British Journal of Psychology (1999), **90**, 125–144 *Printed in Great Britain*

Using the Internet for psychological research: Personality testing on the World Wide Web

Tom Buchanan* and **John L. Smith**
Division of Psychology, University of Sunderland

The Internet is increasingly being used as a medium for psychological research. To assess the validity of such efforts, an electronic version of Gangestad & Snyder's (1985) revised self-monitoring questionnaire was placed at a site on the World Wide Web. In all, 963 responses were obtained through the Internet and these were compared with those from a group of 224 undergraduates who completed a paper-and-pencil version. Comparison of model fit indices obtained through confirmatory factor analyses indicated that the Internet-mediated version had similar psychometric properties to its conventional equivalent and compared favourably as a measure of self-monitoring. Reasons for possible superiority of Internet data are discussed. Results support the notion that Web-based personality assessment is possible, but stringent validation of test instruments is urged.

The growth of electronic communication networks and in particular the Internet has, in recent years, been phenomenal. Estimates (e.g. by Bride, 1996) of the number of people with access to the Internet reach the 50 million mark, and increase daily. Like these millions, psychologists have not been slow to afford themselves of the potential it presents for a range of activities, including research.

Interest was perhaps initially focused on support for traditional research through increased access to information. Allie (1995) describes the technology, software and resources available for researchers using the Internet: these include e-mail, electronic mailing lists, Usenet newsgroups (on-line discussion groups which function somewhat like bulletin boards) and the World Wide Web (also known as the WWW or simply 'the Web'). Given the rapid pace of technological development, for up-to-date information about what is currently available, the most recent editions of guides such as those by Thomas (1996) or Hahn (1996) should be consulted.

In a similar vein, Levy (1995) provides an overview of the history, nature and possibilities offered by the Internet in general and the World Wide Web in particular. Levy focuses on the provision of resources and information via the Web and discusses the implications of this new medium for traditional publishing methods. An example of exactly how the possibilities of the Internet can be exploited is provided by Krantz (1995) who describes the development and use of World Wide Web resources in one psychology department—resources similar to those being

* Requests for reprints should be addressed to Tom Buchanan, Division of Psychology, University of Sunderland, The Business School, St Peters Campus, St Peters Way, Sunderland SR6 0DD, UK.

placed on-line by rapidly increasing numbers of psychology departments and other organizations and individuals.

Information provision and the dissemination of research findings remain important roles of the WWW, as highlighted by developments in electronic publication (e.g. Harnad, 1995) and current debate over its status. However, the imagination of many psychologists has also been caught by another possibility the WWW presents: a new environment in which to do research, as well as to discuss it.

Psychological research on the Internet: Why?

A major attraction of the Internet as a research environment is that it provides access to very large numbers of potential research participants, who can be included in one's study at minimal expense and with relative ease. No outlay is required for laboratory space, testing time, materials and other expenses traditionally associated with research. Once the study has been set up it may effectively be left to run itself with little intervention: data acquisition and even analysis may be entirely automated. The sheer number and variety of people using the Internet also means that populations with special characteristics may be located and contacted (for instance, through Usenet newsgroups) more easily than has previously been the case.

Szabo & Frenkl (1996) discuss the possibilities the Internet and associated technologies present for various types of research. Drawing on ethical guidelines for psychological research (American Psychological Association, 1990) and current practice regarding the standards and modes of behaviour which Internet users expect from each other ('netiquette'; Shea, 1994), they present a set of recommendations for Internet-based research. Similarly Hewson, Laurent & Vogel (1996) present guidelines for conducting electronic analogues of traditional questionnaire-based studies, which they see as essentially analogous to postal surveys, and also discuss issues involved in translating other research designs to the new medium. At this point in time, the majority of Internet-mediated studies appear to be questionnaire based, with personality tests and surveys of various sorts being the most common, although Krantz, Ballard & Scher (1997), Schiano (1997) and Stern & Faber (1997) describe electronic implementations of other research paradigms.

There are various ways in which questionnaire-based designs may be implemented, depending on how participants are recruited, how test materials are delivered to them and how their responses are retrieved. Batinic (1997) offers general guidelines regarding the construction and use of questionnaire studies designed to be administered through e-mail, newsgroups and WWW pages and evaluates the relative merits of each format. He concludes that e-mail surveys are the poorest technique—although Anderson & Gansneder (1995) reported a more positive experience and review several studies in which they have been successfully implemented—and suggests that WWW-based studies are the best technique for on-line research.

In a typical WWW-based study, participants will access a specially designed WWW page containing a fill-out form (the questionnaire). Once this has been done they will select an option to submit their answers, which will then be automatically passed to a program for scoring or e-mailed to the experimenter. Schmidt (1997) describes how

such studies may be constructed, outlining hardware and software requirements, potential problems, recruitment strategies and other factors which should be considered. However, like Batinic (1997), Schmidt notes that as yet the validity of WWW-based studies has not been adequately assessed.

Validity of Internet-based research

This last point is an important one, as it is clear that many psychologists (of both professional and amateur flavours) are discussing, exploring or actively using the web as a research medium. Using the Déja News archive (see Barrett, 1997, for details) we carried out a search of Usenet articles for the period from 19 March 1995 to 26 June 1997 using the search string:

> (Internet OR WWW OR World Wide Web)AND (research OR test OR experiment OR questionnaire OR survey) AND (personality) OR (psychology)

This search yielded 14016 references. Many, if not most, of these are likely to be only ephemerally related to the question at hand, follow-ups to previous messages or duplicate messages posted to different newsgroups. However, even if this figure is vastly inflated, it does illustrate a considerable level of interest and activity.

In comparison, the refereed psychological literature remains almost silent on the issue. Using the same set of keywords, a search of the PsycLIT database of psychology journal articles for the period from January 1991 to March 1997 found only eight, of which just four (Bordia, 1996; Hewson, Laurent & Vogel, 1996; Lukoff, Lu, Turner & Gackenbach, 1995; Ransdell & Anderson, 1995) were directly relevant to the topic of Internet-mediated research.

The only study of which we are aware in which an Internet-mediated questionnaire was directly compared with a pencil-and-paper version is that of Smith & Leigh (1997), using a questionnaire designed to assess the nature and frequency of sexual fantasies experienced by respondents (Ellis & Symons, 1990). Through the newsgroup sci.psychology.research, 72 participants were recruited and did the test on-line.[1] As a comparison group, 56 introductory psychology students did a paper-and-pencil version of the questionnaire.

Smith & Leigh compared the demographic characteristics of the two groups and found that they differed in age and gender composition. However, the groups did not differ significantly in terms of sexual orientation, marital status, ethnicity, education or religiosity. They concluded from this that Internet samples are as representative of the general population as traditional student samples. We would add the qualification that this depends entirely on the sampling strategy being used—one cannot assume that all Internet samples will have similar demographic characteristics. This study essentially compared a sample comprised largely of psychologists with a sample of psychology students. Had their Internet group been recruited through a newsgroup catering to a different special interest, the result might be very different.

