Behavior Research Methods
2005, 37 (1), 148-154

Nonequivalence of on-line and paper-and-pencil psychological tests: The case of the prospective memory questionnaire

TOM BUCHANAN and TARICK ALI
University of Westminster, London, England

THOMAS M. HEFFERNAN
Northumbria University, Newcastle Upon Tyne, England

JONATHAN LING
University of Teesside, Middlesbrough, England

ANDREW C. PARROTT
University of Wales Swansea, Swansea, Wales

JACQUI RODGERS
University of Newcastle Upon Tyne, Newcastle Upon Tyne, England

and

ANDREW B. SCHOLEY
Northumbria University, Newcastle Upon Tyne, England

There is growing evidence that Internet-mediated psychological tests can have satisfactory psychometric properties and can measure the same constructs as traditional versions. However, equivalence cannot be taken for granted. The prospective memory questionnaire (PMQ; Hannon, Adams, Harrington, Fries-Dias, & Gibson, 1995) was used in an on-line study exploring links between drug use and memory (Rodgers et al., 2003). The PMQ has four factor-analytically derived subscales. In a large ($N = 763$) sample tested via the Internet, only two factors could be recovered; the other two subscales were essentially meaningless. This demonstration of nonequivalence underlines the importance of on-line test validation. Without examination of its psychometric properties, one cannot be sure that a test administered via the Internet actually measures the intended construct.

Internet-mediated psychological tests appear to be here to stay. In recent years, there has been considerable recognition of their potential and increasing use in a number of psychological fields. These diverse fields include occupational assessments (e.g., Bartram, 1998), mental health applications (e.g., Buchanan, 2002), career counseling (e.g., Barak, 2003), psychological self-help (e.g., Barak & Buchanan, 2004), and various research applications (e.g., Buchanan, 2000b; Schmidt, 1997).

Whereas some on-line tests make use of innovative question formats and item content, the majority of the instruments used to date closely resemble traditional paper-and-pencil psychometric tests. Typically, a series of questions is presented on one or more Web pages, and test takers use their browsers to indicate their responses by clicking radio buttons, selecting checkboxes, and so on. Some tests used on line are developed specifically for Internet use (e.g., Pasveer & Ellard, 1998); others—probably more common—are on-line versions of tests previously used and validated in paper-and-pencil format. When an existing test is converted from off-line to Web-based format, it is important to establish that it is actually measuring what it is meant to, rather than assuming that because the paper-and-pencil version "worked," so will the on-line version.

Equivalence of On-Line and Off-Line Versions of Tests

Clearly, validity and fitness-for-purpose are more important requirements for on-line tests than any stipulation that they must exactly mirror an off-line version. However, the literature to date indicates that many tests being used on line have previously existed in paper-and-pencil formats. The fact that there is a previously vali-

A portion of the data presented in this article was presented at the 2002 German Online Research conference. Correspondence concerning this article should be sent to T. Buchanan, Department of Psychology, University of Westminster, 309 Regent Street, London W1B 2UW, England (e-mail: buchant@wmin.ac.uk).

Note—This article was accepted by the previous editor, Jonathan Vaughan.

Copyright 2005 Psychonomic Society, Inc.

dated, reliable, psychometrically satisfactory and widely used off-line version of a test already in existence might lead one to assume that the on-line version has the same properties. However, one of us has argued on several occasions (Buchanan, 2002, 2003; Buchanan, Johnson, & Goldberg, 2005; Buchanan & Smith, 1999b) that equivalence between on-line and off-line measures must be demonstrated, rather than assumed. The rationale for this argument is that the characteristics of the testing medium (such as anonymity, the use of computer-mediated communication and resultant phenomena such as *disinhibition effects*, or reduced socially desirable responding) or the samples used (often differing in makeup and motivation from those used in development and validation of the off-line measure) may impact upon a measure's psychometric properties and, ultimately, its power to reliably and validly measure the construct(s) of interest.

