# Screening Massively Large Data Sets For Non-Responsiveness In Web-Based Personality Inventories

Invited talk to the joint Bielefeld-Groningen Personality Research Group
University of Groningen, The Netherlands
May 9, 2001
John A. Johnson, Ph.D.

[This is an extended version of the talk I gave from informal notes.]

## Purpose of the Talk

What I would like to do is first present a rationale for placing a personality inventory on the World Wide Web. Next I will present a brief history of my efforts at developing a Web-based inventory. Finally, I will describe several methods I developed for screening the data to weed out protocols with inappropriate response patterns. This is a significant problem for data collected on the web because of the very large number of respondents and the potential for different types of inappropriate responding. This paper ends with a summary of the ensuing discussion about this research.

## Background and Rationale for Web-Based Testing

In a review of personality assessment via computers (Johnson, 1994a), I outlined the advantages of using a computer to collect, score, and interpret responses to personality items. Before the advent of computer administration and scoring, psychologists typically tallied pencil marks on answer sheets and transferred these tallies to profile sheets for visual inspection and interpretation. Blurry marks and incomplete or multiple responses had to be dealt with on an ad hoc basis. Computer administration increases the speed and accuracy of personality data collection. Blurry marks do not exist, multiple responses are disallowed, and incomplete responses can be disallowed or assigned a missing value to be handled by the scoring program. Counting, summing, and norming errors are eliminated.

The clerical advantages of computer administration and scoring represented a liberating breakthrough for research psychologists like myself who work at a small campus with limited human resources. Nonetheless, even at the time I wrote the paper on using computers for data collection (Johnson, 1994a), I was still keyboarding personality item responses into computer data files by hand because our campus did not have sufficient computer technology for automated data input. My limited resources also meant limited samples—in terms of both number and heterogeneity. However, by the latter 1990s, the advent of the World-Wide Web provided a way to overcome these limitations. Personality item responses on the Web can be tallied and scored automatically by a CGI program, saving the researcher enormous amounts of time. Also, if a narrative-report writing program is included, the respondent can receive instantaneous feedback for completing the on-line inventory. Furthermore, with a Web-based personality inventory the researcher can collect data literally from all over the world rather than from persons who are in physical proximity (Smith & Leigh, 1997).

Despite the liberating potential of the Web, various concerns must be addressed before this potential can be realized. I have described these concerns in a paper on problems associated with using commercial personality inventories for research (Johnson, 1993). One concern is the secrecy surrounding the scoring keys for some of these inventories. This secrecy disallows researchers from conducting even the most fundamental psychometric analyses such as reliability estimation and factor analysis. Furthermore, secrecy surrounding the computer programs used to build narrative reports from commercial personality inventories precludes critical discussion of the merits and drawbacks of different techniques for generating narrative reports. The expense of commercial personality inventories places additional burdens on researchers with limited resources, and copyright restrictions make it virtually impossible to create Web-based versions of these inventories. All of these issues (and additional problems with commercial inventories noted by Goldberg, in press) impede technical progress in personality assessment.

To work around the problems associated with commercial personality tests, Goldberg (1999) has developed, in collaboration with researchers from the Rijksuniversiteit Groningen (The Netherlands) and Universität Bielfeld (Germany) a set of 1,252 items dubbed the International Personality Item Pool (IPIP).

By administering the IPIP with a variety of commercial personality inventories to an adult community sample (in stages to prevent fatigue), Goldberg's research team has been able to identify, empirically, sets of IPIP items that measure the same constructs as the commercial inventories. Scales formed from these item sets possess psychometric properties that match or exceed those of the original commercial scales (Goldberg, in press). These scales are in the public domain on the World-Wide Web at http://ipip.ori.org/ipip/ . My contacts with the Bielefeld research group during my 1990-91 sabbatical year (Johnson & Ostendorf, 1993) and the Groningen research group (Johnson, 1991, 1994b) eventually led me to Goldberg's IPIP research program at the Oregon Research Institute.