Smith & Leigh also compared responses of the two groups to 6 of the 28

[1] In the technique implemented in Smith & Leigh's (1997) study, questions were presented sequentially and interactively. This differs from the HTML form employed in our and most other studies, where questions appear on the screen in a visual format very similar to a traditional test.

questions, and found no significant differences in answers to the individual items. Furthermore, when their Internet and traditional samples were combined, they found the same pattern of sex differences in answers to individual items as did Ellis & Symons (1990). From this they concluded that the Internet can serve as both a primary participant pool and also a supplement to locally recruited participants, suggesting that the two types of participant can be combined within the same sample. We consider that this conclusion is premature: as noted above, this study to some extent compared like with like. The failure to detect differences between these samples using a subset of the questions does not necessarily mean that similar results would be found with all questionnaires and all Internet samples. As with the demographic variables, had this study's Internet participants been recruited through a different newsgroup (particularly one of those dealing with erotica) one would expect the responses of the two groups to be very different.

Smith & Leigh's study is an example of a self-report survey adapted for Internet administration, and as such it is representative of one of the main types of questionnaire-based study currently appearing on the WWW. Qualifications aside, the indications are that this adaptation was successful. Other WWW-mediated instruments set out to measure personality constructs, and the question which arises is whether or not they too can validly do so. While it is clear that the Internet offers potential for research, there are also threats to validity which have yet to be properly explored and assessed. It would be unfortunate if psychologists dismissed this potential by prematurely jumping to the conclusion that Internet questionnaires were mere novelties and lacked serious research credibility (e.g. Gold & Concar, 1996).

Computerized testing

The development of Internet (WWW)-based personality testing is an extension of the rise of psychometric tests administered using stand-alone computers. Almost any modern text on psychological assessment will include a section on computerized testing. Honaker & Fowler (1990) give an account of the history and development of computerized testing and note that most computerized tests and assessment procedures are designed for use in personality evaluation. Many (if not most) such tests are simply translations of traditional paper-and-pencil tests, with questions being presented on screen and responded to via the keyboard (Meier, 1994 notes several prominent examples).

Other types of test are more specialized and are facilitated by the nature of the assessment medium (e.g. adaptive testing, cognitive tasks and procedures which cross into the domain of experimental psychology, Meier, 1994). While such instruments would be equally amenable to WWW-based administration, it is the former type with which this discussion is mainly concerned.

One reason for the popularity of computer testing is that it automates the task of test administration, scoring and, in some cases, even interpretation of tests: Meier (1994) suggests that the increased efficiency offered by computerized tests will ensure their increasing use in test administration, although the interpretation of test results by computers is a different issue altogether, about which many (e.g. Cohen, Swerdlik & Smith, 1992; Kline, 1993*a*) have serious reservations.

With the translation of tests from paper to computer formats, the question arises of equivalence between forms (Skinner & Pakula, 1986). Bartram & Bayliss (1984) review research suggesting that computerized versions of traditional personality tests are generally equivalent to their paper-and-pencil antecedents, given that the tests are (*a*) not speeded and (*b*) require some form of multiple or forced-choice response to textual items, a conclusion endorsed by Cohen *et al.* (1992). However, these authors also discuss instances in which computerized and traditional versions have not been found to be equivalent.

Some studies (e.g. Levine, Ancill & Roberts, 1988; Locke & Gilbert, 1995) have found increased levels of self-disclosure when tests—particularly on personal or sensitive topics—are administered by computer. Others (e.g. Webster & Compeau, 1996) have found that for some measures different patterns of responses are seen using traditional and computerized tests. It is possible that in some cases levels of computer anxiety might affect participants' responses: a test might give different results for people who are not confident in the use of computers, especially if the construct being measured was in some way related to anxiety.

While it is unclear just how big an influence these and other factors may exert, both Meier (1994) and Cohen *et al.* (1992) agree that equivalence of computerized and traditional versions of tests cannot simply be assumed but must be demonstrated for each test. The situation is perhaps best summarized by Cronbach (1990): 'It seems that the conventional and computer versions of a test do usually measure the same variables, but difficulty or reliability can easily change. Whether the computer version is "the same test" must be questioned with each instrument in turn psychologically' (p. 48). This recommendation should apply especially to tests administered via the World Wide Web, as there are particular threats to the reliability and validity of these.

Threats to reliability and validity of Web-based tests

Nature of sample

In terms of age, sex, language, culture, education and socio-economic factors, as well as the construct one is attempting to measure, participants recruited via the Web are likely to vary to a greater degree than those recruited and tested by traditional means. While the population of WWW users is believed to be currently biased toward young males of above-average socio-economic and educational status (Schmidt, 1997), the heterogeneity of Internet users is increasing and one may no longer make assumptions about who or what they are.

One might argue that increased heterogeneity increases the representativeness of the sample and thus the external validity of findings (e.g. Gordon, Slade & Schmitt, 1987), although this in itself is a topic of debate (e.g. Berkowitz & Donnerstein, 1982; Greenberg, 1987). However, one also runs the risk of introducing numerous unknown confounding variables which might have the effect of increasing 'noise' in the data and reducing the proportion of variance in responses accounted for by differences on the causal entitie(s) one is trying to measure. While this is more likely to be a problem in adaptations of traditional experimental paradigms (e.g. Krantz *et*

al., 1997), in studies where personality tests are used one must take account of the possibility that factors other than those the test was designed to measure may contribute to variance on the scale.

Volunteer status of participants

Participants in Internet-mediated experiments typically make an active effort to participate (and sometimes even to seek out the experiment). They do this, presumably, through interest or curiosity, as the rewards they receive are intellectual in nature. These are true volunteers. In contrast, those who most commonly participate in research usually do so for quite different reasons, such as course credit or a financial incentive. Smart (1966) estimated that 80% of studies reported in the psychological literature used undergraduate student samples. We surveyed articles published in *British Journal of Social Psychology* in 1996 and found a similar figure: 33 out of 39 used students as participants. Other research participants may be recruited in occupational or clinical settings. While the motivational and external factors affecting this group may differ from those of college students, they are again less likely to be taking part purely out of interest in the research.

Participants in Web-based studies are therefore likely to have different motivation to take part, whether it be simple curiosity or pursuit of some personal agenda. This difference in motivation may affect results. Oakes (1972) discusses studies in which results obtained using coerced student samples differed from those obtained when true volunteers were recruited from the general populace. This does not necessarily imply that the results of studies using students may be ungeneralizable: under some circumstances, coerced students may be more representative of the general population than particular volunteer groups, 'The generalisability of any particular finding ... may be limited by interaction with behavioural characteristics peculiar to any population to which one is attempting to generalise' (Oakes, 1972, p. 962). Thus one might query whether a test which validly measures a psychological construct in the population with which it was developed will be equally valid when administered via the WWW.

