coefficient and Goldberg's psychometric antonym measure of consistency were computed, and the two consistency measures were correlated to determine the similarity of the kinds of consistency they were assessing. Statistics for the Jackson measure were compared to similar data reported by Jackson (1977). Frequency analyses identified protocols with unacceptably low levels of consistency. Cases that were retained were divided into quartiles on both

measures, and separate, item-level principal components analyses were conducted for the lowest and highest quartiles. The magnitudes of loadings from the high- and low-consistency groups were compared. Finally, scores from the Jackson and Goldberg measures were entered into a principal components factor analysis with the facet scales of the IPIP-NEO to see whether the meaning of consistency might be understood within the FFM.

7. Results

7.1. Duplicate protocols

The SPSS LAG function revealed 747 protocols (sorted first by time and then by nickname) in which all 300 responses were identical to the previous protocol. Also identified were an additional 34 cases in which the first 120 responses were identical. A few additional protocols contained nearly all identical response (e.g., four protocols contained 299 identical responses, one contained 298 identical responses). Protocols with 298, 299, or 300 identical responses to 300 items (or 118, 119, or 120 responses in the first 120 items) are almost certainly coming from the same individual.

The mean number of identical responses in consecutive protocols, out of 300, after these duplicate protocols were removed was 81, with a *SD* of about 20. The area of the frequency curve between 135 and 155 showed a noticeable drop in the number of cases, indicating that a value in this range might make an appropriate cutoff for suspected duplicate protocols. The value chosen, 155, is nearly four standard deviations above the mean. Similar examination of the first 120 items alone led to a cutoff of 66 identical responses in the first 120 items, a value about four standard deviations of the mean of 32. Thus, the total number of protocols judged to be from a prior participant was 918, or 3.8% of the original sample. Removing these protocols reduced the sample to 23,076 participants.

7.2. Long strings of the same response category

The longest strings of the same response category and the number of participants with those longest strings are shown in Table 1. If one applies a scree-like test (Cattell, 1966) of sudden drops in the frequency of the longest response category strings, the following values appear to be potential maxima for the 1–5 points, respectively, on the Likert response scale: 9, 9, 8, 11, and 9. These are similar to Costa and McCrae's (in press) suggested maxima for their NEO-PI-R: 6, 9, 10, 14, and 9. Using the scree-suggested values would reduce the sample by 3.5%, whereas using Costa and McCrae's values would reduce the sample by 6.3%. If Costa and McCrae's maxima were applied to the author's archival sample of 450 paper-and-pencil NEO-PI-R protocols, only four cases (.9%) would be excluded, indicating that inattentive use of the same response category is more likely to happen on the Web than on a paper-and-pencil measure. I opted to use Costa and McCrae's suggested cut points, which

Table 1
Longest consecutive strings of each response category

Longest consecutive string	Response categories					
	0 (Missing)	1 Very inaccurate	2 Moderately inaccurate	3 Neither	4 Moderately accurate	5 Very accurate
1	21158	1801	355	2238	117	1653
2	911	5984	3274	7237	1045	6133
3	271	6819	7194	7138	4888	6982
4	136	4982	6415	3798	7591	4873
5	98	1734	3046	1420	4474	1556
6	65	**740**	1479	598	2462	837
7	17	406	662	279	1291	510
8	20	274	326	122	561	219
9	9	191	**169**	72	323	**132**
10	20	54	68	**41**	131	52
11	4	27	35	25	71	38
12	14	20	19	12	42	21
13	6	9	8	9	26	15
14	4	8	6	5	**13**	10
>14	343	27	20	82	41	45

Note. Values in boldface represent longest string observed by Costa and McCrae (in press). Underlined values represent maxima suggested by a scree-like test.

probably eliminated more non-attentive responders, albeit at the cost of more false positives. This conservative decision reduced the sample to 21,621 participants.

7.3. Missing responses

The average number of missing responses in the sample at this point was 3.6 ($SD = 17.5$), or 1.2% of 300 items. This figure is an order of magnitude larger than the .1–.5% missing responses in the archive of paper-and-pencil CPI, HPI, and NEO inventories. The percentage is inflated by 101 participants who left half or more of the responses blank, but even if those protocols are eliminated, the mean number of missing responses is still 2.6 ($SD = 9.2$), or. 87%. On the positive side, from the sample of 21,621, 33.9% had no missing responses, 60.8% had fewer than two missing responses, and 75.6% had fewer than three missing responses. An examination of the frequency curve showed a sharp decrease in cases after 10 missing responses, so protocols with less than 11 missing responses were retained. This eliminated 2.9% of the protocols, leaving 20,993 cases.