## History of the Web Version of the IPIP-NEO

### Pilot Testing Computer Algorithms

During my sabbatical year in Bielefeld I had created computer algorithms in FORTRAN for producing five-factor model (FFM; Wiggins, 1996) based narrative reports (Johnson, 1991, 1993, 1994a). Between 1995 and 1996 I taught myself the hyper-text markup language (HTML) and practical extraction and report language (Perl) needed to transport these algorithms to the World-Wide Web. At this point in time I was unaware of Goldberg’s IPIP project, so I had to choose an existing inventory for Web testing. I own hundreds of test booklets for three major personality inventories: the California Psychological Inventory (CPI; Gough, 1987), Hogan Personality Inventory (HPI; Hogan & Hogan, 1992), and the revised NEO Personality Inventory (NEO PI-R; Costa & McCrae, 1992). Although I have used all three inventories successfully in research, I chose the NEO PI-R because it seems to be the most robust predictor of FFM-based acquaintance ratings (Johnson, 2000a). I coded the NEO PI-R into HTML and pilot tested the Web page with my introductory psychology students.

The HTML method for presenting items and scoring responses employed the relatively simple TABLE and FORM formats. After some introductory instructions, NEO items were presented in a column on the left side of a table. Five small circles called *radio buttons* appeared next to each item in the right-hand column; these radio buttons were labeled *strongly disagree, disagree, neutral, agree*, and *strongly agree*. Numerical values of 1, 2, 3, 4, and 5 were associated with these respective responses (or 5, 4, 3, 2, and 1 for items scored in reverse direction) by the HTML underlying the Web page. Respondents could not see these numbers, although persons familiar with HTML and features of their Web browsers could select "view source" to see the software code underlying the Web page. Respondents were constrained to clicking only one of the five radio boxes for each item. When the respondent reached the end of the Web page, clicking the "submit button" sent the numerical responses to a common gateway interface (CGI) script written in Perl. This CGI script (a) summed the responses into scale scores, (b) transformed the scores into T-scores with a mean of 50 and standard deviation of 10, (c) created a Web page containing a narrative report explaining the meanings of the scores for the respondent, and (d) automatically sent me an email containing these scores.

Other papers (Johnson, 1991, 1993, 1994a) describe my program's basic algorithms for producing narrative reports from personality scores, although these earlier programs created text files for printing rather than Web pages. Essentially, the program produces a Web page with a detailed description of each personality dimension from the FFM (Neuroticism, Extraversion, Openness to Experience, Agreeableness, and Conscientiousness). Included within these descriptions is what research has revealed about persons with relatively high or low scores on each of the five factors. At the end of each detailed description of the major factors a paragraph with the following form appears: "Your score on [name of factor] is [high, average, low], indicating that [brief summary of what research has revealed about persons with the score]."

Underneath the detailed descriptions of the five major factors are shorter descriptions of the six facets of dimension. For example, short descriptions of Anxiety, Anger, Depression, Self-Consciousness, Immoderation, and Vulnerability appear under the description of Neuroticism. After each facet description is a sentence of the form: "Your level of [name of facet] is [high, average, low]." High scores were defined by T-scores greater than 55 (i.e., greater than .5 standard deviation above the mean) and low scores by T-scores less than 45, according to the norms in the NEO PI-R manual (Costa & McCrae, 1992).

The initial pilot test indicated that presenting all 240 items on one web page overtaxed many computers with relatively slow processors, limited memory, or slow Internet connections. Respondents reported that their computers "froze up" after they completed only part of the inventory. In 1997 I corrected this problem by revising my Web site to present only 60 items on the first Web page and then store the HTML for the remaining items within a series of CGI script programs. Clicking the submit button at the end of 60 items temporarily stored the item responses and triggered the next CGI script to generate, on the fly, a Web page presenting the next 60 items. Clicking the submit button after the final item sent all responses to the final CGI script. Although some respondents occasionally reported technical problems with the Web site, this on-line presentation of items appeared to run smoothly 99% of the time.