Environmental factors and nature of the testing environment

Bartram & Bayliss (1984) list among the advantages of computerized tests the fact that they offer greater control over testing conditions and thus greater objectivity than situations in which a human tester administers the instrument. In the Web paradigm, this advantage is lost and the situation reversed. The researcher has absolutely no control over the conditions under which a test is completed: someone doing the test in, for example, a noisy computer lab will experience a different set of environmental stimuli or distractions to somebody completing the test on a computer in their home or workplace. Non-stable attributes of the individual (such as mood state, fatigue or intoxication) might also increase unexplained variance in responses. Such factors are not only beyond the control of researchers: they are beyond their awareness as well. Variables which might contribute to such effects include the anonymity of participants and the psychological distance between themselves and the experimenter. While evidence (see above) exists that the former

of these might increase honesty and self-disclosure, some authors (e.g. Kline, 1993*a*) place value on establishing rapport between test takers and administrators.

Technological factors

The test is likely to be completed by people using different browser software packages, each differently configured, running on different computer platforms with different displays. The presentation of the test will thus be different for every participant; Bartram & Bayliss (1984) suggest that there may be reasons to query the equivalence of tests administered via different hardware and software platforms. Also dependent on hardware, software and network configuration is the speed (time lag) with which documents may be accessed. Slow connections could lead to time delays which might frustrate participants.

Multiple completions and mischievous responding

Another problem arising from lack of control over the testing situation relates to independence of observations, as it would be entirely possible for the same people to complete the test more than once, perhaps answering in different ways (Schmidt, 1997; Smith & Leigh, 1997). Conventional tests are not free of problems such as dissimulation or faking (hence the development of 'lie scales'). However, while such dissimulation may be a problem, the particular danger for Web tests lies in the possibility of multiple completions. Difficulties may well arise from people accidentally submitting responses twice, or varying their answers to see how their final scores are affected—an activity facilitated by the fact that browser software typically permits one to go back and change answers to such forms. Steps must thus be taken to ensure independence of observations.[2]

Summary

One might consider some of the 'threats' outlined above to be sources of variance which could actually increase the generalizability of results. However, this is only the case if they are sources of variance which do not compromise accurate measurement of the construct of interest. Therefore, these factors make the validation of Web tests even more critical than the validation of normal computerized tests.

Establishing reliability and validity of Web tests

When an entirely new test is constructed for use on the Internet, the normal process of validation must take place (although procedures may be somewhat altered). The work of Pasveer & Ellard (1997) is an example of this. However, it is likely that most Internet-based tests will, at least initially, be translations of traditional paper-and-pencil instruments. It is with the latter variety that this discussion is concerned.

[2] One might also consider the possibility of faking on the part of a dishonest experimenter. It would seem easy simply to fabricate a data file and, given the anonymity of participants and absence of any physical experimental materials, difficult to check the authenticity of the data. However, examination of the WWW server's log files would indicate the number of times the experimental programme has been run—effectively faking a dataset would involve accessing the experiment many times from many different Internet addresses.

There is an obvious parallel between the translation of traditional tests to (stand-alone) computerized formats, and their translation to WWW-based formats. As noted above, the reliability and validity of at least some computerized tests has been shown to compare favourably with traditional formats. Establishing these properties has been done in a similar fashion to traditional measures—a computerized version of a traditional test can effectively be treated as an alternate or parallel form of that test. In addition to establishing the psychometric properties of the computerized test, Kline (1993*a*) considers it essential to demonstrate a satisfactory correlation (0.9 or above) between individuals' scores on the two forms if they are to be considered identical.

The validation of Web-based tests is somewhat more difficult, largely due to the way in which these tests are used. There are two major ways in which an Internet-mediated test differs from its paper-and-pencil equivalent: (*a*) the format of presentation and (*b*) the nature of the participants and the circumstances under which the test is likely to be taken. For the latter of these reasons, a Web test cannot simply be considered an alternate form and treated as described above.

Due to the anonymity of participants recruited via the Internet, and their wide geographical dispersion, bringing them in to complete a questionnaire under traditional face-to-face testing conditions would be difficult in the extreme. The reverse strategy of testing people first under traditional conditions and then with a WWW-based instrument is also not viable, as such a test would differ from a 'normal' Web test in terms of many of the special characteristics (e.g. motivation, self-selection, anonymity and sample heterogeneity) outlined above. Similarly, the assessment of test–retest reliability would be difficult, if not impossible, given the nature of the sample. One would have to obtain contact details for each participant and then follow them up at a later date. This means they would be doing the second test under different circumstances and through different motivation to the first. If it is the case that these factors might influence responding, then any reliability coefficient obtained in this way would be attenuated. However, assessment of reliability based on measures of internal consistency may be easily achieved through normal procedures such as determination of coefficient alpha.

Beyond establishing that a Web test is reliably measuring something, it must also be established that it is measuring the same variable(s) as its conventional counterpart. As noted above, it is probably not feasible to do this by examining correlations between participants' scores on the two versions. There are, however, some ways in which the psychometric properties of a Web-based test may be compared with its traditional equivalent.

When a test is revised, or a new version of an existing test is developed, an obvious question to ask is whether it is still measuring the same construct(s). This question may be partially answered by comparing the factor structures of both instruments[3] (e.g. Meesters, Muris, Bosma, Schouten & Beauving, 1996). If the tests are

[3] Were one interested in establishing whether differences (if any were found) between the Internet-mediated and traditional forms of test were caused by the method of data collection or the method of recruitment, a third condition should be included in which participants recruited by traditional means were required to do the electronic version. However, this would tell us little about the functional equivalence of Web-based and traditional tests, as the test in the third condition would essentially be a stand-alone computerized test.

equivalent, the same number of factors should account for similar proportions of variance and the same items should load on each factor (Tabachnick & Fidell, 1989), and the relationships between the factors should be the same. This may be tested through confirmatory factor analysis: if a model derived from exploratory factor analyses of a traditional test provides a good degree of fit to data obtained with a WWW equivalent, then we might say with some confidence that the two test versions had similar psychometric properties. The purpose of the present research is to implement this strategy to investigate the equivalence of Web-based and traditional test formats.

Method

Selection of appropriate scale

Four factors were considered in the choice of an appropriate scale: psychometric properties, length and format, previous publication and the construct being measured.

The first requirement of the instrument to be used is that it should have well-established and satisfactory psychometric properties. Indices of reliability should be known (so that comparisons can be made across the formats) and preferably high. As the factor structure of the test is to be used as a point of comparison, this should be known and should have been shown to be relatively stable across a number of samples.

Secondly, the format of the test should be suitable for the intended mode of administration: a series of items to which the test taker responds by selecting one from a range of options. The nature of the medium, and the findings noted above relating to equivalence of computerized and traditional test formats, require that the test not be speeded. Furthermore, it is desirable that the test be relatively brief, for the practical reason that a short test is less likely to induce boredom or fatigue (and thus increase the drop-out rate) in the volunteer participants.