A number of studies have been published demonstrating that on-line tests can "work" satisfactorily and that, in many cases, on-line versions of tests are equivalent to traditional paper-and-pencil versions of the same instruments. For example, Meyerson and Tryon (2003) demonstrated that an on-line version of a sexual boredom scale had correlations with other scales that mirrored those of an original off-line version and almost identical reliability coefficients. This indicated that the two versions of the tests were essentially parallel and, thus, psychometrically equivalent. Similarly, a series of studies in which the psychometric properties of a self-monitoring scale were examined demonstrated that the on-line measure had satisfactory reliability and a factor structure comparable to that of an off-line version. Differences in the scores of criterion groups and replication of previous off-line findings also provided some evidence of construct validity (Buchanan, 2000a; Buchanan & Smith, 1999a, 1999b). There have been other studies in a similar vein: Barak and English (2002) Buchanan (2001), and Epstein and Klinkenberg (2001), among others, have reviewed this literature. Findings from this body of research have been encouraging—so encouraging, in fact, that they led Epstein and Klinkenberg to suggest that "because research has demonstrated that translating a measure into a computerized format does not necessarily change its reliability and validity, new Internet-based investigations need not be so focused on demonstrating the equivalency of paper-and-pencil instruments to computerized versions that are identical in every other way" (p. 310). We believe, however, that there are some difficulties with this suggestion.

We agree that translation to computerized format will not *necessarily* change the reliability and validity of a test. However, we cannot be sure that some tests will not change. We agree that *proof of concept* that on-line and off-line tests can be equivalent has been established, so it is no longer a major research priority. However, we do not believe that this will be the case for each and every test, even if the versions are identical in every way other than the administration medium. For that reason, researchers still should examine the properties of the tests that they use on line.

This idea is supported by a number of studies that have shown differences between on-line and off-line versions of the same tests in terms of both the score distributions achieved and the psychometric properties of the tests (e.g., factor structure). In many cases, these differences can be attributed to differences between samples tested on line and off line (e.g., Buchanan, 2003).

However, such differences have also been observed in comparisons of groups of people (e.g., students randomly assigned to on-line or off-line testing conditions) who should not vary in terms of the constructs being addressed. In such cases, sample differences should not be a contributing factor to any on-line–off-line effects found. For example, Joinson (1999) administered a social desirability questionnaire to student participants assigned to either on-line or off-line testing conditions and found lower levels of socially desirable responding in the on-line condition. Davis (1999) tested similar samples via the World-Wide Web (WWW) or paper-and-pencil techniques and found higher levels of self-focused negative thought in the WWW sample. Fouladi, McCarthy, and Moller (2002) randomly assigned student participants to either on-line or off-line testing conditions and found differences between the two groups in mean scores on mood-related scales. Barak and Cohen (2002) randomly assigned high school students to complete a multidimensional career preference inventory either on line or in paper-and-pencil format. Again, they found differences between the conditions, with elevated scores in the WWW condition. In none of these cases was there any reason to believe that the "true scores" on the underlying constructs addressed by the measures should differ between the groups that were compared. The most likely explanation in each case is that the differences found were in some way a function of mode of administration, rather than sample differences.

The differences reported in score distributions have important implications for the ways on-line tests are used, most especially with respect to the use of normative data (see Buchanan, 2003, for more discussion of this point). The most important implication is that normative data gathered in traditional settings may well be unsuitable as a basis for interpretation of scores from on-line versions of tests. It may prove that for some tests (e.g., Bartram & Brown, 2003; Cronk & West, 2002), off-line norms are appropriate for use, but this is a question that needs to be addressed empirically for individual tests.

However, differences in score distributions are relatively unimportant when compared with the other type of difference that has been reported. As long as one knows that, say, people will score higher on a measure of neuroticism administered on line and takes that into account in one's treatment of the data, there may be little to worry about. Considerably more problematic would be a situation in which a test does not actually measure what

one thinks—for example, when some or all of the items of a particular scale do not load on the expected latent construct. A number of such examples have been reported in the literature to date. Buchanan et al. (2005), working with an on-line version of a five-factor personality inventory, found that the latent structure of the inventory appeared to have changed slightly: A small number of the items loaded on factors other than those they had loaded on in the off-line development sample. In a separate evaluation of another Web-based five-factor inventory, Johnson (2000) similarly found that a small number of the facet (subscale) level constructs appeared to load on unexpected dimensions. Woolhouse and Myers (1999), who compared on-line and pencil-and-paper versions of a Jungian personality inventory, also found differences in the factor structures obtained for the two modes of assessment.

Does On-Line–Off-Line Equivalence Matter?

Any differences found between on-line and off-line versions of scales are usually relatively minor. This leads one to question of whether they actually matter. When one administers a traditional test to different samples, one would not be surprised to find slightly different loadings of items on latent variables, perhaps as a result of differences between the samples. This may well be the case for the examples outlined above, and given that the differences are usually small, the measurement ability and predictive value of the tests in question may not be compromised.