Throughout this period of pilot testing, my goal was simply to see if it were possible to administer and score a long personality inventory on the Web. I realized that I would eventually have to deal with copyright issues if I were to use this on-line version of the NEO PI-R for serious research. My plan had been to get the on-line inventory working before contacting Costa and McCrae for permission to conduct research with it. However, at this point in time, I became aware of Goldberg's 300-item proxy for the NEO PI-R. I therefore substituted Goldberg's items for the items in the original NEO PI-R and changed the response options from *strongly disagree, disagree, neutral, agree,* and *strongly agree* to *very inaccurate, moderately inaccurate, neither accurate nor inaccurate, moderately accurate,* and *very accurate.* I dubbed the new on-line inventory the IPIP-NEO.

## Further Developments with the IPIP-NEO

Preliminary norms for generating the T-scores for the IPIP-NEO were based on means of the adults in Goldberg's (1999) community sample, adjusted for the slight differences between this sample's scores on the NEO PI-R and national norms for the NEO-PI-R reported in the professional manual (Costa & McCrae, 1992). Separate norms were used for each sex, graded by age. Norms for persons under 21 were estimated by adjusting the adult norms according to differences between the two age groups reported by Costa & McCrae (1992) for the NEO PI-R. These estimated norms were considered preliminary until sufficient data were collected to establish genuine norms from Internet participation. On September 1, 1998, I uploaded the IPIP-NEO to the World-Wide Web at <http://cac.psu.edu/~j5j/test/ipipneo1.htm>.

I continued to work on my CGI PERL scripts throughout my 1999 sabbatical leave. I added to the introductory web page a section where respondents could indicate if they were completing the IPIP-NEO for course extra credit or if they were visiting from outside the university. I included a special checkbox for visitors who linked to the inventory from a set of Web pages on personality that I helped construct for the Annenberg/Corporation for Public Broadcasting Project (http://www.learner.org/exhibits/personality/ ). I added "traps" in the first CGI script to send respondents who failed to indicate their age, sex, or visitor status back to the first page to complete that information.

During the month of August 1999, I analyzed all scores that had been collected between September 1, 1998 and August 6, 1999. Information in the headers of emails containing the IPIP-NEO scores was used to eliminate duplicate emails from individuals who apparently clicked the final submit button more than once. Scores were also screened to insure that they fell within one point of the minimum score for each scale. This procedure retained scores from individuals who skipped a few items on the IPIP-NEO, but eliminated scores from individuals who left many answers blank. (Blank items are scored 0 by default, whereas the lowest score of a completed item is 1.) After eliminating duplicate and incomplete protocols, 4,472 valid protocols remained. Of these protocols, 639 were identified as links from the Annenberg/CPB site and the remaining from unspecified other sites. Data from Penn State students were not included in these analyses. The sample included 2,026 males and 2,446 females. Reported ages ranged from 11 through 90; the average age for males was 34.1 years (SD=12.3) and for females, 31.8 years (SD=12.0).

I prepared a descriptive summary of these data for the 2000 meeting of the Eastern Psychological Association (Johnson, 2000b). In this summary I noted that my Internet sample scored higher in Neuroticism and Openness to Experience and lower in Agreeableness and Conscientiousness than Goldberg's Eugene-Springfield community sample. This finding is perhaps in accord with a previous suggestion (Kraut, et al., 1998) that Internet use may be linked to depressive tendencies. A factor analysis of the IPIP-NEO's facets produced the anticipated five-factor structure, with 27 of the 30 facet scales showing a primary loading on the expected factor.

By August 6, 1999, I had taught myself enough additional PERL to have the last CGI script save item responses directly to a computer file rather than email the scores. On August 18-22 I traveled to the Oregon Research Institute to share my findings with Lewis Goldberg and discuss plans for future research. Much of these discussions concerned the development of improved programming and automated procedures for screening data. It appears that in some cases, individuals are not getting feedback instantly (as they should) after clicking the final submit button, which leads them to click and send their data several times. Although duplicate entries are easily detected after the fact, much time could be saved if the mechanism for storing data could prevent multiple submissions. Other concerns included persons who are leaving too many items blank or using the same response option too often (e.g., mostly answering with 3s). Goldberg also suggested incorporating intra-person reliability checks to examine the internal consistency of responding.