A third requirement is that the test should be one which is published and freely available in the public domain, and for which the authors have granted permission for research use. This is mainly for reasons of test security. As Bartram & Bayliss (1984) point out, wide publication of a test (for example, by placing it on the Internet) could well reduce its utility in applied settings, especially if the details of how it is scored are also made available. Thus tests which are likely to be used in, for example, personnel selection or clinical and educational settings should be avoided for current purposes, and any test used should be made accessible for as short a time as possible.

The final point relates to the subject matter of the test, and to feedback. Users of traditional tests have a duty to provide meaningful feedback to test takers (Kline, 1993*a*) in a responsible and sensitive manner. It is perhaps even more important (though more difficult) to provide feedback from an automated test: apart from ethical and, in some countries, legal requirements, one must remember that most participants will be taking part mainly due to curiosity about their scores on the test. The simplest method of providing feedback is to present the score, and sufficient information for the test taker to interpret it. However, for ethical reasons this necessitates that the test should measure something relatively innocuous: distress might easily be caused by bluntly informing participants they had achieved a high score on a test of authoritarianism or psychoticism, or a low score on an intelligence test, or a pattern of test results indicating that they ran some health risk.

The Self-Monitoring Scale

An instrument which appeared to satisfy the requirements outlined above was the SMS–R (Gangestad & Snyder, 1985), the revised version of Mark Snyder's Self-Monitoring Scale (Snyder, 1974), which is, in the words of Briggs & Cheek (1986) '... a popular measure of personality' which has 'served as the centrepiece for a number of published articles' (p. 129).

The scale is a measure of the tendency to observe and regulate expressive behaviours and self-presentation. Individuals high in self-monitoring are sensitive to social and situational cues, and adjust

their behaviour accordingly. Low self-monitors, on the other hand, lack either the ability or motivation to do this and tend to behave in ways consistent with their stable personality attributes or internal states (Snyder & Gangestad, 1986). Over the years, versions of the scale have seen wide use and been translated into a number of languages including Spanish (Hosch & Marchioni, 1986), Chinese (Hamid, 1993), Greek (Malikiosi & Anderson, 1992) and Japanese (Ishihara & Mizuno, 1992).

Reliability and validity of the SMS–R. The reliability of the SMS–R appears satisfactory. Both Gangestad & Snyder (1985) and Briggs & Cheek (1988) report a coefficient alpha of 0.70 which, although less than ideal, does meet the criteria laid out by (e.g.) Nunally (1978) and Kline (1993*a*). Anderson (1991) reports a test–retest correlation of 0.55, which may be considered adequate given that the time interval was relatively long (two years).

The construct validity of the measure is generally considered to be well-established (e.g. Snyder, 1987). According to Briggs & Cheek (1988), 'Many studies have reported associations between the Self-Monitoring Scale and a wide range of important and conceptually relevant criteria' (p. 663).

Factor structure of the SMS–R. The factor structure of the Self-Monitoring Scale has been examined by a number of authors. It was criticism (e.g. by Briggs, Cheek & Buss, 1980) of the psychometric properties of the original 25-item scale (Snyder, 1974) which led to the development of the revised 18-item measure for which improved factorial purity is claimed (Gangestad & Snyder, 1985; Snyder & Gangestad, 1986). It is the English-language version of the revised 18-item measure, using a true–false response format, which is used here, and with which the following discussion is concerned.

Snyder & Gangestad (1986), through principal axes factor analysis, extracted three factors similar to those found in the 25-item scale by Briggs, Cheek & Buss (1980). Briggs & Cheek (1988) and Lennox (1988) have both published two-factor solutions which, according to Miller & Thayer (1989) are 'virtually identical' (p. 150). However, Miller & Thayer used confirmatory factor analysis to test the fit of one-, two- and three-factor models. They concluded that while both two- and three-factor solutions provided 'an excellent fit to the data' (p. 153), the level of fit was best for the three-factor solution. Similarly, in a confirmatory factor analysis by Hoyle & Lennox (1991), the best fit was found for a model with three intercorrelated latent variables. The consensus would seem to be that this last solution is the best fitting: a view endorsed by Gangestad & Snyder (1991) who accept that a three-factor solution in which at least two of the factors are correlated may be derived.

Given this fact, one might be surprised that a single score is typically derived from the scale. This has been a focus of the criticism (e.g. by Briggs & Cheek, 1986) and vigorous defence by Gangestad & Snyder (e.g. 1985, 1991). A main thread in the argument of the critics is that self-monitoring is best interpreted in terms of these three rotated factors (e.g. Briggs & Cheek, 1988) which have been labelled as 'other-directedness', 'extraversion' and 'acting ability' (e.g. Lennox, 1988). While Gangestad & Snyder agree that these three factors can be derived and do validly tap particular behavioural tendencies (e.g. Gangestad & Snyder, 1991), they counter the critics with the argument that the underlying self-monitoring construct is located in factor space very close to the first unrotated factor. This factor accounts for the majority of variance on the scale and correlates strongly with total scale score. In contrast, the second unrotated factor is estimated as correlating only weakly with the total scale score (Gangestad & Snyder, 1991).

For our purposes, the 'real' structure and nature of the self-monitoring construct matters little. The important question is whether or not a Web-mediated version of the test behaves in the same way as the pencil-and-paper version, and whether the same factorial structure can be found in both.

Development of Web version. The WWW-based questionnaire was constructed along lines similar to those described by Schmidt (1997) and presented in accord with the guidelines suggested by Szabo & Frenkl (1996). Data were acquired via an HTML (hypertext mark-up language, the code in which WWW documents are written) form and passed to a program which processed the input, saved the data to a file and provided feedback to the participant. These mechanisms, however, were invisible to participants, whose experience was as follows. The first thing seen was a screen entitled 'On-line Personality Test', bearing sufficient information about what would be involved in the study to ensure informed consent while maintaining naivety as to the particular construct being measured and the overall purpose of the study. Anonymity was assured, and participants were asked to complete the test only once. Contact details and affiliation of the experimenters were also included, as was an option to give us feedback through an interactive form.

In order to proceed beyond this introductory screen, a link labelled 'CONTINUE' had to be selected. On doing this, participants were presented with an interactive form bearing the questions from the SMS–R, and a set of instructions adapted from those used with the paper-and-pencil version. Respondents were required to respond to each question by selecting 'true' or 'false' from the options given. Some demographic details were also requested: sex, age, and whether or not they were students.

At the foot of the page, an option was presented which, on selection, would submit the form for processing. Assuming that all questions had been answered (if any blanks were left, a message was printed pointing this out along with a request to go back and complete the form), a page titled 'Results and Debriefing' was then displayed. On this, the questionnaire was named and a brief description of the construct being measured was given. The score achieved was printed, along with guidelines on how to interpret it.

Some information about the purpose of the study was also given, as was a full citation for Snyder (1987) for anyone interested in finding out more. At the bottom of the page, after contact details for the experimenters and a link to the feedback form, a message was printed thanking participants for their help and asking them not to do the test again.