For example, in Buchanan et al.'s (2005) work on an on-line five-factor inventory, the finding that the factor structure of their inventory differed slightly from the predicted model motivated the identification of a subset of items that would make up a factor-univocal version (one in which all the items of a subscale loaded substantively on the appropriate factor, and not on any other factors). Factor-univocal scales are desirable in measures of orthogonal constructs, such as the five-factor model of personality. Both the original and the revised (factor-univocal) versions of the inventory correlated similarly with various behavioral self-report measures. In terms of predictive power, differences between the two versions were trivial. In terms of the extent to which they "worked," therefore, the minor difference found between the original on-line and off-line versions seemed to have no impact: The unmodified on-line version still appeared to measure the expected constructs.

However, we are aware of at least one case in which assuming that an on-line measure was equivalent to the paper-and-pencil instrument would have led to serious problems. The present article comprises an examination of the factor structure of an instrument used in an on-line study of links between recreational drug use and memory (Rodgers et al., 2001, 2003). In that study, participants (some of whom were drug users, some of whom were not) completed a number of questionnaires. One of the measures used was the Prospective Memory Questionnaire (PMQ; Hannon, Adams, Harrington, Fries-Dias, & Gibson, 1995). This is a self-report measure that asks people to provide estimates of the frequency with which they have various failures in prospective memory (memory to do things at some point in the future, such as taking a book back to the library or making an important phone call).

Hannon et al.'s (1995) measure of prospective memory has four subscales. The Long-Term Episodic (LT) subscale relates to memory for irregularly scheduled tasks, which require completion some hours or days after a cue to perform it (e.g., "I forgot to send a card for a birthday or anniversary"). The Short-Term Habitual (ST) subscale addresses memory for tasks to be completed shortly after the relevant cue that occur on a regular basis (e.g., "I forgot to put a stamp on a letter before mailing it"). The Internally Cued (IC) subscale addresses memory for tasks that do not have a clear external cue (e.g., "I forgot what I wanted to say in the middle of a sentence"). The final, somewhat different subscale is the Techniques to Remember (TR) subscale. This measures the use of strategies to aid prospective memory (e.g., "I rehearse things in my mind so I will not forget to do them").

A number of investigators have assessed links between scores on self-report memory questionnaires and recreational use of the drug ecstasy (3, 4-methylenedioxymethamphetamine; MDMA). The PMQ has been one of the instruments used in this manner (Heffernan, Ling, & Scholey, 2001). The data reported here were acquired as part of a project intended to address the drug/memory link, using Internet research techniques (see Rodgers et al., 2001, 2003, for full details). Given that we were not aware of any previous on-line research using the PMQ, we sought to establish whether it was actually suitable for on-line use. The key question addressed in the present analysis was whether the four constructs the PMQ sets out to address are actually reflected in data acquired using the on-line version of the measure.

METHOD

Materials

A set of WWW pages and Perl CGI scripts was created for the purposes of data acquisition, hosted on the University of Westminster Web server and accessible via a number of different addresses (e.g., www.drugresearch.org.uk). In addition to the PMQ, the participants were asked to complete the Everyday Memory Questionnaire (EMQ), which is a 28-item inventory measuring the frequency of memory slips, such as forgetting where things are usually kept (Sunderland, Harris, & Baddeley, 1983), and a modified version of the Recreational Drug Use Questionnaire (Parrott, Sisk, & Turner, 2000), which asks respondents to estimate their level of use of ecstasy and a number of other drugs. For all questions regarding drugs, a "prefer not to answer" option was also included. The participants also answered a number of demographic questions (age, sex, location, occupation, and education) and questions relating to their participation (how they found out about the study, whether they were currently under the influence of any substance, and whether there was any reason their data should not be used in analyses). All of these instruments were presented as forms on a single Web page.

Different response formats (clicking on radio buttons or selecting options from a drop-down menu) were used within the questionnaires. It has been argued (Dillman & Bowker, 2001) that drop-down menus raise problems because they do not replicate any existing paper-and-pencil format. However, for the questionnaire of primary interest here (the PMQ), drop-down menus were not used, and radio buttons arranged in a manner closely replicating the paper version were employed. For a full account of the questionnaires other than the PMQ, see Rodgers et al. (2003).

Procedure

The participants were recruited using a variety of methods. These included messages posted to relevant Usenet discussion groups (e.g., alt.drugs.ecstasy), links from other on-line experiments, notices on WWW pages, announcements in our home institutions, and e-mails to personal contacts. Different WWW addresses were given in different recruitment methods (e.g., www.drugresearch.org.uk and survey.drugresearch.org.uk).