7.4. Internal consistency

Frequency curves for the two consistency indices are shown in Figs. 1 and 2. Whereas the curve for the Jackson coefficient scores is negatively skewed, the

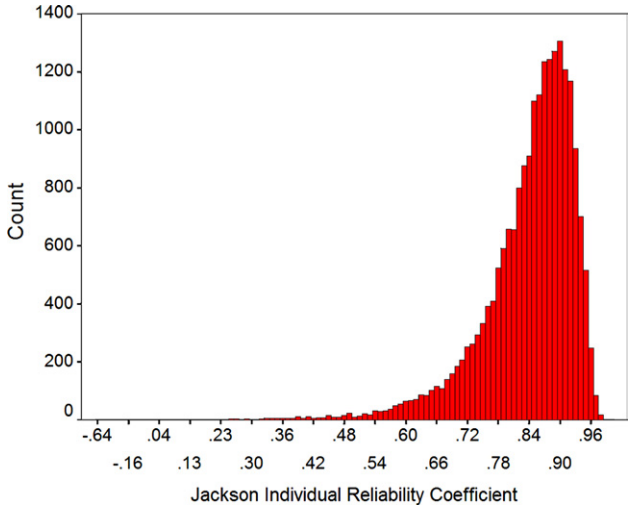

Fig. 1. Frequency curve for the Jackson measure of protocol consistency.

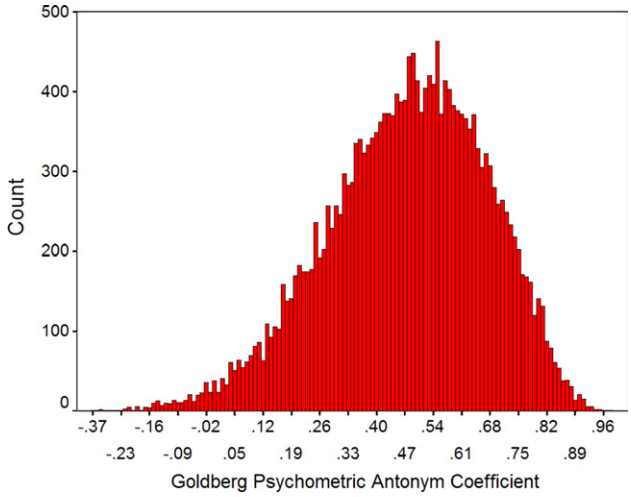

Fig. 2. Frequency curve for the Goldberg measure of protocol consistency.

Goldberg coefficient scores were nearly normally distributed. The skewed distribution for the Jackson coefficient is what one would expect if most participants were responding with appropriate consistency. The near-normal distribution of the Goldberg antonym coefficients resembles the distribution of many personality traits. The Jackson and Goldberg consistency indices correlated moderately ($r = .49$) with each other, although the magnitude of the correlation is probably attenuated by the skew of the Jackson measure. The moderate correlation indicates that the coefficients measure distinct, though related, forms of response consistency.

The mean for Jackson's individual reliability coefficient in the present sample was .84 (range = −.64 to +1.00; $SD = .10$). These results are highly similar to those found by Jackson (1977) for his Jackson Vocational Interest Inventory (JVIS), indicating that current Internet sample is no less consistent than a paper-and-pencil sample. (Even *without* screening out cases with the same response category to many consecutive items or many missing responses, the mean individual reliability coefficient was found to be .83). Jackson's suggested .30 cut point for excluding inconsistent cases was used, which eliminated 41 protocols (.2% of 20,993).

The average for Goldberg's antonym coefficient, with the sign reversed, was .47 (range = −.37 to +.98; $SD = .20$). No comparable data for Goldberg's coefficient have been published, but an analysis of 24,000 pseudo-random cases in SPSS yielded the expected mean coefficient near zero, $M = -.02$ ($SD = .18$). The shape of the psychometric antonyms distribution, coupled with Costa and McCrae's (1997) evidence that scores from low-consistent protocols may be as valid as scores from highly consistent protocols, suggests caution on eliminating too many protocols from the bottom of the distribution. Therefore, only protocols with antonym coefficients less than −.03 were eliminated. With these protocols removed as well as protocols with Jackson individual reliability coefficients less than .30, 20,767 protocols remained (13 protocols were identified as too inconsistent by both indices). The next set of results evaluate whether low-consistent protocols in the remaining sample are indeed less valid than high-consistent protocols.

7.5. Relation between consistency and factor structure

Item responses from the lowest and highest quartiles for both the Jackson and Goldberg measures were subjected to a principal components analysis. When five factors were retained, most items showed their highest loadings on their keyed scales, regardless of whether the sub-sample was the lowest-consistency quartile or the highest-consistency quartile. The few items that did not show their primary loading on their keyed scales did so for both high- and low-consistency sub-samples. The primary loadings in the highest-consistency quartiles averaged about .46, whereas primary loadings in the lowest-consistency quartiles averaged about .35 (see Table 2), but the FFM factor structure was equally discernable, regardless of protocol consistency.