Procedure

Once the interactive questionnaire had been constructed and extensively tested, participants were recruited by means of messages posted to Usenet newsgroups selected for their relevance to the study (alt.usenet.surveys, sci.research, sci.psychology, alt.psychology.personality and alt.psychology.help—all newsgroups in which messages soliciting participants for on-line tests had previously been seen), a procedure in accord with that suggested by Szabo & Frenkl (1996). The same message was also placed on a Web page which has pointers to some on-line experiments.

Each message (entitled 'Personality test available on-line') contained a brief invitation to participate in the study, the URL (Internet location) where the questionnaire could be found, and contact details for the experimenters. These messages were posted four times, at two-weekly intervals from 23 December 1996 to 3 February 1997. Once it was deemed that sufficient responses had been acquired, no further messages were posted (to avoid exhausting the goodwill of newsgroup readers). Data acquisition continued at a slower rate until 7 April 1997, when the document at the advertised URL was changed to a brief message outlining the goals of the study and thanking participants. The number of responses finally obtained was 1181.

Data screening and processing. As noted above, a study such as this entails the risk that some of the observations are actually multiple responses from the same people. When somebody submits a response to a Web-based questionnaire, unless they have been asked for some identifying information there is no way of telling who they are. However, it is possible to identify the unique Internet address of the computer they are using, and this information was logged during the study.

This information can be used to identify any multiple responses from the same computer, so that all but the first (chronological) response coming from any such address may be deleted from the data file. This approach is not foolproof, as a malicious individual who knew what they were doing could access the study from different computers at different times (the likelihood of this happening is probably low). It is also somewhat conservative, as multiple submissions could come from the same address (for instance, a heavily used machine in a computer laboratory) and still be valid. We decided, however, to err on the side of caution. Duplicate and multiple submissions from the same addresses were therefore screened out, leaving a total of 963 unique and valid responses.

Comparison group. For comparison purposes, a second sample was recruited and tested using traditional means. These participants were undergraduate students at the University of Sunderland, who were recruited and tested in small groups after classes using the paper-and-pencil version of the SMS–R.

Participants

Of the 963 participants tested using the Internet questionnaire (sample 1), 491 were male and 472 were female; 405 were students of some type, and participants ranged in age from 11 to 67 ($M = 32$ years,

SD = 10.79). One participant claimed to be five years old, but this was judged likely to be a typing mistake.

Determining where these participants were geographically located is difficult—even when the recorded Internet addresses contain a country identifier, there is no guarantee that the person using that machine is physically present in, or is a native of, that country. However, cursory examination of the logged addresses suggests that while responses were obtained from around the globe (with responses coming from machines located, in for example, the UK, Canada, Australia, France, Finland, New Zealand, Germany, South Africa, Japan, Hong Kong, Israel and Taiwan) the majority of respondents were probably from the USA.

The comparison group, sample 2, comprised 224 undergraduate volunteers, of whom 35 were male and 176 female (13 did not state their gender). All members of this group were students. Ages were only recorded for 74 of these participants, who ranged in age from 18 to 53 years (M = 27 years, SD = 8.10).

Results

Reliability

Coefficient alpha was determined for sample 1 and found to be 0.75. This compares favourably with the value of 0.70 consistently reported (e.g. by Briggs & Cheek, 1988; Hoyle & Lennox, 1991; Gangestad & Snyder, 1985) for the paper-and-pencil version of the test. Alpha was also determined for sample 2 and found to be 0.73.

Fit of factor solution

As outlined above, there is some consensus that a model with three intercorrelated latent variables underlies the self-monitoring scale. The fit of such a model to our data was thus tested in a confirmatory factor analysis using the package AMOS 3.6. In the model, items were specified as loading on the same factors as those identified (on the basis of a survey of the rotated factor loadings found in five exploratory analyses) in the analysis by Lennox (1988): factor 1 (other-directedness) comprises items 2, 8, 10, 11 and 18; factor 2 (extraversion), comprising items 12, 14, 22 and 23; and factor 3 (acting ability), which comprises items 1, 4, 5, 6, 8, 10, 18, 21 and 24. The three factors were permitted to correlate.

Confirmatory factor analysis uses the chi-square statistic to test the null hypothesis that the model fits the data (Crowley & Fan, 1997). When the model was tested with sample 1, the chi-square statistic was highly significant, with χ^2 (132, N = 963) = 669.7, $p < .000$, indicating a poor fit. However, it is known (e.g. Arbuckle, 1997; Crowley & Fan, 1997) that with large samples (as is the case here) the high power of the rest may lead to rejection of a model which actually provides an acceptable level of fit to the data.

For this reason, a number of other indices of model fit have been developed, and these are used in conjunction with the chi-square statistic to assess fit. In the confirmatory factor analysis of Miller & Thayer (1989), the fit indices reported were goodness of fit (GFI), adjusted goodness of fit (AGFI) and root mean square residual (RMS). For GFI and AGFI, values approaching 1 indicate a good fit, while RMS should approach zero (Arbuckle, 1997).

For sample 1, the values obtained for GFI, AGFI and RMS were 0.925, 0.903 and 0.014, respectively. This is a reasonable, if not particularly compelling, level of fit.

However, recall that we are interested in how well the Internet test compared with the traditional version, not the absolute veracity of the theoretical model of the structure of self-monitoring. Miller & Thayer (1989) tested the fit of the three-factor model to a sample obtained using the traditional version (albeit with a restricted item set with two items omitted) of the SMS–R. The values of GFI, AGFI and RMS reported were 0.905, 0.873 and 0.082 respectively, indicating that the level of fit found in that study was not as good as the level of fit found for sample 1.

The same analysis was performed for our comparison group, sample 2. Here the value obtained for GFI was 0.900, AGFI was 0.871 and RMS was 0.015. Again, all indicate a level of fit inferior to that obtained for the sample tested with the Internet questionnaire.

The other major confirmatory factor analysis which has tested the fit of the three-factor model is that of Hoyle & Lennox (1991). Given that Miller & Thayer did not use the complete SMS–R, this analysis is probably the best to use for comparison purposes: Hoyle & Lennox tested a model identical to that specified in our analysis.

While Hoyle & Lennox were less enthusiastic about the fit of the three-factor model, they did conclude that it was the best of those tested. The indices of fit which they report are the chi-square statistic, the Bentler–Bonnet normed fit index (NFI, or Δ) the Tucker–Lewis coefficient (TLI, or ρ and the mean absolute value of the residuals (which is not reported by AMOS). For the NFI and TLI, values approaching 1 are again considered indicative of good fit (Arbuckle, 1997).