The participants first saw an informed consent page describing the study and the kind of questions that would be asked and giving a link to a statement on anonymity and confidentiality. This assured the participants that no information from which they could be personally identified would be requested and that they could select "prefer not to answer" options (or withdraw completely) when appropriate. Those who wished to continue clicked on a button reading "I understand the nature of the study and wish to continue" as an indicator of informed consent. The participants then saw a page bearing brief instructions, demographic items, the EMQ, the PMQ and drug use questionnaires, and questions about their participation. Having completed all the items, they then clicked on a button labeled "Finished" at the bottom of the page. The participants who had not answered all the questions then saw a page asking them to return to the form and fill it out completely prior to resubmission. (The number of incomplete submissions was recorded and used in analyses of drug effects; Rodgers et al., 2003.) The participants who had answered all the items saw a debriefing page. This thanked them, outlined the purpose of the study, and provided links to several Web sites with information about drugs and also a link to a Web page where a summary of the results would be posted upon conclusion of the study. An e-mail contact address was also provided for respondents who wished to give us feedback or ask questions.

Data Screening and Processing

To control for possible multiple submissions from the same people (Buchanan & Smith, 1999b; Schmidt, 1997), we recorded all IP addresses of participants accessing the site. For ethical reasons, IP addresses were not stored in the same file as information about drug use; this was done to maximize the nonidentifiability of the participants by splitting off their IP address from any information about their illegal drug use. Instead, the CGI script used to process the data compared the IP of each fresh submission with the list of IPs of previous participants, and those that duplicated previous addresses were automatically flagged in the data file. Any submissions duplicating previous IP addresses were excluded from analysis. This is a technique that is becoming less satisfactory with the increase in dynamic IP addressing and the use of proxy servers by Internet Service providers—different individuals who are assigned the same IP will be excluded. However, given the subject (illegal drug use) of the study, more "ironclad" procedures, such as e-mail/password techniques, seemed undesirable, since they may be more likely to compromise the privacy of respondents (and to reduce response rates). Also flagged up were instances in which the participants indicated they were under the influence of some substance or that there was some reason their data should not be used. Application of these criteria led to the exclusion of 435 of the initial 1,199 submissions.

Fraudulent or mischievous data entry is harder to detect. One technique often employed is to use demographic information to screen out implausible responses (e.g., very young respondents claiming to have doctoral degrees). This led to the exclusion of one response (a person in the 16–20 age group claiming to have postgraduate education). Other data provided were consistent with the view that people were answering seriously—for example, nobody selected Antarctica as a location or claimed to have been recruited via a Web site on which we did not advertise.

Participants

Seven hundred sixty-three responses met our inclusion criteria. Of these, 465 (60.9%) were female. The modal age group was 21–25 (32%). The majority of the respondents came from Europe (71%). Many were well educated, having some university or college education (31%). About half the sample (49%) indicated that they were students.

RESULTS

Analyses pertaining to drug use and memory are reported elsewhere (Rodgers et al., 2003). The present analysis focuses on the psychometric properties of the PMQ. To compare the factor structure of the PMQ in our data with the four-factor structure reported by Hannon et al. (1995), we repeated their analysis (principal components followed by Varimax rotation) specifying extraction of the first four factors. Item loadings on these factors are shown in Table 1 (left-hand side).

Examination of the highest loading of each item indicates that those making up the LT and TR subscales group together in the expected way. However, the highest loadings of the items making up the ST and IC subscales appear to be scattered across three different factors in each case. In many instances, there are substantive loadings on more than one factor. The ST and IC subscales are clearly not factor univocal.

Differences between samples was previously advanced as a possible reason for differences in data from on-line and off-line test administrations. The demographic composition of the present sample is somewhat different from the sample used by Hannon et al. (1995), which constituted largely students. Accordingly, the analysis was repeated for a subset of the present sample—those who reported their occupations as students of some sort. The factor structure for this subset is also shown in Table 1 (right-hand side). The latent structure for the student subsample is very similar to that for the whole sample: Again, the LT and TR factors are clearly defined, whereas the ST and IC items are again scattered over three different factors and do not form discrete clusters. A third factor analysis, based only on nonstudents, also produced very similar findings.

