Behavior Research Methods
2005, 37 (4), 691-702

Measuring health beliefs on the Internet: A comparison of paper and Internet administrations of the Multidimensional Health Locus of Control Scale

CLAIRE HEWSON and JOHN P. CHARLTON
University of Bolton, Bolton, England

A growing number of studies have supported the use of unidimensional psychometric test instruments administered via the Internet; however, support for the use of multidimensional scales is weak. The present study compares paper and Internet administrations of the Multidimensional Health Locus of Control (MHLC) Scale (Wallston & Wallston, 1981). In terms of reliabilities and factor structures, the Internet data were found to be at least as good as the paper data. MHLC scores were comparable for paper and Internet administrations, although the Internet sample scored significantly lower on the Powerful Others subscale. Overall, the results show that administration of the MHLC Scale via the Internet can produce data comparable to that obtained by pen-and-paper methods. However, it is concluded that generalization of these findings beyond the psychometric test instrument and sampling procedures used here is not warranted.

The scope of the Internet as a primary research tool in the social and behavioral sciences is now becoming well documented (e.g., Birnbaum, 2000; Hewson, Yule, Laurent, & Vogel, 2003; Mann & Stewart, 2000). The potential advantages, disadvantages, caveats, and pitfalls have been recapitulated many times over (e.g., Birnbaum, 2004; Hewson, Laurent, & Vogel, 1996; Reips, 2000; Smith & Leigh, 1997). Within this context, a growing number of studies are emerging that attempt to validate Internet-mediated research (IMR) procedures across various domains. Although an increasing number of studies are reporting IMR data to be valid and reliable, when compared with data collected via traditional modes (Krantz & Dalal, 2000), generalizability of these results from one test instrument or procedure to another cannot be assumed (Buchanan & Smith, 1999). The present article contributes to this area by comparing online and offline administrations of the Multidimensional Health Locus of Control (MHLC) Scale (Wallston & Wallston, 1981). This research is of particular interest because, so far, support for the use of multidimensional scales administered via the Internet has been weak.

Validation of IMR Procedures

There are a number of reasons why administration of a research procedure or test instrument over the Internet may produce different results than does administration by non-Internet methods. Differences in sample composition, mode of administration, and levels of researcher control could impact upon the results obtained in IMR (Barbeite & Weiss, 2004; Hewson, 2003; Krantz, Ballard, & Scher, 1997).

Early concerns about the lack of generalizability of IMR data due to the biased nature of the Internet user population and, thus, of Internet-accessed samples (e.g., Bordia, 1996; Schmidt, 1997) are now less prevalent, probably owing to both the observed rapid growth in the size and diversity of the Internet user population (e.g., ISC, 2004) and the growing number of studies that have shown equivalence of Internet and non-Internet data *despite* differences in sample composition (e.g., Best, Krueger, Hubbard, & Smith, 2001; Birnbaum, 2002; Buchanan & Smith, 1999; Krantz et al., 1997; Riva, Teruzzi, & Anolli, 2003). One of the key appeals of IMR is the ability to reach a large number of potential participants cost and time effectively, using procedures such as posting participation requests to newsgroups, to mailing lists, and on Web pages (Barbeite & Weiss, 2004; Musch & Reips, 2000). In psychological IMR at least, these approaches (we include those that involve sending requests to e-mail addresses obtained from Internet databases, such as mailing lists) have prevailed (e.g., Birnbaum, 2001; Browndyke, Santa Maria, Pinkston, & Gouvier, 1998; Buchanan, 2000; Buchanan & Smith, 1999; Coomber, 1997; Corley & Scheepers, 2002; Eichstaedt, 2002; Im & Chee, 2004; Kaye & Johnson, 1999; Krantz et al., 1997; Laugwitz, 2001; Riva et al., 2003; Smith & Leigh, 1997; Szabo, Frenkl, & Caputo, 1996).[1] Interestingly (but perhaps not surprisingly), these studies provide evidence that Internet samples accessed using these methods tend to differ in *systematic* ways from the undergraduate student samples often encountered in psychological

We thank several anonymous reviewers for providing very helpful comments on earlier versions of the manuscript. We also thank Kenneth Wallston for permission to place the MHLC Scale on the Internet. Correspondence concerning this article should be addressed to C. Hewson, Department of Psychology and Life Sciences, University of Bolton, Deane Rd., Bolton, BL3 5AB, England (e-mail: ch5@bolton.ac.uk).

research (for evidence that traditional psychological research relies heavily on undergraduate student samples, see Buchanan & Smith, 1999; Hewson et al., 2003; Smart, 1966). Thus, Internet samples tend to be more diverse in nationality (or *geographical location* as an approximate indicator of nationality; Birnbaum, 1999; Buchanan & Smith, 1999; Krantz et al., 1997; Senior et al., 1999), of higher educational level (Birnbaum, 1999, 2000), more balanced in terms of gender (Bailey, Foote, & Throckmorton, 2000; Buchanan, 2000; Buchanan & Smith, 1999; Riva et al., 2003; Smith & Leigh, 1997), and broader in age range (Buchanan, 2000; Buchanan & Smith, 1999; Eichstaedt, 2002; Krantz et al., 1997; Riva et al., 2003; Senior et al., 1999; Smith & Leigh, 1997).