As with our sample 1, Hoyle & Lennox obtained a significant value for the chi-square statistic: $\chi^2(132, N = 1113) = 966.83$, $p < .000$. Note however that this is higher than the value we obtained for sample 1, indicating that the fit of the model is poorer in Hoyle & Lennox's analysis. This suggestion is reinforced by the fact that the NFI and TLI values (0.756 and 0.759 respectively) obtained for sample 1 are higher than those reported by Hoyle & Lennox (0.616 and 0.591). These statistics were also computed for our comparison group, sample 2: NFI (0.670) was poorer than for sample 1, but TLI (0.791) was slightly higher.

The fit indices obtained in these analyses are summarized in Table 1. Across all points of comparison other than the chi-square statistic (which is problematic due to the large sample size), the fit indices obtained using sample 1 are better than those previously reported in tests of model fit using data acquired by traditional means. With the exception of the chi-square and TLI statistics, the fit indices obtained using sample 1 are also better than those obtained using our comparison group, sample 2.

Relationship of unrotated factors to full scale score

Recall that Snyder & Gangestad (1986) consider the first unrotated factor to capture much of the meaning and most of the variance of the entire scale. Is this still the case with the Internet version? To answer this question, a principal axes factor analysis with no rotation (in order to replicate that reported by Snyder & Gangestad, 1986) was performed upon sample 1. Factor scores were generated and the correlations between these and total scale scores examined. The first unrotated factor derived from sample 1 correlated very strongly ($r = 0.97$) with total score on the SMS–R, while the second unrotated factor correlated only weakly ($r = 0.14$) with total scale

Table 1. Summary of fit indices for different samples

	Sample			
Fit index	Sample 1 (Internet) ($N = 963$)	Sample 2 (comparison) ($N = 224$)	Hoyle & Lennox (1991) ($N = 1113$)	Miller & Thayer (1989) ($N = 266$)
χ^2	669.70	229.57	966.83	228.21
d.f.	132	132	132	101
GFI	0.925	0.900	—	0.905
AGFI	0.903	0.871	—	0.873
RMS	0.014	0.015	—	0.082
NFI	0.756	0.670	0.616	—
TLI	0.759	0.791	0.591	—

Note. Cells are empty where the relevant indices were not reported.

score. Gangestad & Snyder (1991) estimate values of 0.84 and 0.15 respectively for these correlations. When the same analyses were performed for the comparison group, sample 2, almost identical values (0.97 and 0.12 respectively) to those for sample 1 were obtained.

Pattern of loadings

Given that the first unrotated factor relates so strongly to the total score derived from the scale, it is instructive to examine the pattern of item loadings upon it. If the first factor is measuring the same construct for both Internet and conventional tests, items should load upon it in the same way. Tabachnick & Fidell (1989) describe a procedure whereby the pattern and magnitude of loadings in different data sets may be compared: compute the correlation between the two sets of loadings. A strong positive correlation ($r = 0.89$) was found between item loadings on the first unrotated factor in sample 1 and the corresponding loadings presented by Snyder & Gangestad (1986).

Norms

Finally, the means and distributions of scores gathered via the Web and traditional means were examined. Summary statistics are presented in Table 2. While there is no reason to suggest that means for the different populations should be the same (it has been reported, for example, that Chinese students have higher self-monitoring scores than New Zealanders; Hamid, 1993), the characteristics of the samples appear to be similar: indeed, an independent samples t test indicated no significant difference in means ($t(1185) = -1.23$, $p < .218$).

Table 2. Descriptive statistics for Internet and comparison groups

	Sample 1 (Internet)	Sample 2 (comparison)
N	963	224
Mean	9.232	9.576
SD	3.813	3.545
Kurtosis	−.559	−.324
Skew	−.110	−.095

Discussion

The psychometric properties of the Internet-mediated version of the SMS–R appear to compare favourably with its conventional equivalent. Indeed, it is tempting to claim that the Web-based test has better psychometric properties. Reliability is slightly higher than in either our comparison group or any of the major analyses of the scale reported in the literature. When the three-factor model commonly claimed to underlie the SMS–R was tested with confirmatory factor analysis, the level of fit found was comparable to that obtained for our comparison group and seemingly better than that reported in the literature for an identical model. In common with the analyses of Snyder & Gangestad (e.g. 1986), the first unrotated factor was found to account for the vast bulk of variance on the scale. The similarity of the item loadings upon this factor to those reported by Snyder and Gangestad suggests that the two formats of the test are measuring substantially the same thing: indeed, the higher correlation found here between the first factor and the total scale score suggests that the Internet version may actually provide a better measure.

So why might a Web-based test appear to provide a better measure of a personality trait than its conventional equivalent? As noted previously, increased levels of honesty and self-revelation have sometimes been found when computerized assessments are employed. This might facilitate more accurate measurement of a construct.

Another possible reason relates to the heterogeneity of the sample. If a test score is underpinned by a number of different factors, and participants' scores on a particular factor are relatively homogeneous, then that factor will not emerge as a strong source of variance (Kline, 1993*b*). Thus, factor solutions derived from particular groups may not be entirely representative of the wider population (Child, 1990). This effect might to some extent operate here. The analyses of both Hoyle & Lennox (1991) and Miller & Thayer (1989) were based on student samples. While it is clear that these samples were heterogeneous with regard to the factors underlying self-monitoring scores, it is possible that the Internet sample (which included many non-students and people from more diverse backgrounds) was even more so. It may thus be that due to the greater heterogeneity of the Internet sample a clearer picture emerges of the test's factor structure, not that the Internet test in itself is any better than its conventional equivalent.

This demonstrates one of the potential boons of the Internet for research. Psychology, famously described by McNemar (1946) as a science of sophomores, has a deserved reputation for relying on student samples. Because of this, assumptions about the generalizability of findings have been challenged (e.g. Smart, 1966). However, less than half the participants recruited through the Internet were students, suggesting that this methodology offers access to a wider population of potential participants than has often been the case (note, however, that this access extends only to that small portion of the world's population which has Internet access).

But let us not get ahead of ourselves, for the process of test validation is always a long one. We have presented some evidence that a WWW-based test may have comparable psychometric properties to its conventional antecedent, and that it reliably measures some construct. The factor structure of the instrument suggests that this is the same construct as that measured by the paper-and-pencil test, a test whose validity as a measure of self-monitoring is well established.

However, it is important to note that our failure to detect any differences between the factor structures or sample means does not necessarily mean that the Web version is valid as a measure of self-monitoring. Such a conclusion would require that the two participant groups be comparable in terms of the distribution of the underlying self-monitoring construct. This is something about which we have no independent information. Were the two groups to differ in reality, a failure to detect differences would actually indicate that one (or both) of the instruments lacked validity. While this study can tell us something about the reliability of the Internet-mediated test, it cannot tell us anything about its construct validity.[4] For that, further work is required: comparing the scores of groups of Internet-recruited participants who should differ in their self-monitoring tendencies.