Next, the 30 IPIP-NEO facet subscale scores and the two measures of protocol consistency were submitted to a principal components analysis. Loadings from this analysis are presented in Table 3. The Jackson individual reliability index showed a notable negative loading on the Neuroticism factor and positive loading on the Openness to Experience factor. The Goldberg antonym index of protocol consistency showed a substantial negative loading on the Neuroticism factor and a secondary positive loading on the Openness factor. That stable, open individuals provide more consistent response to personality items than unstable, narrow individuals should not be surprising, given that Openness to Experience has been linked to verbal intelligence (McCrae & Costa, 1985).

Table 2
Average factor loadings for low- and high-consistency protocols

Scale/factor	Measure of protocol consistency							
	Jackson individual reliability				Goldberg psychometric antonyms			
	Low consistency		High consistency		Low consistency		High consistency	
	Keyed scale	Other scales	Keyed scale	Other scales	Keyed scale	Other scales	Keyed scale	Other scales
Extraversion	.32	−.02	.52	−.04	.35	−.02	.51	−.04
Agreeableness	.32	.06	.45	.04	.36	.05	.42	.05
Conscientiousness	.34	.01	.50	.07	.37	.00	.47	.07
Neuroticism	.36	.00	.50	−.07	.37	.00	.48	−.07
Openness	.30	.03	.38	.00	.34	.01	.35	.01
Average	.33	.02	.47	.00	.36	.01	.45	.00

Note. Average factor loadings from a five-factor, varimax-rotated solution to a principal components factor analysis. Loadings under "keyed scale" represent mean factor loadings for the 60 items on the scale defining its respective factor, e.g., average loading of 60 Extraversion-keyed items on the Extraversion factor. Loadings under "other scales" represent mean factor loadings for the remaining 240 items, e.g., average loading of 240 non-Extraversion-keyed items on the Extraversion factor.

7.6. Overlap between exclusion rules

Table 4 presents crosstabulations showing how many cases from the original, full sample would be excluded by two different exclusion rules. Not unexpectedly, between 500 and 900 cases were identified as duplicates by two of the four rules for identifying repeat participants because the rules are not independent measures. Other than the duplicate protocol indices, the other exclusion criteria were remarkably independent. Only a small number of protocols were identified as invalid by any pair of exclusion rules.

8. Discussion

The present study investigated the degree to which the unique characteristics of a Web-based personality inventory produced uninterpretable protocols. It was hypothesized that the ease of accessing a personality inventory on the Web and the reduced accountability from anonymity might lead to a higher incidence (compared to paper-and-pencil inventories) of four types of problematic protocols. These problems are as follows: (a) the submission of duplicate protocols (some of which might be slightly altered), (b) protocols in which respondents use long strings of the same response category without reading the item, (c) protocols with an unacceptable number of missing responses, and (d) randomly or carelessly completed inventories that lack sufficient consistency for proper interpretation. Evidence for a higher incidence was found for the first three problems, but not for protocol inconsistency. Invalid protocols appeared to be easily detectable, and the occurrence of some forms of invalidity may be preventable.

Table 3
Loadings from principle component analysis of facet and protocol consistency scores

	Extraversion	Agreeableness	Conscientiousness	Neuroticism	Openness
Facet					
Friendliness	**.83**	.21	.18	−.16	.01
Gregariousness	**.86**	−.03	.04	−.04	−.04
Assertive	**.56**	−.39	.45	−.19	.20
Activity	.25	−.18	**.67**	−.03	.06
Excitement-seeking	**.62**	−.40	−.18	−.01	.19
Cheerful	**.73**	.16	.03	−.21	.17
Trust	.46	**.51**	−.03	−.24	.03
Morality	−.12	**.78**	.22	−.06	.00
Altruism	.46	**.69**	.18	.02	.22
Cooperativeness	−.04	**.80**	−.06	−.22	−.02
Modesty	−.31	**.52**	−.17	.24	−.21
Sympathy	.17	**.69**	−.05	.18	.35
Self-efficacy	.14	.04	**.63**	−.52	.23
Order	−.14	.18	**.65**	.09	−.29
Dutifulness	−.10	**.61**	.54	−.18	−.04
Achievement	.11	.05	**.80**	−.19	.14
Self-discipline	.04	.14	**.77**	−.27	−.13
Cautious	−.44	.41	**.45**	−.32	−.10
Anxiety	−.28	.02	.03	**.85**	.01
Anger	−.12	−.40	.10	**.72**	−.01
Depression	−.36	−.09	−.26	**.72**	.07
Self-consciousness	−.53	.25	−.26	**.53**	−.11
Immoderation	.19	−.23	−.27	**.52**	.11
Vulnerability	−.16	.07	−.23	**.83**	−.07
Imagination	.10	−.08	−.18	.12	**.69**
Artistic interests	.22	.28	.09	.08	**.64**
Emotionality	.20	.22	.18	.51	**.54**
Adventurousness	.39	−.08	.01	−.32	**.50**
Intellect	−.07	−.09	.18	−.28	**.76**
Liberalism	−.06	−.01	−.30	−.01	**.51**
Protocol consistency					
Jackson	.01	.11	.10	**−.32**	**.36**
Goldberg	.02	−.04	.19	**−.49**	**.21**