Between the time I began collecting item responses (August 1999) and the present time (May 2001), roughly 175,000 IPIP-NEO protocols were collected. By estimates described below, about 90% of these were judged to be valid, non-duplicate protocols. Most of these data remain unanalyzed. However, the first set of 23,994 protocols was subjected to an intensive data screening to detect duplicate protocols and to attempt to distinguish protocols marked by valid, responsive answering from protocols produced by various forms of inappropriate responding.

## Screening Rules for Non-Responsiveness

### Data Set

The data set used in these analyses comes from 23,994 IPIP-NEO protocols collected between August 6, 1999 and March 18, 2000. Protocols were subjected to a seven-stage screening for non-respondent answering or duplication of previous protocols. This screening procedure is described below. After screening, 21,588 non-duplicate, individual protocols (7,859 male, 1,3729 female) were judged to be valid. The average age of the final sample was 26.2 (SD=10.8). The average age for males was 26.3 (SD=10.6) and for females, 26.2 (SD=10.9). The median age for both males and females was 23.

### Seven-Stage Data Screening

I designed the first stage in the screening procedure to deal with respondents who submitted their completed inventory more than once by clicking the back button on their browser after receiving their results and then clicking the submit button again (sometimes after changing some of their responses). A count of the number of responses in each protocol that duplicated responses in the previously completed protocol revealed 526 protocols in which all 300 responses were identical to the previous protocol. Only a few additional protocols contained large numbers of duplicate response (e.g., four protocols contained 299 duplicate response, one contained 298 duplicate responses). I judged a cutoff of 155 or more duplicate responses to be an adequate cutoff point for eliminating protocols. (The 155 cutoff is nearly two SDs above the mean for duplicate responses, 85.6). The reduced N, 23,443, represents 97.7% of the original sample size.

The second stage in the screening procedure considered the number of unanswered items in each protocol. The average number of blank responses was 3.7 (SD=17.7). The mean for the longest string of consecutive missing responses was 1.9 (SD=16.6). I decided to eliminate protocols with more than five consecutive missing responses (n=452), leaving 22,991 protocols. With this reduced data set, the average number of duplicate responses fell to 81.2, and the cutoff for duplicate responses described above, 155, now represented four standard deviations above this mean.

The third stage in the screening process searched for duplicate protocols that did not appear consecutively in time. These duplications would appear if someone finished the test and began to retake it, other individuals completed the test, and then the first person resubmitted the test, possibly with some answers changed. To detect this type of duplication, I sorted all protocols by nickname and recomputed the number of duplicate responses with the previous protocol. Once again using the cutoff of 155 duplicate responses, I found 341 suspected duplicates that were eliminated, leaving 22,650.

The fourth screening stage employed a different criterion for identifying potential duplicate protocols. Instead of looking at the total number of duplicate responses with the previous protocol (sorted by nickname), I looked only at the number of duplicate responses within the first 120 items. This screening identified one protocol with 120 duplicate responses that had not been detected by the <155 criterion. I eliminated this protocol as well as three protocols with 66 duplicate responses and one with 76 duplicate responses, leaving 22,645 cases.

The fifth stage of screening eliminated protocols with excessive missing answers. The average number of missing response in the sample at this point was 1.8 (SD=2.5). Protocols with less than 11 missing responses constituted 98.7% of the sample, so <11 missing responses was used as a cutoff. This left 22,361 cases.

The sixth stage of screening focused on apparent non-responsive answering, defined as long strings of answers with the same response category, e.g., two-dozen "Very Accurate" responses in a row. I generated frequency curves for the longest strings of each response category and computed the mean, range, and SD for these longest strings. The means for each response category were between .97 and 3.68, with SDs between 1.43 and 4.32. The longest consecutive strings ranged from 58 consecutive "Somewhat Accurate" to 257 "Neither Accurate nor Inaccurate" answers. These results and the frequency curves led me to reject protocols containing strings of "Very Inaccurate," >14; "Moderately Inaccurate," >15; "Neither Accurate nor Inaccurate," >59; "Somewhat Accurate," >19; and "Very Accurate," >27. These cutoffs retained at least 99% of protocols in each case, resulting in a sample size of 22,295.