Recall Nunally's caution that 'one validates not a measuring instrument, but rather some use to which the instrument is put' (1978, p. 87). The question which remains, therefore, is whether an Internet test will actually 'work' in the setting for which it is intended. In the current example, can our electronic implementation of the SMS–R differentiate between high and low self-monitors among the population of Internet users? While the evidence outlined above would tend to support the notion that it can, this is a notion which remains to be empirically tested (and which is being addressed in our current work).

One must also consider the sampling strategy which is used and the population from which an Internet sample is recruited, as this will affect the generalizability of results (Oakes, 1972). Like Smith & Leigh (1997), we recruited participants mainly from psychologically oriented newsgroups. These participants may thus have had much in common (e.g. shared interests and experiences) with our traditional sample of psychology students. Had we recruited participants via newsgroups likely to be frequented predominantly by people high or low in self-monitoring, a different pattern of results would be expected. Sampling strategies must therefore be carefully considered in implementing (and interpreting the results obtained from) a Web-based

[4] The question of validity cannot be addressed by examining the associations of the two sets of self-monitoring scores with the other variables measured, as self-monitoring is not believed to be meaningfully associated with either age or sex.

test, and it would seem that research on the kinds of samples obtained via different recruitment techniques is required.

Other questions remain to be answered. For instance, what is the subjective experience of a participant in such a study? Comments volunteered by participants in this study suggest this may vary from person to person. We received feedback, via the automated feedback form, from 37 respondents. This was transcribed into the ETHNOGRAPH package for coding purposes. As we had imposed no structure on this feedback, responses were very varied and we received both positive and negative comments of several different types (as the number of responses was small, we do not propose to report on these data in further detail in this paper, but will explore the themes identified in future work).

One might also pose questions as to why people choose to take part. Of the people who see the recruitment notice, how many visit the test's Web page? Of the people who visit the Web page, how many complete it and submit their answers for analysis? What sets them apart from those who do not? The demographic characteristics of our sample suggest that one important variable is gender. Given that the population of Internet users is currently believed to be predominantly male (Schmidt, 1997), the near equivalence of male and female participants in our Internet sample is surprising. One reason for this might be that women are more strongly represented in the newsgroups where we posted recruitment notices, or tend to be more interested in psychology and personality testing. This view is supported by the fact that more women tend to study psychology (Holdstock & Radford, 1998), a finding clearly reflected by the gender ratio in our traditionally recruited sample.

While it is clear that much work is required before Internet tests can be accorded a status equivalent to that of conventional instruments, it seems likely that Internet-mediated research can be done. What makes investigations of its validity crucial are the facts that such research is currently being done, and that for many of the instruments being used in this work, no evidence exists of reliability or validity. A parallel exists here with tests designed for use on stand-alone computers, where 'The advent of the cheap microcomputer... made it possible for people with little understanding of the principles and processes involved in the development of psychometric testing instruments to produce their own "tests"' (Bartram & Bayliss, 1984, p. 232). The advent of the World Wide Web has made it possible not only for people to do this, but to bring their 'tests' to the world—and it is likely that the number of people with the programming skills required to construct a Web-based test exceeds the number of those with the psychological expertise and experience required to design and assess the validity of such an instrument. The dangers of this situation are illustrated by the Usenet discussion group alt.psychology.personality, where much of the discourse centres around unvalidated on-line tests with little foundation in psychological theory. The neophyte stumbling across this group might well conclude that personality assessment begins and ends with such instruments.

We believe the Internet offers great potential for psychological research. However, if this potential is to be exploited, we must take care that on-line testing does not fall into disrepute. The only way in which this can be done is by taking pains to validate stringently the instruments and research paradigms we use. In conclusion, we suggest that the evidence presented here lends some degree of support to the notion

that Web-based personality assessment is a real possibility. However, as with stand-alone computerized tests (Cohen *et al.*, 1992), the question of whether an Internet test can be considered equivalent to its traditional antecedent must be answered individually for each test, and each use to which it is put. For the SMS–R, we have taken one step along this road.

Acknowledgements

We gratefully acknowledge the assistance of Paul Whitely and Nicky Berry in collecting a portion of the data for sample 2, and three anonymous reviewers for their helpful comments on an earlier version of this paper.

References

Allie, D. A. (1995). The Internet and research: Explanation and resources. *The Journal of Mind and Behaviour*, **16**, 339–368.

American Psychological Association (1990). Ethical principles of psychologists. *American Psychologist*, **45**, 390–395.

Anderson, L. R. (1991). Test–retest reliability of the Revised Self-monitoring Scale over a two-year period. *Psychological Reports*, **68**, 1057–1058.

Anderson, S. E. & Gansneder, B. M. (1995). Using electronic mail surveys and computer-monitored data for studying computer-mediated communication systems. *Social Science Computer Review*, **13**, 33–46.

Arbuckle, J. L. (1997). *AMOS User's Guide Version 3.6*. Chicago: SmallWaters Corporation.

Barrett, D. (1997). NET SMARTS—Don't hunt blind through Usenet for the information you need. Let the search engine in Déja News do the walking. *Keyboard*, **23**, 108.

Bartram, D. & Bayliss, R. (1984). Automated testing: Past, present and future. *Journal of Occupational Psychology*, **57**, 221–237.

Batinic, B. (1997). How to make an Internet based survey? In W. Bandilla & F. Faulbaum (Eds), *SoftStat '97 Advances in Statistical Software 6*, pp. 125–132. Stuttgart: Lucius & Lucius.

Berkowitz, L. & Donnerstein, E. (1982). External validity is more than skin deep. *American Psychologist*, **37**, 245–257.

Bordia, P. (1996). Studying verbal interaction on the Internet: The case of rumour transmission research. *Behavior Research Methods, Instruments and Computers*, **28**, 149–151.

Bride, M. (1996). *Teach yourself the Internet*. London: Hodder Headline.

Briggs, S. R. & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality*, **54**, 106–148.

Briggs, S. R. & Cheek, J. M. (1988). On the nature of self monitoring: Problems with assessment, problems with validity. *Journal of Personality and Social Psychology*, **54**, 663–678.

Briggs, S. R., Cheek, J. M. & Buss, A. H. (1980). An analysis of the Self-Monitoring Scale. *Journal of Personality and Social Psychology*, **38**, 679–686.

Child, D. (1990). *The essentials of factor analysis*, 2nd ed. London: Cassell Educational.

Cohen, R. J., Swerdlik, M. E. & Smith, D. K. (1992). *Psychological testing and assessment*, 2nd ed. Mountain View, CA: Mayfield Publishing.

Cronbach, L. J. (1990). *Essentials of psychological testing*, 5th ed. New York: Harper Collins.

Crowley, S. L. & Fan, X. (1997). Structural equation modelling: Basic concepts and applications in personality assessment research. *Journal of Personality Assessment*, **68**, 508–531.

Ellis, B. J. & Symons, D. (1990). Sex differences in sexual fantasies: An evolutionary psychological approach. *Journal of Sex Research*, **27**, 527–555.

Gangestad, S. W. & Snyder, M. (1985). 'To carve nature at its joints': On the existence of discrete classes in personality. *Psychological Review*, **92**, 317–340.