It would appear, then, that Internet samples display a tendency to be more diverse than traditional undergraduate samples on a number of dimensions (see also Krantz & Dalal, 2000) and that, in at least some psychological research contexts, findings have been shown to be robust in response to such sample variation. Indeed, it has been argued that IMR allows psychological researchers to move beyond the traditional student sample and, thus, obtain data that are more widely generalizable than would otherwise be possible (e.g., Krantz & Dalal, 2000). However, different sampling approaches will almost certainly give rise to different types of samples (as may the same approaches administered at different times), and researchers must bear this in mind when deciding upon the most appropriate procedures for any given study. One successful approach to date has been to target specialist newsgroups in order to reach difficult-to-access populations (e.g., Birnbaum, 2001; Coomber, 1997). Unfortunately, however, at this stage still relatively little is known about the relationship between sampling approach and sample composition in IMR, and this issue requires further investigation.

Mode of administration could also impact upon the data obtained in IMR. There is some evidence that participants responding on the Internet are less susceptible to social desirability effects (Joinson, 1999), are more candid in their responses (Joinson, 2001), and display lower levels of *risk aversion* (Shavit, Sonsino, & Benzion, 2001). Thus, some evidence for mode of administration effects exists. However, the vast majority of IMR validation studies to date have reported equivalence between Internet and non-Internet implementations, despite differences in both mode of administration and sample composition (e.g., see the studies cited above). Furthermore, those studies that have controlled for sample equivalence also have typically failed to show a mode effect (e.g., Cronk & West, 2002; Epstein, Klinkenberg, Wiley, & McKinley, 2001; Herrero & Meneses, in press; Huang, 2006; Knapp & Kirk, 2003; Metzger, Kristof, & Yoest, 2003; Meyerson & Tryon, 2003; Smither, Walker, & Yap, 2004). This would suggest that mode of administration, at least in the contexts studied to date, often can have minimal impact when IMR and traditional approaches are compared. However, given that some studies have shown evidence of mode effects (as well as those cited above, see Fouladi, McCarthy, & Moller, 2002; Linnman, Carlbring, Åhman, Andersson, & Andersson, 2006[2]), this issue is clearly worthy of further investigation in order to clarify the types of effects that may occur and in which contexts.

The problem of reduced levels of researcher control over stimulus materials (e.g., Krantz, 2001), participation environment (e.g., Barbeite & Weiss, 2004), and participant behavior (e.g., Hewson, 2003) has been identified as a key issue in IMR. Different hardware and software configurations may easily cause stimulus display variability, and researchers can never be entirely sure that participants have followed instructions as directed. Although the potential impact of such uncontrolled variations should not be underestimated, the number of studies to date that have shown that IMR procedures can produce high-quality data indicates that, in many cases, these factors do not appear to pose a serious threat. It is also worth pointing out that IMR procedures have the potential to enhance data validity over traditional (pen-and-paper) methods by *increasing* levels of researcher control through incorporation of such procedures as response completeness checking and collection of *metadata* (e.g., completion patterns). Furthermore, it has been suggested that increased procedural variability in IMR studies allows for wider generalizability of results, if an effect is found, beyond what would normally be possible in a more controlled traditional setting (Reips, 2002).

Although the above suggests that, in many contexts, IMR procedures can lead to high-quality data, at least comparable to that obtained in more traditional settings, researchers should be wary about simply adapting a questionnaire or experimental procedure to an Internet environment and supposing that this will generate valid and reliable data. Validation studies of IMR procedures are essential in order to gain confidence in the data that can be acquired through this medium. The present study set out to contribute to this area by comparing Internet and traditional administrations of a widely used health beliefs questionnaire (Wallston & Wallston, 1981). In general, questionnaire and survey data gathered online have been shown to be as valid and reliable as those gathered offline (e.g., Anderson, Kaldo-Sandström, Ström, & Strömgren, 2003; Buchanan & Smith, 1999; Davis, 1999; Riva et al., 2003; Smith & Leigh, 1997; Stanton, 1998; Szabo et al., 1996; Voracek, Steiger, & Gindl, 2001). However, some studies have failed to show equivalence of Internet and non-Internet administrations of questionnaire-based research. Buchanan (2001) has reviewed validation studies in personality research, noting that those showing equivalence have tended to use *unidimensional* scales (e.g., Davis, 1999), whereas those that have failed to find equivalence (e.g., Johnson, 2000) have used *multidimensional* scales. Thus, the question of whether multidimensional scales can be validly implemented on the Internet remains.