Note. Boldface facet loadings indicate the highest factor loading. Boldface protocol consistency loadings suggest location of protocol consistency in five-factor space (Hofstee et al., 1992).

8.1. Detecting and preventing multiple submissions

Identifying duplicate and near-duplicate protocols was readily accomplished by comparing the number of duplicate responses between protocols sorted by time of completion and user-supplied nickname. The total number of protocols judged to be from a prior participant was 3.8% of the original sample, a figure remarkably close to the 3.4% repeat responders to an Internet survey reported by Gosling et al. (2004). At least for the type of inventory and Web format used in the present study, one might expect that about 3.5–4% of the cases will be from the same person. The procedures

Table 4
Numbers of cases from original sample ($N = 23{,}944$) eliminated by two exclusion criteria

	Dup300Time	Dup300Nick	Dup120TIme	Dup120Nick	Consec1	Consec2	Consec3	Consec4	Consec5	Missing	Jackson
Dup300Nick	545										
Dup120TIme	546	544									
Dup120Nick	544	869	546								
Consec1	24	47	25	44							
Consec2	7	10	7	10	9						
Consec3	5	8	5	8	4	3					
Consec4	2	2	2	2	2	0	2				
Consec5	2	4	3	5	47	2	9	2			
Missing	13	31	12	27	38	10	11	2	14		
Jackson	0	5	0	3	10	5	14	2	8	61	
Goldberg	2	11	2	9	10	8	19	0	1	49	39

Note. Abbreviations for exclusion criteria are as follows: Dup300Time, duplicate responses to 300 items, sorted by time of completion. Dup300Nick, duplicate responses to 300 items, sorted by nickname. Dup120Time and Dup120Nick, duplicate responses to first 120 items, sorted by time of completion or nickname. Consec1–Consec5, consecutive use of response categories 1–5. Jackson, Jackson's individual reliability coefficient. Goldberg, Goldberg's psychometric antonym measure.

and statistics presented here can serve as guidelines for screening duplicate protocols in future studies.

More proactive techniques for preventing or identifying duplicate protocols have been suggested by other researchers. One idea is to control access by requiring potential participants to request by e-mail a unique password before participating. If participants must log in with their email address as their user id and their uniquely assigned password, repeat participation can be easily traced (except for individuals who request additional passwords with other e-mail addresses they use). One problem with this restriction is that it destroys anonymity, which will probably lead to less participation. Also, it would require sophisticated accounting software and either an auto-respond program to answer participation requests or significant time from researchers to answer each request by hand. Johnson (2000b) estimated that, on average, someone completed the IPIP-NEO every 5 min.

Fraley (2004) suggests other, less restrictive methods for dealing with multiple-participation. One is to include an item that says something like "I have completed this inventory before," with the "yes" option preselected so that respondents must deselect it if they are completing the inventory for the first time. There may be merit in this idea, although I have found with similar validity checks in the IPIP-NEO that participants sometimes ignore checkboxes that are not part of the actual personality inventory. Fraley also suggests recording the IP address of participants. However, this procedure would misidentify protocols as duplicates when two or more persons share the same computer, which happens often in homes, computer cafés, libraries, and computer classrooms. Furthermore, recording IP addresses compromises anonymity. Fraley also recommends placing a message instructing participants to click the submit button only once and to expect a brief delay, to prevent impatient participants from clicking the submit button multiple times. The IPIP-NEO contains such a warning, but this warning was probably ignored by some participants, given the 747 protocols whose responses were completely identical to another's. Gosling et al. (2004) suggest that some cases of multiple participation can be prevented by providing participants with a link to all forms of feedback, so curious participants can see the full range of possible feedback. The best course of action may be to use both prevention and detection techniques to reduce multiple participation.