In terms of reliability, Hannon et al. (1995) reported alphas "ranging between .78 and .90" (p. 292) for the PMQ subscales. In our sample, reliabilities for the LT, TR, ST, and IC subscales were .85, .89, .68, and .86, respectively (Cronbach's alpha). The high alpha for IC is a little surprising in view of the suggestion that these items do not delineate a clear construct. However, some of the items within this set do seem to form clusters, and in addition, many of them also have substantive loadings on the first factor (the factor defined by the LT items). The

Table 1
Loadings of PMQ Items on the First Four Extracted Components Following Varimax Rotation

		Full Sample (N = 763)				Students Only (n = 375)			
Item	Subscale	1	2	3	4	1	2	3	4
1	LT	**.69**	.01	.09	.04	**.57**	.03	.07	.12
2	LT	**.62**	.08	.04	.28	**.58**	.09	.18	.08
3	LT	**.43**	.12	.10	.15	**.53**	−.01	.10	.04
4	LT	**.69**	.09	.11	.14	**.60**	.11	.04	.02
5	LT	**.56**	.05	−.01	.05	**.50**	.08	.12	−.05
6	LT	**.71**	.05	.13	.07	**.68**	.06	.06	−.02
7	LT	**.49**	.32	.12	.15	**.53**	.28	.05	.11
8	LT	**.59**	−.06	.24	.03	**.50**	−.16	.09	.11
9	LT	**.60**	.07	.16	.12	**.54**	−.04	.24	.09
10	LT	**.62**	.15	.22	−.03	**.64**	.10	.12	.01
11	LT	**.67**	.14	.11	.10	**.68**	.04	.08	.17
12	LT	**.50**	.18	.08	.08	**.59**	.12	.05	.16
13	LT	**.53**	.14	.27	−.01	**.56**	.04	.04	.11
14	LT	**.57**	.14	.08	.09	**.47**	.06	−.13	.03
15	ST	.19	.03	**.28**	.20	.18	.18	−.03	**.50**
16	ST	.12	.06	−.06	**.22**	**.30**	−.02	.12	−.02
17	ST	−.10	.10	.24	**.58**	.12	.07	**.30**	.22
18	ST	−.06	.08	**.50**	.39	−.08	.02	.29	**.54**
19	ST	.07	.08	**.38**	.31	.03	.03	**.55**	.07
20	ST	.17	.10	**.75**	−.10	.13	.05	−.02	**.66**
21	ST	.04	−.01	**.82**	.22	.09	.11	.22	**.56**
22	ST	.23	.10	**.28**	.19	.14	.01	**.40**	.06
23	ST	.21	.08	**.33**	.18	.10	−.02	**.46**	.17
24	ST	.21	.05	.17	**.34**	−.03	.01	**.47**	.22
25	ST	.12	.13	**.76**	−.03	.02	.09	.19	**.67**
26	ST	−.01	.09	.16	**.67**	.08	.01	**.39**	.30
27	ST	.09	−.02	**.56**	−.06	.11	−.06	**.55**	.20
28	ST	**.41**	.07	.16	.20	**.43**	.16	.06	.25
29	IC	.42	.16	.08	**.62**	.38	.26	**.58**	−.23
30	IC	.51	.17	.08	**.59**	.45	.29	**.59**	−.25
31	IC	.44	.21	.10	**.63**	.38	.33	**.56**	−.14
32	IC	.41	.21	.04	**.63**	.46	.29	**.57**	−.14
33	IC	**.45**	.20	.10	**.45**	**.43**	.12	.40	−.18
34	IC	**.57**	.15	−.01	.35	**.57**	.14	.25	.00
35	IC	.21	.15	**.61**	.40	.15	.17	.38	**.49**
36	IC	.35	−.04	**.49**	.27	**.43**	.06	.39	.15
37	IC	**.40**	.06	**.40**	.13	.32	.26	.14	.05
38	IC	.39	.03	**.60**	−.05	.26	.24	−.07	**.50**
39	TR	.13	**.76**	.00	−.04	.10	**.77**	−.16	.10
40	TR	.11	**.75**	.00	.09	.05	**.79**	−.06	.07
41	TR	.09	**.61**	.24	−.06	.10	**.63**	−.07	.19
42	TR	.00	**.78**	−.02	−.06	−.03	**.79**	−.06	.06
43	TR	.21	**.67**	.04	.24	.01	**.69**	.19	.03
44	TR	.04	**.39**	.21	.27	.03	**.43**	.12	.13
45	TR	.25	**.57**	.05	.35	.22	**.57**	.25	−.01
46	TR	.06	**.66**	.03	.29	.05	**.59**	.21	−.02
47	TR	.05	**.71**	−.01	.05	.09	**.68**	−.07	.06
48	TR	.06	**.52**	.08	.32	.12	**.54**	.16	−.04
49	TR	.10	**.64**	.06	.24	−.04	**.65**	.13	.06
50	TR	.29	**.53**	.12	.22	.16	**.59**	.19	−.03
51	TR	.13	**.61**	.09	.11	.09	**.53**	.01	.13
52	TR	.09	**.66**	.01	−.09	.00	**.60**	−.13	.15

Note—Highest loadings of each item are shown in bold. LT, long-term episodic subscale; ST short-term habitual subscale; IC, internally cued subscale; TR, techniques to remember subscale.

alpha values for the student subsample are similar (.85, .68, .85, and .89, respectively).