The seventh stage of screening demanded a certain level of internal consistency in responding. I considered item response theory models (e.g., Reise, 1999) for this stage, but judged them to be overly restrictive for data screening. 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 30 facet 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 30 scales, and corrected for decreased length by the Spearman-Brown formula. Jackson refers to this value as an "individual reliability" coefficient. The mean individual reliability coefficient in the present sample was .84 (SD=.10). A fairly lenient minimum cutoff of .40 was chosen, which disqualified only .7% of the sample.

In Goldberg's method, all items on the inventory are inter-correlated to identify the 30 unique pairs of items with the highest negative correlations. Goldberg refers to such pairs (e.g., #31, "Fear for the worst" and #154 "Think that all will be well") as “semantic antonyms.” [At the reading of this paper, Wim Hofstee correctly noted that these pairs are actually *psychometric* antonyms rather than *semantic* antonyms.] Consistent responders should tend to answer the semantic antonyms in opposite directions, such that a correlation across the antonyms within one protocol should be negative. The average antonym correlation was -.47 (SD=.20). The sign on these correlations was reversed so that a higher number indicated more consistency. A minimum correlation cutoff of .07 was chosen, which disqualified 3.2% of the sample. The Jackson and Goldberg consistency indices correlated .49 with each other, indicating that they were identifying many of the same inconsistent responders. After protocols that failed to pass either Jackson's or Goldberg's consistency tests were eliminated, the final sample size was 21,588. This reduced sample was used for re-norming and development of an IPIP-NEO short form, described below.

## Other Developments, Future Plans, and Suggestions from Audience

### Other Developments

Shortly after screening the sample of 21,588 respondents described above, I calculated scale means and standard deviations for males and females, for ages 21 years or older and for under 21 years of age. I entered these new normative data as reference points for determining high, average, or low scores in the scoring routine.

On August 8, 2000, I added to the IPIP-NEO website a drop-down list of countries so that each respondent could indicate the country to which he or she felt to " belong the most, whether by virtue of citizenship, length of residence, or acculturation." I also added a CGI that generates a quasi-random, 23-character long nickname if the respondent does not provide his or her own nickname. This feature was added to help identify duplicate entries for persons who did not choose their own nicknames.

As far as I could tell, the CGI scripts for the IPIP-NEO have continued to execute successfully over 99% of the time. Nonetheless, given the large number of people who attempt to complete the IPIP-NEO every day, a significant number of respondents continued to experience problems completing the IPIP-NEO. During the last few months of the year 2000 I sent my CGI code to several persons to try to determine if the code contained errors that might cause the program to crash. I received one suggestion (which I implemented) to have the CGI catch individuals who try to spell out their age instead of typing a number. I also tried exporting my scripts to another site I manage and ran them under Perl 5 instead of Perl 4. A test run on the alternate site revealed a possibly problematic coding for the way I listed my email address, so I changed that. No other errors could be found, so why the scripts occasionally fail remains a mystery.

In November 2000, I was contracted by a consulting firm to develop a shorter version of the IPIP-NEO that they could use in personnel selection. Their hope was that I could reduce the length of the inventory by at least 50% while maintaining alpha reliability estimates of at least .70 for every scale. The following is a section from my report to that company that describes my three-phase strategy for developing the shorter scales.

The first phase was a pure empirical-statistical procedure on the full sample that eliminated items with the lowest corrected item-total correlations, without considering item content. The Reliability application from SPSS Base 10.0 (SPSS, 1999) was used to generate corrected item-total correlations for each of the 10 items in each facet scale. The item with the lowest item-total correlation was identified, eliminated, and the Reliability application run again until four items remained. Before any items were eliminated, coefficient alphas for the 10-item facet scales ranged from a low of .72 for E4 Activity level to a high of .92 for N2 Anger. When reduced to 4-item scales, alphas of at least .70 were maintained for every scale except C3 Dutifulness, which was .69.

The second phase involved examining the item content of every 4-item scale for three properties: near-duplicate items, references to disabilities or other areas that may result in legal problems, and fidelity to the content of items on Costa and McCrae's original NEO PI-R. In response to Wolfe's (1993) admonitions and email complaints about repetitiveness from persons who had taken the test, I replaced any item whose wording I judged to be too close to one of the other three items with a substitute that maintained the highest level of alpha reliability. For example, on O4 Adventurousness, two finalists for the four-item scale were "Dislike changes" and "Don't like the idea of change." The latter was replaced with "Prefer variety to routine."