8.2. Determining when long strings of the same response category indicate inattentiveness

When a participant uses the same response category (e.g., "Strongly Agree," on a 1–5 Likert scale) for all 60 items on a screen of inventory items, he or she is obviously not attending to the content of the items. But content-inattentive respondents may not use the same response category throughout the entire inventory or even an entire screen. They may alternate between moderately long strings of different response categories. They may become inattentive for only a portion of the inventory, completing most of the inventory appropriately but using strings of the same response category when they are tired (usually at the end of the inventory—see

Morey & Hopwood, 2004). The question is how to determine whether a string of the same response category represents attentive or inattentive responding.

Costa and McCrae (in press) answer this question for their NEO-PI-R by pointing to the longest strings of each response category occurring in a sample of nearly 1000 volunteers that they claim were fully cooperative and attentive. If these participants were fully attentive, the longest string of each response category might be considered the outermost limit for attentive responding. If any of the participants were actually inattentive, some of these suggested outermost limit values would be too high. Costa and McCrae make no strong claims about the values they report, suggesting instead that strings longer than their observed maxima simply be viewed as warnings for potential protocol invalidity.

Although the IPIP-NEO was designed to serve as a proxy for the NEO-PI-R, with scales measuring similar constructs and a similar pattern of alternating between forward- and reversed-scored items from different scales, the suggested maxima for the NEO-PI-R cannot be assumed to apply automatically to the IPIP-NEO. And because the present sample clearly does not consist of fully attentive participants, using Costa and McCrae's technique of identifying the longest string of the same response will not work.

The alternative procedure developed here was to use scree-test-like judgment of the frequency curves of the longest response category strings. Interestingly, this technique identified cut points that were exactly the same as Costa and McCrae's for two of the five Likert categories. Another two category cut points were so close that the number of protocols excluded would not differ much from the number excluded by Costa and McCrae's. Only Costa and McCrae's maximum value for Likert category 1 was noticeably different from the scree-test determined value, excluding over 800 more cases. Although Costa and McCrae's more stringent cutoffs were used in the present study to better insure the elimination of inattentive responders, the decision probably resulted in many false positives. A more accurate estimate of the actual number of inattentive participants who used long strings is provided by the scree rule, which identified 3.5% of the protocols as invalid. By either standard, the rate of this kind of inattentive responding far exceeded what was observed (.9%) in the archival sample of NEO-PI-R protocols, supporting one of the hypotheses of this study—that this type of inattentive responding is more prevalent on Web-based measures than pencil-and-paper measures.

Some participants who used the repeating response pattern may have been more interested in seeing what kind of feedback is generated than in obtaining feedback applicable to them. Giving participants a chance to see what feedback looks like without completing the inventory may prevent such invalid response patterns from occurring (see Gosling et al., 2004).

8.3. Eliminating and preventing protocols with too many missing responses

Book-length treatises have been written on how to handle missing data in research (e.g., Little & Rubin, 1987). Assuming that some missing responses are allowable, one must decide how many missing responses are acceptable, regardless of the method

used for estimating what those missing responses might have been. Even the most sophisticated IRT models (Little & Schenker, 1995) provide only estimates of missing responses, and the greater the number of estimates, the more likely that error will occur. With a large Internet sample, one can afford to discard many cases while still retaining a relatively large number of participants for group-level statistical analyses. Another scree-like test, similar to the one described for examining long strings of the same response category, indicated a sharp decrease in the number of cases with 11 or more missing responses were retained. Eliminating these protocols (2.9% of the sample at this point) left 20,993 cases, and this was certainly sufficient for further group-level statistics. Researchers desiring to salvage protocols by estimating responses to more than 10 missing data points in 300 items are free to do, although if these missing responses occur consecutively they should consider the possibility of inattentive responding (see Table 1). The average number of missing responses in this Internet sample, even after discarding cases in which half or more of the answers were left blank, exceeded the rate of missing responses in the archival sample of paper-and-pencil tests, supporting the hypothesis that skipping items occurs more frequently on the Web.

While some items on long inventories may be left blank intentionally, others are left blank accidentally. As respondents scroll down the screen, they may scroll further than intended and miss items that disappear off the top of the screen. On advantage of Web-based measures over paper-and-pencil measures is that they can be programmed to warn respondents about missing responses. The program can display a message such as "Did you mean to leave item × blank? You are free to do so, but answering all items will improve the validity of your feedback."

8.4. Assessing the consistency of protocols as a sign of protocol validity

Psychologists have long considered a degree of personal consistency or coherence to be a requirement for properly understanding someone's personality (Johnson, 1981). In fact, despite enormous differences in conceptualizing personal consistency—contrast Lecky (1969) with Cervone and Shoda (1999)—many would say that personality *is* some sort of self-consistency (Johnson, 1997b). Randomly inconsistent behavior simply can not be described in the language of personality. For a personality protocol to accurately represent the personal consistencies of everyday life that we know as *personality*, a respondent must respond consistently to personality items in a manner that is monomorphic to (corresponds to the basic structure of) his or her everyday consistencies.