The other instruments used (EMQ and the drug use questionnaire) were not multifactorial in nature and, so, were not amenable to examination of latent structure. However, reliability data for the EMQ are available. The value of Cronbach's alpha in the present sample was .94, which compares well with published values for the traditional version (e.g., .899; Cornish, 2000).

DISCUSSION

In the present on-line sample, two of the PMQ's subscales (LT and TR) appear to form coherent constructs that, on the basis of the item content, seem likely to address the areas specified in Hannon et al.'s (1995) model of prospective memory. The items of the other two subscales (ST and IC) do not appear to cluster together in the way one would expect if they loaded on discrete latent constructs. Therefore, in the present data set, there are no grounds for saying that these subscales measure *anything*, let alone the constructs delineated by Hannon et al. To have assumed that on-line and off-line versions of this test would be equivalent would clearly have been wrong.

The implication for the work of Rodgers et al. (2003) was, therefore, that the IC and ST subscales should not be included in their analysis. Had they been included, erroneous conclusions regarding links between recreational drug use and these aspects of prospective memory might have been reported. The same is true for other projects in which other on-line scales have been used: If the scales do not work as one expects, misleading results may be obtained. The obvious solution is to check that any on-line scales one uses *do* work as expected, as a precursor to any other analyses. Differences found, if any, may be minor (as in most cases) and of relatively little practical significance. In other cases (as in the present data), they may be large and important to know about.

The question then arises of why the factor structure observed in our sample varied from that described by Hannon et al. (1995). There are two possible classes of answers: differences in the samples used and differences in the method of test administration.

Hannon et al.'s (1995) development samples consisted largely of students (85% across two samples), with a small minority of the participants coming from brain-injured and alcoholic populations. The present on-line sample, however, is somewhat more diverse: Only around half were students. It is possible that findings are not fully generalizable from one of these populations to the other. As well as simple differences in score distributions, differences between samples may lead to different views of the latent structure of a questionnaire. Buchanan and Smith (1999b) suggested that Internet samples, by virtue of their heterogeneity, might provide clearer pictures of the factor structure of tests than traditional student samples do. The rationale for this suggestion is that more heterogeneous samples are likely to display more variance on the latent constructs of a test, thereby making the covariance structure clearer (a parallel situation is attenuation of correlations due to the restricted range on one or the other of the correlated variables). Variance on the latent constructs the PMQ sets out to address was possibly lower for Hannon et al.'s (mainly student) de-

velopment samples than in the present data, leading to a different observed factor structure.

However, when our analysis was repeated for a subset of the present sample that more closely resembled Hannon et al.'s (1995) sample (students), a pattern of factor loadings similar to that of our full sample was observed. The LT and TR subscales emerged, but the others did not (Table 1, right-hand side). This analysis suggests that differences in sample composition are not likely to explain the discrepancy between the present findings and Hannon et al.'s model of prospective memory. A more likely explanation is, therefore, that the effect arises in some way from differences in the mode of questionnaire administration.

The clearest difference between the two administration modes is that our implementation of the PMQ was administered on line, whereas the original version was administered in paper-and-pencil format. A possible explanation for the present findings is, therefore, that the observed nonequivalence is due to some factor associated with the Internet: Completion under conditions of anonymity and reduced experimental demands (Hewson, Laurent, & Vogel, 1996), lack of supervision or social presence (Bartram & Brown, 2003), changes in attentional focus, lowered socially desirable responding, increased self-disclosure (Joinson, 1999), variance in environmental or intraindividual conditions (Buchanan & Smith, 1999b; Reips, 2000), different degrees or types of motivation (Buchanan & Smith, 1999b), or some usability-related characteristic of the interface are all factors that have been suggested might impact upon online psychological research.