As I examined all scales for near-duplicate items, I also attempted to insure that the content of all items was similar as possible to Costa and McCrae's items. Content similarity, couple with empirical correlations with the original NEO PI-R scales, increases the probability that evidence for the NEO PI-R's validity would apply to the shortened IPIP-NEO.

The only scale with questionable content similarity was O6 Liberalism, which, compared to its counterpart in the NEO PI-R, O6 Openness to Values, contained a disproportionate number of items dealing with crime, law, and order. It also contained an item referring to belief in one true religion, which is legally problematic. Replacing the dubious items produced a revised four-item O6 facet scale with an alpha reliability of .64. This is the only alpha other than the .69 for C3 Dutifulness that failed to reach the .70 level for the full sample. The .70 level could be achieved by adding two items to O6 and one item to C3, but I think the unbalance with two facet scales longer than the other 28 is more problematic than failure to reach the arbitrary .70 level. The highly successful Hogan Personality Inventory (HPI; Hogan & Hogan, 1992) contains many facet scales with alphas lower than .70.

The third phase of scale development included the computation of alphas for the five domain scales and the computation of alphas for domain and facet scales separately by sex. The alphas for each sex separately are nearly identical to alphas for the full sample for most scales. Alphas falling below .70 for males included E4 Activity Level (.69), O3 Emotionality (.68), and O6 Liberalism (.64). C3 Dutifulness for males was .70, compared to .69 in the full sample. Alphas falling below .70 for females included O3 Emotionality (.66), O6 Liberalism (.63), A6 Sympathy (.67), and C3 Dutifulness (.69). Again, I judged that these alphas are acceptably high, and that adding items to just a few facet scales would create an unbalance that would be more problematic than failure to reach the arbitrary .70 alpha level.

## Future Plans

I have a number of things on my “to do” list for the IPIP-NEO. I am still concerned about the occasional reports of the program crashing before a respondent has received results. I plan to continue to solicit help from expert programmers to attempt to eliminate this problem.

I would also like to compare IPIP-NEO responses collected on line with other information about the respondents, including demographic background, scores on other psychological measures, and significant life events (Buchanan, Goldberg, & Johnson, 1999). In particular, I would like to couple the collection of data on the IPIP-NEO with non-self-report data such as peer ratings of personality. Goldberg and I discussed this issue during my visit to the Oregon Research Institute, but I have not yet settled upon a viable method for connecting non-self-report data to IPIP-NEO responses.

One of the reasons the IPIP was established was to create public-domain versions of commercial instruments so that research could be conducted unimpeded by prohibitive costs and copyright restrictions. The procedure for developing these public domain proxies involves comparing scores on the commercial inventories to responses on to the IPIP. To this end, I have targeted two inventories for which I think IPIP proxies would be useful: Kolb's (1999) Learning Style Inventory (LSI) and Riso's (1999) Riso-Hudson Enneagram Type Indicator (RHETI). This past year I have begun analyzing relationships between responses to these two instruments and the IPIP-NEO.

Another project I have in mind is making available a data archive on the World Wide Web for other researchers who might be interested in conducting their own analyses of my IPIP-NEO data. Eventually I would like to secure funding to create a true “collaboratory” (Finholt & Olson, 1997), which would allow a community of personality researchers to post items from the IPIP and/or their own personality items to a common Web site to conduct research (Johnson, 2001).

Once I have screened the remaining 150,000+ cases for duplicates and non-responsiveness, I plan to investigate possible differences across countries to which people reported belonging. If significant differences are found, this could form the basis of an empirical report as well as more differentiated norms for the narrative report-scoring program. I also plan to examine the data for differences across a greater number of age groups. (Currently, only two age categories are used for norms: ages greater than 20 and ages less than 21). Again, differences could form the basis of an empirical report and more differentiated age norms for the narrative report (McCrae, et al., in press).