Different standards of consistency on personality measures have been proposed according to different measurement perspectives (Lanning, 1991). Item Response Theory (IRT; Reise, 1999) uses "person-fit statistics" to assess how well an individual's item endorsement pattern fits theoretical expectations based upon item endorsement patterns in a population. In IRT, items are scaled according to how often they are endorsed in the population. For a protocol to have good person-fit, the respondent must endorse mostly items that correspond to their estimated trait level. For example, if an individual with a low estimated trait level endorsed items rarely

endorsed by others in the population, the protocol would have poor person-fit and might be considered invalid. IRT assessments of protocol validity are not only stringent but also mathematically complex and, from a practical point of view, computationally intensive. Therefore, the alternative methods of assessing consistency proposed by Goldberg and Jackson were used in the present study. Future developments may show IRT modeling to be an effective method for detecting invalid protocols (Knotts, 1998).

The shape of the frequency curves for both the Jackson and Goldberg measures, especially the latter, suggested that consistency itself might be regarded as a trait of personality and not simply an index of protocol validity. The frequency curve for the Jackson measure was quite skewed with a mean near the high end of the scale, which helped to justify eliminating some of the extremely low scores. Using Jackson's suggested cutoff of .30 eliminated only .2% of the sample. Overall, the Web-based sample was actually a little more consistent on Jackson's measure than Jackson's participants on his paper-and-pencil inventory, disconfirming the hypothesis that Web measures produce more inconsistency than paper-and-pencil measures.

The Goldberg measure, on the other hand, showed a nearly symmetrical, bell-shaped distribution, making it difficult to decide upon a lower bound for acceptable consistency. Using a cutoff of −.03 on the Goldberg measure eliminated an additional .9% of the sample. A psychometric antonym consistency coefficient near zero may seem like extreme leniency in allowing inconsistency, but the trait-like shape of the frequency curve raised doubts about whether this kind of inconsistency actually implied invalidity. By allowing many inconsistent protocols to remain in the sample, it was possible to test whether inconsistency impacted upon the expected factor structure of the IPIP-NEO.

For both consistency measures, the more consistent respondents did not produce a clearer factor structure than less consistent responders. This finding dovetails with Costa and McCrae's (1997) and Kurtz and Parrish's (2001) conclusions about the lack of impact of consistency on validity. Collectively, the consistency measures' frequency curves, their lack of impact on factor structure, and their loadings on the Neuroticism and Openness to Experience factors supports the notion that these measures of protocol inconsistency reflect more about personality than protocol validity. Ideally we would like to examine the relation between protocol consistency and validity by testing how well inventory scores from less- and more-consistent protocols predict relevant non-self-report data such as acquaintance judgments of personality (Hofstee, 1994). Methods for gathering acquaintance ratings validly on the Web were not available for the current study. When they do become available, studies of the moderating effects of all of the internal indices on self-acquaintance agreement study can be undertaken.

8.5. What about misrepresentation?

The current study did not attempt to assess the incidence of any kind of misrepresentation. The IPIP-NEO has neither "fake good" nor "fake bad" scales, and no external criteria were collected to verify, externally, how accurately a protocol

represented someone's personality. Even if acquaintance ratings are gathered in a future study, participants' knowledge that acquaintances will be providing ratings may well produce a different frequency of misrepresentation than what would be found in a completely anonymous Web-based measure. One possible way to study people's desire to construct uncharacteristic identities would be to give them options of completing the test "as themselves" or as a simulator and asking them to indicate which approach they are using.

Despite reports of individuals constructing identities on the Internet that differ dramatically from the way they are seen by knowledgeable acquaintances, motivational considerations argue against widespread misrepresentation on most Web-based personality inventories. In the words of Fraley (2004, p. 285), "People are fickle when they are surfing the Internet…. It is unlikely that people would waste their time in your experiment just to give you bad or silly data. Most people will participate in your study because they are hoping to learn something about themselves." If these motivational assumptions are correct, most respondents to Web-based inventories will "be themselves," which is to say they respond to the items with the same social-linguistic habits they use in normal conversations, generating the same personality impressions they typically make in everyday life (Johnson, 2002).

9. Conclusions

Of more substance and practical importance than the specter of radical misrepresentation on Web-based personality measures are issues such as detecting multiple participation and protocols that are completed too carelessly or inattentively to be subjected to normal interpretation. The incidence of: (a) repeat participation, (b) selecting the same response category repeatedly without reading the item, and (c) skipping items all exceed the levels found in paper-and-pencil measures. Nonetheless, preventing and detecting these threats to protocol validity can be accomplished with the methods presented in this article.