The Present Study

The MHLC Scale (Wallston & Wallston, 1981) was administered via paper and Internet modes in order to compare the psychometric properties of the scale for each method.

The MHLC Scale was selected because it is a multidimensional scale (and support for the use of multidimensional scales in IMR so far is weak), has not yet been validated in an IMR context, and has been extensively used in health studies, so norms, scale reliabilities, and factor structures are available for comparison with the present data (Wallston & Wallston, 1981). The key aims of the study were to assess the robustness of the MHLC Scale in terms of its factor structure and reliability when administered via the Internet, to compare obtained MHLC scores with published norms, and to consider differences in sample composition resulting from Internet and non-Internet sampling approaches. Since the first aim of the study was to assess administration of the MHLC Scale in a context that has practical relevance to the way IMR actually is and can most usefully be implemented—that is, by using the Internet to recruit participants—and the second aim was to compare Internet-accessed and traditional samples, the sampling methods used were those most commonly employed in each context. Thus, the Internet sample was recruited via participation requests posted to a range of newsgroups, and the non-Internet sample was a convenience sample of undergraduate psychology students and members of the public.[3]

METHOD

Pilot Study

The pen-and-paper version of the MHLC Scale was administered to a convenience sample of 28 participants. The Internet version was placed on a Web server, and a participation request including the Web page URL was posted to the newsgroup ed.general. The latter procedure generated 12 responses over a period of 7 days. The only issues to emerge during this phase were a report from 1 Internet participant that the questionnaire failed to display correctly in the Netscape 6.2 browser and an ensuing discussion among newsgroup readers about the nature and quality of the study. The latter issue will be taken up further in the Discussion section as a potential threat to data validity in IMR.

Main Study

Participants. A non-Internet sample of 200 participants was obtained: 100 were students recruited during psychology lecture sessions, and the remainder were a convenience sample recruited from members of the public; they consisted primarily of friends, acquaintances, and work colleagues of the research associate who oversaw data collection.[4] Although perhaps not an optimal method of recruiting participants, such a method adds to the ecological validity of the study, since this approach to recruitment is commonly used throughout psychology. Age ranged from 18 to 50+ years, with 55 males and 144 females (1 nonresponse to this question). An Internet sample of 167 participants was obtained by posting to a range of newsgroups (freeserve.chat, freeserve.discuss, ie.general, ntl.talk, alt.sci.sociology, alt.history, alt.politics, alt.psychology, sci.psychology.misc, sci.psychology.theory, alt.psychology.help, alt.psychology.jung, alt.psychology.nlp, sci.psychology.psychotherapy, and sci.psychology.personality). Age ranged from 18 to 50+ years, with 91 males and 73 females (3 nonresponses to this question).

Materials. The MHLC Scale, Form A, was used for both Internet and paper administrations. This scale contains 3 six-item subscales: Internality, Powerful Others Externality, and Chance Externality. For each subscale, a higher score indicates a greater tendency to attribute one's own state of health to the factor indicated by the subscale name (i.e., Self, Powerful Others, and Chance). The range of possible scores for each subscale is 6–36. The Internet version of the scale was produced using hypertext markup language (HTML) to generate a form suitable for placing on a Web page. The form consisted of a number of questions designed to elicit demographic information, followed by the 18 six-point Likert-type MHLC Scale items relating to health beliefs. The form code incorporated embedded Javascript commands to measure browser type, IP address, time from loading page to submitting data, and date and time of completion. A Common Gateway Interface (CGI) script written in Perl was used to process incoming form data,[5] which were sent both to one researcher's e-mail account and to a file on the server. A debrief page (HTML document) was sent automatically by the CGI script after the participants had submitted their data. The paper version of the questionnaire was a printed copy of the HTML form and was, therefore, identical to the online version in layout.