Although there has been much speculation about factors that might possibly affect Internet-mediated data, there has been little work to date that actually pins such effects down. An exception was reported by Baron and Siepmann (2000), who describe a study in which responses from on-line and paper-and-pencil respondents were found to differ. Baron and Siepmann attributed this finding to differences in the way the questions were laid out in the two versions of their questionnaire. This implies that the presentation of items in an on-line questionnaire should be as similar as possible to any off-line original (as was the case in the present study). However, as Baron and Siepmann noted, because of the way Web browsers work, it is very difficult to control how a page is presented to the viewer. Dillman and Bowker (2001) also noted this as a likely cause of measurement error for Web surveys. Although they suggested ways to minimize measurement error, realistically one must acknowledge that there is likely to be at least minor variance in how materials appear to respondents. A pragmatic approach to resolving the question is to attempt to develop Web-based research materials that are robust to such variation: questionnaires that measure the desired constructs irrespective of the *noise* that is inevitably introduced by Web-based presentation (it has been argued [Reips, 2000] that this technical variance is actually a good thing, since it is likely to increase the generalizability of findings).

Any or all of the factors outlined above might contribute to the present findings, as might other variables with which they are confounded in the present study. As well as the assessment medium, our procedure varied from Hannon et al.'s (1995) in a number of other ways. These range from the potentially very influential (e.g., presence of an experimenter or other people in the testing situation and higher statistical power due to the considerably larger sample in the present study) to the relatively minor (e.g., completion of other questionnaires alongside the PMQ). In order to disentangle the possible reasons for the present findings, further work is required in which these variables are systematically varied (e.g., completion of questionnaires in either paper or electronic formats, either alone or in the presence of an experimenter). Such work is required not only for the PMQ, but also for a wide range of other instruments, in order to identify the mechanisms underpinning any possible *Internet administration effect* and, perhaps, specific psychological constructs likely to be affected in this way (e.g., computer anxiety and, perhaps, negative affect in general; Buchanan, 2003).

Another methodological question arises from the present findings. If the on-line and off-line versions of a questionnaire such as the PMQ give different impressions of its latent structure, which is "correct" in terms of providing the most accurate picture of the real psychological variables underlying the questionnaire responses? There is evidence that due to the size and diversity of samples available, the WWW can play a valuable role in test development (e.g., Pasveer & Ellard, 1998) and, furthermore, that samples obtained on line may be more generalizable to a target population than to a traditional student sample (e.g., Horswill & Coster, 2001). It may well be that the WWW will become the method of choice for psychometric questionnaire development and administration in the future. This again, however, depends on both the identification of any possible impact that Internet administration may have on measurement properties and the demonstration of test validity.

The suggestion that the WWW may be a better environment for test evaluation and administration also raises the possibility that the present data actually provide a better picture of the latent constructs underpinning responses to the items of the PMQ and that the psychometric model for prospective memory may require further explication. Clearly, no conclusions can be drawn about this on the basis of the present data, due to the confounding of possible explanations for the differences. More work is required to adequately assess the covariance structure of this instrument in both on-line and off-line samples. It is likely that such work would provide insights into the nature of prospective memory and its subsidiary constructs. Likewise, it would be instructive to perform such comparisons with samples varying in demographic characteristics and lifestyle factors (an obvious example in the context of the present study being the level of recreational drug use), to determine whether similar factor structures are obtained across different groups.

In conclusion, we found that the on-line version of the PMQ did not have the same latent structure as its pencil-and-paper antecedent. Whatever the reason, this observation has important practical implications. The data presented in this article indicate that despite the general finding that on-line versions of traditional tests usually seem to measure the same things, there may well be instances in which this is not the case. In the case of the PMQ, two of the four subscales manifestly did not "work" with the present sample. In cases in which researchers use psychological tests on line, therefore, we recommend that an examination of the test's psychometric properties be performed as a matter of course. Only in this way can we ensure the veracity of findings obtained with on-line measures.

REFERENCES

Barak, A. (2003). Ethical and professional issues in career assessment on the Internet. *Journal of Career Assessment*, **11**, 3-21.

Barak, A., & Buchanan, T. (2004). Internet-based psychological testing and assessment. In R. Kraus, J. S. Zack, & G. Stricker (Eds.), *Online counseling: A handbook for mental health professionals* (pp. 217-239). Amsterdam: Elsevier, Academic Press.

Barak, A., & Cohen, L. (2002). Empirical examination of an online version of the self-directed search. *Journal of Career Assessment*, **10**, 387-400.

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

Baron, J., & Siepmann, M. (2000). Techniques for creating and using Web questionnaires in research and teaching. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 235-265). San Diego: Academic Press.