Gangestad, S. W. & Snyder, M. (1991). Taxonomic analysis redux: Some statistical considerations for testing a latent class model. *Journal of Personality and Social Psychology*, **61**, 141–146.

Gold, K. & Concar, D. (1996). Elusive EQ: is emotional intelligence more than just a clever talking point? *New Scientist*, Supplement, 27 April, pp. 8–9.

Gordon, M. E., Slade, L. A. & Schmitt, N. (1987). Student guinea pigs: Porcine predictors and particularistic phenomena. *Academy of Management Review*, **12**, 160–163.

Greenberg, J. (1987). The college sophomore as guinea pig: Setting the record straight. *Academy of Management Review*, **12**, 157–159.

Hahn, H. (1996). *Harley Hahn's student guide to Unix*, 2nd ed. New York: McGraw-Hill.

Hamid, P. N. (1993). Self-monitoring and ethnic group membership. *Psychological Reports*, **72**, 1347–1350.

Harnad, S. (1995). The post Gutenberg galaxy: How to get there from here. *Information Society*, **11**, 285–292.

Hewson, C. M., Laurent, D. & Vogel, C. M. (1996). Proper methodologies for psychological and sociological studies conducted via the Internet. *Behavior Research Methods, Instruments and Computers*, **28**, 186–191.

Holdstock, L., & Radford, J. (1998). Psychology passes its 1997 exams. *The Psychologist*, **11**, 117–119.

Honaker, L. M. & Fowler, R. D. (1990). Computer-assisted psychological assessment. In G. Goldstein & M. Hersen (Eds), *Handbook of psychological assessment*, 2nd ed., pp. 521–545. New York: Pergamon.

Hosch, H. M. & Marchioni, P. M. (1986). The Self-Monitoring Scale: A factorial comparison among Mexicans, Mexican Americans and Anglo Americans. *Hispanic Journal of Behavioral Sciences*, **8**, 225–242.

Hoyle, R. H. & Lennox, R. D. (1991). Latent structure of self-monitoring. *Multivariate Behavior Research*, **26**, 511–540.

Ishihara, S. & Mizuno, K. (1992). A study of revised Self-Monitoring Scale. *Japanese Journal of Psychology*, **63**, 47–50.

Kline, P. (1993*a*). *The handbook of psychological testing*. London: Routledge.

Kline, P. (1993*b*). *Personality: The psychometric view*. London: Routledge.

Krantz, J. H. (1995). Linked Gopher and World Wide Web services for the American Psychological Society and Hanover College Psychology Department. *Behavior Research Methods, Instruments, and Computers*, **27**, 193–197.

Krantz, J. H., Ballard, J. & Scher, J. (1997). Comparing the results of laboratory and World Wide Web samples of the determinants of female attractiveness. *Behavior Research Methods, Instruments, and Computers*, **29**, 264–269.

Lennox, R. D. (1988). The problem with self-monitoring: A two-sided scale and a one-sided theory. *Journal of Personality Assessment*, **52**, 58–73.

Levine, S., Ancill, R. J. & Roberts, A. P. (1988). Assessment of suicide risk by computer-delivered self-rating questionnaire: Preliminary findings. *Acta Psychiatrica Scandinavica*, **80**, 216–220.

Levy, C. M. (1995). Mosaic and the information superhighway: A virtual tiger in your tank. *Behavior Research Methods, Instruments, and Computers*, **27**, 187–192.

Locke, S. D. & Gilbert, B. O. (1995). Method of psychological assessment, self disclosure, and experiential differences: A study of computer, questionnaire and interview assessment formats. *Journal of Social Behavior and Personality*, **10**, 255–263.

Lukoff, D., Lu, F., Turner, R. & Gackenbach, J. (1995). Transpersonal psychology research review: Researching religious and spiritual problems on the Internet. *Journal of Transpersonal Psychology*, **27**, 153–170.

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, **43**, 289–374.

Malikiosi, L. M. & Anderson, L. R. (1992). Reliability data on a Greek translation of the Revised Self-Monitoring Scale. *Psychological Reports*, **71**, 544–546.

Meesters, C., Muris, P., Bosma, H., Schouten, E. & Beauving, S. (1996). Psychometric evaluation of the Dutch version of the Aggression Questionnaire. *Behavior Research and Therapy*, **34**, 839–843.

Meier, S. (1994). *The chronic crisis in psychological measurement and assessment: A historical survey*. San Diego: Academic Press.

Miller, M. L. & Thayer, J. F. (1989). On the existence of discrete classes in personality: Is self-monitoring the correct joint to carve? *Journal of Personality and Social Psychology*, **57**, 143–155.

Nunally, J. C. (1978). *Psychometric theory*, 2nd ed. New York: McGraw-Hill.

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

Pasveer, K. A. & Ellard, J. H. (1997). *The making of a self-trust questionnaire: Help from the WWW.* Paper presented at the meeting of the Society for Computers in Psychology, Philadelphia, November.

Ransdell, S. E. & Anderson, M. D. (1995). Establishing a SCiP list for year-round discussion: Keeping the information ball rolling. *Behavior Research Methods, Instruments, and Computers*, **27**, 116–119.

Schiano, D. J. (1997). Convergent methodologies in cyber-psychology: A case study. *Behavior Research Methods, Instruments, and Computers*, **29**, 270–273.

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

Shea, V. (1994). Core rules of netiquette. *Educom Review*, **29**, 58–62.

Skinner, H. A. & Pakula, A. (1986). Challenge of computers in psychological assessment. *Professional Psychology: Research and Practice*, **17**, 44–50.

Smart, R. (1966). Subject selection bias in psychological research. *Canadian Psychologist*, **7**, 115–121.

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

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology*, **30**, 526–537.

Snyder, M. (1987). *Public appearances/Private realities.* New York: W. H. Freeman.

Snyder, M. & Gangestad, S. W. (1986). On the nature of self-monitoring: Matters of assessment, matters of validity. *Journal of Personality and Social Psychology*, **51**, 125–139.

Stern, S. E. & Faber, J. E. (1997). The lost e-mail method: Milgram's lost-letter technique in the age of the Internet. *Behavior Research Methods, Instruments, and Computers*, **29**, 260–263.

Szabo, A. & Frenkl, R. (1996). Consideration of research on Internet: Guidelines and implications for human movement studies. *Clinical Kinesiology*, **50**, 58–65.

Tabachnick, B. G. & Fidell, L. S. (1989). *Using multivariate statistics*, 2nd ed. New York: Harper & Row.

Thomas, B. J. (1996). *The Internet for scientists and engineers.* Oxford: Oxford University Press.

Webster, J. & Compeau, D. (1996). Computer-assisted versus paper-and-pencil administration of questionnaires. *Behavior Research Methods, Instruments, and Computers*, **28**, 567–576.

Received 20 November 1997; revised version received 24 March 1998