Other protocols may be uninterpretable because the respondent answers many items randomly, or is not linguistically competent enough to understand items, or purposely responds to items in contradictory ways to see what happens. Given the motivational considerations discussed above, intentional inconsistency would be expected to be rare, and data from the present study indicate that Web respondents are no less consistent than respondents to paper-and-pencil measures. Some inconsistency due to language problems can be mitigated by using items that are simple and comprehensible (Wolfe, 1993) and judged to clearly imply the trait being measured (Hendriks, 1997).

In conclusion, although the rates of certain kinds of inappropriate responding may invalidate a slightly higher percentage of unregulated, Web-based personality measures than paper-and-pencil measures, steps can be taken to reduce inappropriate responding, and invalid protocols can be detected. The much larger and potentially more diverse samples that can be gathered via the World Wide Web (Gosling et al., 2004) more than make up for the slightly higher incidence of invalid protocols.

Future research assessing protocol validity by comparing self-report results to judgments of knowledge acquaintances may further improve our ability to detect and eliminate invalid protocols. As we gain confidence in our methods for detecting potentially invalid protocols, we can program these detection rules directly into Web-based measures to automatically flag suspect protocols.

Acknowledgments

Some of these findings were first presented in an invited talk to the Annual Joint Bielefeld-Groningen Personality Research Group meeting, University of Groningen, The Netherlands, May 9, 2001. I thank Alois Angleitner, Wim Hofstee, Karen van Oudenhoven-van der Zee, Frank Spinath, and Heike Wolf for their feedback and suggestions at that meeting. Some of the research described in this article was conducted while I was on sabbatical at the Oregon Research Institute, supported by a Research Development Grant from the Commonwealth College of the Pennsylvania State University. I thank Lewis R. Goldberg for inviting me to the Oregon Research Institute and for his suggestions for assessing protocol validity. Travel to Austin, Texas, where this research was presented to the Association for Research in Personality was partially supported by the DuBois Educational Foundation. I thank Sam Gosling and Oliver John for their helpful comments on an earlier version of the manuscript.

References

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (in press). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment*.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory2 (MMPI-2): Manual for administration and scoring*. Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., & Rouse, S. V. (1996). Personality: Individual differences and clinical assessment. *Annual Review of Psychology, 47*, 87–111.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276.

Cervone, D., & Shoda, Y. (Eds.). (1999). *The coherence of personality: Social-cognitive bases of consistency, variability, and organization*. New York: Guilford Press.

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.

Costa, P. T., Jr., & McCrae, R. R. (1997). Stability and change in personality assessment: The revised NEO personality inventory in the year 2000. *Journal of Personality Assessment, 68*, 86–94.

Costa, P. T., Jr., & McCrae, R. R. (in press). The revised NEO Personality Inventory (NEO-PI-R). In: S. R. Briggs, J. M. Cheek, & E. M. Donahue (Eds.). *Handbook of adult personality inventories*. New York: Kluwer.

Dunnette, M. D., McCartney, J., Carlson, H. C., & Kirchner, W. K. (1962). A study of faking behavior on a forced choice self-description checklist. *Personnel Psychology, 15*, 13–24.

Fraley, R. C. (2004). *How to conduct behavioral research over the Internet*. New York: Guilford Press.

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

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. (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: Kluwer.

Goldberg, L. R., & Kilkowski, J. M. (1985). The prediction of semantic consistency in self-descriptions: Characteristics of persons and of terms that affect the consistency of responses to synonym and antonym pairs. *Journal of Personality and Social Psychology, 48*, 82–98.

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.

Gough, H. G., & Bradley, P. (1996). *CPI manual: Third edition*. Palo Alto, CA: Consulting Psychologists Press.

Hendriks, A. A. J. (1997). *The construction of the Five Factor Personality Inventory*. Unpublished doctoral dissertation, University of Groningen, The Netherlands.

Hofstee, W. K. B. (1994). Who should own the definition of personality? *European Journal of Personality, 8*, 149–162.

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

Hogan, R. (1987). Personality psychology: Back to basics. In J. Aronoff, A. I. Rabin, & R. A. Zucker (Eds.), *The emergence of personality* (pp. 79–104). New York: Springer Publishing Company.

Hogan, R., & Hogan, J. (1992). *Hogan Personality Inventory manual*. Tulsa, OK: Hogan Assessment Systems.

Jackson, D. N. (1976). *The appraisal of personal reliability*. Paper presented at the meetings of the Society of Multivariate Experimental Psychology, University Park, PA.