Bartram, D. (1998, January). *Distance assessment: Psychological assessment through the Internet.* Paper presented at the 1998 British Psychological Society Division of Occupational Psychology Conference, Eastbourne, U.K.

Bartram, D., & Brown, A. (2003, January). Online testing: Mode of administration and the stability of OPQ 32i scores. Paper presented at the British Psychological Society Occupational Psychology Conference, Bournemouth, U.K.

Buchanan, T. (2000a). Internet research: Self-monitoring and judgments of attractiveness. *Behavior Research Methods, Instruments, & Computers*, **32**, 521-527.

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

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

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

Buchanan, T. (2003). Internet based questionnaire assessment: Appropriate use in clinical contexts. *Cognitive Behaviour Therapy*, **32**, 100-109.

Buchanan, T., Johnson, J. A., & Goldberg, L. R. (2005). Implementing a five-factor personality inventory for use on the Internet. *European Journal of Psychological Assessment*, **21**, 115-127.

Buchanan, T., & Smith, J. L. (1999a). Research on the Internet: Validation of a World-Wide Web mediated personality scale. *Behavior Research Methods, Instruments, & Computers*, **31**, 565-571.

Buchanan, T., & Smith, J. L. (1999b). Using the Internet for psychological research: Personality testing on the World-Wide Web. *British Journal of Psychology*, **90**, 125-144.

Cornish, I. M. (2000). Factor structure of the everyday memory questionnaire. *British Journal of Psychology*, **91**, 427-438.

Cronk, B. C., & West, J. L. (2002). Personality research on the Internet: A comparison of Web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments, & Computers*, **34**, 177-180.

Davis, R. N. (1999). Web-based administration of a personality questionnaire: Comparison with traditional methods. *Behavior Research Methods, Instruments, & Computers*, **31**, 572-577.

Dillman, D. A., & Bowker, D. K. (2001). The web questionnaire challenge to survey methodologists. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 159-178). Lengerich, Germany: Pabst.

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

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

Hannon, R., Adams, P., Harrington, S., Fries-Dias, C., & Gibson, M. T. (1995). Effects of brain injury and age on prospective memory self-rating and performance. *Rehabilitation Psychology*, **40**, 289-297.

Heffernan, T. M., Ling, J., & Scholey, A. B. (2001). Prospective memory deficits in "ecstasy" users. *Human Psychopharmacology: Clinical & Experimental*, **16**, 607-612.

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

Horswill, M. S., & Coster, M. E. (2001). User-controlled photographic animations, photograph-based questions, and questionnaires: Three Internet-based instruments for measuring drivers' risk-taking behavior. *Behavior Research Methods, Instruments, & Computers*, **33**, 46-58.

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

Joinson, A. (1999). Social desirability, anonymity, and Internet-based questionnaires. *Behavior Research Methods, Instruments, & Computers*, **31**, 433-438.

Meyerson, P., & Tryon, W. W. (2003). Validating Internet research: A test of the psychometric equivalence of Internet and in-person samples. *Behavior Research Methods, Instruments, & Computers*, **35**, 614-620.

Parrott, A. C., Sisk, E., & Turner, J. J. D. (2000). Psychobiological problems in heavy "ecstasy" users. *Drug & Alcohol Dependence*, **60**, 105-110.

Pasveer, K. A., & Ellard, J. H. (1998). The making of a personality inventory: Help from the WWW. *Behavior Research Methods, Instruments, & Computers*, **30**, 309-313.

Reips, U.-D. (2000). The Web experiment method: Advantages, disadvantages, and solutions. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 69-117). San Diego: Academic Press.

Rodgers, J., Buchanan, T., Scholey, A. B., Heffernan, T. M., Ling, J., & Parrott, A. C. (2001). Differential effects of ecstasy and cannabis on self-reports of memory ability: A web-based study. *Human Psychopharmacology: Clinical & Experimental*, **16**, 619-625.

Rodgers, J., Buchanan, T., Scholey, A. B., Heffernan, T. M., Ling, J., & Parrott, A. C. (2003). Patterns of drug use and the influence of gender on self-reports of memory ability in ecstasy users: A web-based study. *Journal of Psychopharmacology*, **17**, 389-396.

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

Sunderland, A., Harris, J. E., & Baddeley, A. D. (1983). Do laboratory tests predict everyday memory? A neuropsychological study. *Journal of Verbal Learning & Verbal Behavior*, **22**, 341-357.

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

(Manuscript received March 9, 2003;
revision accepted for publication July 9, 2004.)