I also plan to continue investigating the problem of detecting non-responsiveness. In particular, I would like to examine the effects of excluding protocols failing to meet the various criteria of responsiveness as outlined in my seven-stage screening process. Ultimately I would like to see how these screening criteria affect validity, but until non-self-report validity data can be gathered, I will examine the effects of excluding cases on internal characteristics such as factor structure.

Finally, one of my original and long-standing concerns with this project has been assessing the validity of individual narrative reports (Johnson, 1994a). Normally we validate personality tests by examining covariation between test scores and meaningful criteria. Although based upon test scores, a narrative report itself is not numeric and therefore cannot be correlated with criteria. I have discussed alternative strategies for validating narrative reports, including Moreland's (1985) good suggestions, but it remains to be seen how these alternative strategies can be carried out on responses to a Web-based personality inventory.

Suggestions from the Audience

The participants in this joint Groningen-Bielfeld Research Group meeting made a number of good points and suggestions regarding the research described in this paper. I hope I have the attribution right in summarizing their comments and suggestions.

Karen van Oudenhoven-van der Zee wondered about the advisability of adjusting the scale norms for an Internet sample that might be more neurotic, disagreeable, and unconscientious than the general public. She wondered whether we are doing people a service when we tell them mostly nice things about themselves. My answer was that I thought that norms should be set to the most appropriate reference group we can identify as opposed to the general public, and that appropriate norms can be established only in the context of validation studies. I also noted that, except perhaps for the Neuroticism domain, the high and low ends of all the dimensions refer to both positive and negative characteristics.

Wim Hofstee and others thought it wise to compare the factor structures for the unscreened and screened and also suggested comparing the mean scores for the unscreened and screened samples. I said this was a good idea, although I would probably want to first eliminate protocols with too many missing responses. Hofsteee also suggested that I could detect deviant response patterns by correlating a respondent's answers to social desirability values for the items. Looking back at one of Goldberg's papers, I see that this is one of his standard checks for responsiveness, and it is certainly worth pursuing.

Alois Angleitner noted that Costa & McCrae (1992) reported in their NEO PI-R manual specific lengths of strings using the same response category that should be regarded as non-responsive answering. Indeed, the manual says the following: "Occasionally, respondents are uncooperative and complete the NEO-PI-R in a careless or random fashion. … One common form of random responding patterns can be evaluated by visually inspecting the answer sheet to determine whether the same response option has been used over a long series of items. Based upon the results of item response patterns in a volunteer sample (Costa & McCrae, in press-b), endorsements of *strongly agree* to more than 6 consecutive items, *disagree* to more than 9 consecutive items, *neutral* to more than 10 consecutive items, *agree* to more than 14 consecutive items, or *strongly agree* to more than 9 consecutive items invalidate formal scoring and interpretation of the NEO-PI-R" (p. 6). These are more conservative than the criteria I had chosen to reject cases (< 14, 15, 59, 19, and 27, respectively). I plan to request a copy of Costa & McCrae's unpublished book chapter to examine their rationale for their cutoff points and to examine the effect of different cutoffs for eliminating protocols on the properties of the scales (homogeneity, factor structure).

Alois Angleitner also suggested administering Goldberg's adjective markers with the IPIP-NEO and adding an item to the demographic section to determine birth order. When I gather acquaintance ratings on persons who complete the IPIP-NEO, I will certainly consider one of Goldberg's sets of markers, and I will probably add an item or two about siblings to determine birth order and family size.

My paper reading session concluded with a spirited discussion of validating narrative reports. Wim Hofstee suggested that the narrative report might have to be validated in piecemeal fashion, in which each personality description within the report is rated for accuracy. He suggested obtaining feedback from respondents on-line as they read and react to their report. Frank Spinath raised some important points about the ability of a narrative report to say more than the person has already said in his or her endorsement of items. If the report merely re-states the items endorsed by the respondent, there's no need to get the respondent's feedback, but if the report says more, then there is a question of whether the respondent can give appropriate feedback. My own feeling is that feedback from a respondent can be useful, but I would rather develop methods for ascertaining the validity of narrative reports from knowledgeable acquaintances.

References

Buchanan, T., Goldberg, L. R., & Johnson, J. A. (1999, November). *WWW personality assessment: Evaluation of an on-line five factor inventory*. Paper presented at the meeting of the Society for Computers in Psychology, Los Angeles, CA.