Jackson, D. N. (1977). *Jackson Vocational Interest Survey manual*. Port Huron, MI: Research Psychologists Press.

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. (1981). The "self-disclosure" and "self-presentation" views of item response dynamics and personality scale validity. *Journal of Personality and Social Psychology, 40*, 761–769.

Johnson, J. A. (1990). *Unlikely virtues provide multivariate substantive information about personality*. Paper presented at the 2nd Annual Meeting of the American Psychological Society, Dallas, TX.

Johnson, J. A. (1997a). Seven social performance scales for the California Psychological Inventory. *Human Performance, 10*, 1–30.

Johnson, J. A. (1997b). Units of analysis for description and explanation in psychology. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 73–93). San Diego, CA: Academic Press.

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

Johnson, J. A. (2000b). *Web-based personality assessment*. Paper presented at the 71st Annual Meeting of the Eastern Psychological Association, Baltimore, MD.

Johnson, J. A. (2002). Effect of construal communality on the congruence between self-report and personality impressions. In P. Borkenau & F.M. Spinath (Chairs) (Eds.). *Personality judgments: Theoretical and applied issues*. Invited symposium for the 11th European Conference on Personality, Jena, Germany.

Johnson, J. A. (2004). The impact of item characteristics on item and scale validity. *Multivariate Behavioral Research, 39*, 271–300.

Knotts, L. S. (1998). Item response theory and person-fit analyses of the Revised NEO Personality Inventory conscientiousness domain. *Dissertation Abstracts International, 59*(6), 3063B.

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

Kurtz, J. E., & Parrish, C. L. (2001). Semantic response consistency and protocol validity in structured personality assessment: The case of the NEO–PI–R. *Journal of Personality Assessment, 76*, 315–332.

Lanning, K. (1991). *Consistency, scalability, and personality measurement*. New York: Springer.

Lecky, P. (1969). *Self-consistency: A theory of personality*. Garden City, NY: Doubleday Anchor.

Lippa, R. (1976). Expressive control and the leakage of dispositional introversion–extraversion during role-played teaching. *Journal of Personality, 44*, 541–559.

Lippa, R. (1978). Expressive control, expressive consistency, and the correspondence between expressive behavior and personality. *Journal of Personality, 46*, 438–461.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Little, R. J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 39–75). New York: Plenum Press.

McCrae, R. R., & Costa, P. T., Jr. (1985). Openness to experience. In R. Hogan & W. H. Jones (Eds.), *Perspectives in personality* (Vol. 1, pp. 145–172). Greenwich, CT: JAI Press.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychological Bulletin, 52*, 194–216.

Mills, C., & Hogan, R. (1978). A role theoretical interpretation of personality scale item responses. *Journal of Personality, 46*, 778–785.

Morey, L. C., & Hopwood, C. J. (2004). Efficiency of a strategy for detecting back random responding on the Personality Assessment Inventory. *Psychological Assessment, 16*, 197–200.

Orpen, C. (1971). The fakability of the Edwards Personal Preference Schedule in personnel selection. *Personnel Psychology, 24*, 1–4.

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). New York: Academic Press.

Piedmont, R. L., McCrae, R. R., Riemann, R., & Angleitner, A. (2000). On the invalidity of validity scales: Evidence from self-reports and observer ratings in volunteer samples. *Journal of Personality and Social Psychology, 78*, 582–593.

Reise, S. P. (1999). Personality measurement issues views through the eyes of IRT. In S. E. Embretson & S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 219–241). Mahwah, NJ: Erlbaum.

Schinka, J. A., Kinder, B. N., & Kremer, T. (1997). Research validity scales for the NEO–PI–R: Development and initial validation. *Journal of Personality Assessment, 68*, 127–138.

Tellegen, A. (in press). *Manual for the Multidimensional Personality Questionnaire*. Minneapolis: University of Minnesota Press.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. New York: Simon and Schuster.

Turkle, S. (1997). Constructions and reconstructions of self in virtual reality: Playing in the MUDs. In S. Kiesler (Ed.), *Culture of the Internet* (pp. 143–155). Hilldale, NJ: Lawrence Erlbaum Associates.

Wiggins, J. S. (1997). In defense of traits. In R. Hogan, J. A. Johnson, & S. R. Briggs (Eds.), *Handbook of personality psychology* (pp. 95–115). San Diego, CA: Academic Press (Originally presented as an invited address to the Ninth Annual Symposium on Recent Developments in the Use of the MMPI, held in Los Angeles on February 28, 1974.).

Wolfe, R. N. (1993). A commonsense approach to personality measurement. In K. H. Craik, R. Hogan, & R. N. Wolfe (Eds.), *Fifty years of personality psychology* (pp. 269–290). New York: Plenum.