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

Costa, P. T., Jr., & McCrae, R. R. (in press). The NEO Personality Inventory. In S. R. Briggs and J. Cheek (Eds.), *Personality measures: Vol. 1*. Greenwich, CT: JAI Press. [Since the time this chapter was cited in the NEO manual, the title, editors, and publishing company have changed. The new reference is: *The Revised NEO Personality Inventory (NEO PI-R).* In: S.R. Briggs, J. Cheek, and E.M. Donahue (Eds.): Handbook of personality inventories. New York: Plenum, in press.

Finholt, T. A., & Olson, G. M. (1997). From laboratories to collaboratories: A new organizational form for scientific collaboration. *Psychological Science*, **8**, 28-36.

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, and F. Ostendorf (Eds.), *Personality Psychology in Europe* (Vol. 7, pp. 7-28). Tilburg, The Netherlands: Tilburg University Press, pp. 7-28.

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

Gough, H. G. (1987). *California Psychological Inventory administrator's guide*. Palo Alto, CA: Consulting Psychologists Press.

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

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

Johnson, J. A. (1991, June). *Interpreting the California Psychological Inventory with the AB5C model*. Invited paper presented to the Department of Psychology, University of Groningen, The Netherlands.

Johnson, J. A. (1993). *Generating computer narrative reports for the CPI, HPI, and other omnibus personality inventories with the AB5C model*. Manuscript originally prepared for a symposium on computer-generated personality reports, 101st Annual Convention of the American Psychological Association, August, 1993, Toronto, Canada, but the symposium did not take place. Available from the author at Penn State DuBois, DuBois, PA 15801.

Johnson, J. A. (1994a). *Computer narrative interpretations of individual profiles*. Chapter originally prepared for R. Hogan, J. Johnson, and S. Briggs (Eds.), *Handbook of personality psychology*, but withdrawn due to space limitations. Available from the author at Penn State DuBois, DuBois, PA 15801.

Johnson, J. A. (1994b). Multimethod replication of the AB5C model of personality traits. In B. De Raad, W. K. B. Hofstee, & G. L. M. Van Heck, *Personality psychology in Europe, Volume 5: Selected papers from the sixth European conference on personality held in Groningen, The Netherlands, June 1992.* (pp. 42-49). Tilburg, The Netherlands: Tilburg University 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, March). *Web-based personality assessment*. Paper presented at the 71st Annual Meeting of the Eastern Psychological Association, Baltimore, MD.

Johnson, J. A. (2001). *Proposal for funding a personality collaboratory*. Unpublished manuscript available from the author at Penn State DuBois, DuBois, PA 15801.

Johnson, J. A., & Ostendorf, F. (1993). Clarification of the five factor model with the Abridged Big Five-Dimensional Circumplex. *Journal of Personality and Social Psychology*, *65*, 563-576.

Kolb, D. A. (1999). *Learning Style Inventory*. Boston: Hay/McBer Training Resource Group.

Kraut, R., Lundmark, V., Patterson, M., Kiesler, S., Mukopadhyay, T., & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist*, **53**, 1017-1031.

McCrae R. R., Costa, P. T. Jr., de Lima, M. P., Simoes, A., Ostendorf, F., Angleitner, A., Marusic, I., Bratko, D., Caprara, G. V., Barbarenelli, C., Chae, J., & Piedmont, R., L. (in press). Age differences in personality across the adult lifespan: Parallels in five cultures. Developmental Psychology.

Moreland, K. L. (1985). Validation of computer-based test interpretations: Problems and prospects. *Journal of Consulting and Clinical Psychology*, *53*, 816-825.

Riso, D. R. (1999). *The Riso-Hudson Enneagram Type Indicator (Version 2.5)*. New York: The Enneagram Institute.

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

Smith, M. A., & Leigh, B. (1997). Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods, Instruments, & Computers*, **29**, 496-505.

SPSS, Inc. (1999). *SPSS Base 10.0 applications guide*. Chicago, Author.

Wiggins, J. S. (Ed.). (1996). *The five-factor model of personality: Theoretical perspectives*. New York: Guilford Press.

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