Procedure. The non-Internet participants were handed the paper version of the scale and were asked to complete this in their own time (students were invited to participate at the start of lectures and completed and returned the scale at that time). After completing the questionnaire and returning it to the researcher, the participants were verbally debriefed. The Internet participants responded to a participation request posted to newsgroups, which gave them the address (URL) of the Web page that allowed them to access the online version of the questionnaire. The Internet participants were required to tick an informed consent box and then respond to the questions and statements, using a mixture of text boxes and radio buttons (the non-Internet participants were required to tick radio button boxes and write in text boxes). The online questionnaire did not fit onto one screen, so the participants were required to scroll down to answer all the questions. The paper version consisted of 2 one-sided pages stapled together. At the end of the online questionnaire was a *send data* button, which the participants clicked if they wished to send their data to the researcher; when this button was clicked, a thank you and debrief screen was displayed, and the participants were then directed to the homepage of the researchers' institution. In the paper version, instead of a send data button, there was a sentence thanking the participants for taking part in the study, and they were then verbally debriefed.

RESULTS

Data Screening

The Internet data were screened in order to detect multiple submissions. First, the data were examined for submissions within a few minutes of each other. Six such cases were found, and these were clearly multiple submissions, since the responses were identical, including browser type and version, IP address, and operating system. Second, remote (IP) addresses were checked for any duplicates (submissions from the same address). Seven were found. Six of these were the multiple submissions that had already been detected as having been submitted close in time. For the other, the two sets of responses from the same IP address were found to be very different (including demographic details), and both were maintained in the final data set. Although some authors have used the more conservative method of removing all but one data set coming from the same IP address (e.g., Birnbaum, 2001), it was felt in this case that such an approach was overcautious.

After removal of multiple submissions, completion times (measured as the time interval between the page's being loaded and the send data button's being pressed) ranged from 1.5 to 47 min, with a mean completion time of 4.5 min. Completion times were screened for *out-of-*

range responses (those that were 1.5 times the interquartile range above the upper quartile or below the lower quartile), and eight outliers were found in the upper range. The outliers were examined for anomalies, such as evidence of random responding or identical responses to all the items, and given that the data looked genuine, were retained. Since statistical analysis of completion times was not subsequently carried out, the lack of a normal distribution was not crucial here.

After multiple submissions had been removed, the final Internet data set consisted of 167 participants.

Psychometric Properties of the Scale

Reliability. To compare the internal consistency of the MHLC Scale across samples, Cronbach's alpha coefficients were computed for the three subscales. Table 1 shows that the coefficients for all three subscales were higher for the Internet data than for the paper data. For the Internet data, all the coefficients were equal to or above the .70 level considered to be the minimum acceptable (Kline, 1993) and were within or above the .67–.77 range achieved by Wallston and Wallston (1981) for the three subscales with their development samples. None of the coefficients for the paper data reached an acceptable level, although the Powerful Others coefficient was just within the range quoted by Wallston and Wallston. It can therefore be concluded that the internal consistency statistics for the Internet data were better than those for the paper data.

Factor structure. Using the maximum likelihood method of estimation,[6] EQS 5.7a (Bentler, 1995) was used to perform confirmatory factor analysis (CFA) on the Internet and paper data. Studies of the MHLC Scale's factor structure either have confirmed the three-subscale structure of the original instrument or, sometimes, have shown a two-subscale structure, with the two external subscales (Powerful Others and Chance) merging to form a single factor (Chaplin et al., 2001). In the present study, analyses specifying both types of structure were performed. However, for both samples, a three-subscale structure provided the best fit to the data, and therefore, only these analyses are considered here. It has also been noted that scores on the two external subscales exhibit a positive correlation of around .20 (Wallston & Wallston, 1981). Lagrange Multiplier tests subsequent to initial runs specifying independence of all three factors suggested a relationship between scores on the external subscales. This was true for runs on data for both samples. Therefore, the statistics reported below are for analyses that specified a correlation between the Powerful Others and the Chance factors.

In Table 2, the independence χ^2 statistic reflects the goodness of fit between the input covariance matrix and the matrix implied by a model assuming no relationships between variables. For both analyses, the present high values of the independence χ^2 statistic indicated a mismatch between input covariance matrices and models assuming no relationships between variables. It was therefore concluded that there was sufficient structure in both input matrices to make analyses meaningful.

Table 1
Internal Consistency Statistics (Cronbach's Alpha)

	Sample	
Subscale	Internet (*n* = 167)	Paper (*n* = 198)
Internality	.79	.63
Powerful others	.71	.68
Chance	.70	.60

Table 2
Fit Indices for the Internet and Paper Samples

	Sample	
Index	Internet (*n* = 167)	Paper (*n* = 198)
Independence χ^2	856.921 (153)	662.683 (153)
χ^2	248.633 (131)*	251.693 (131)*
χ^2/df	1.898	1.921
CFI	.833	.763
RMSEA	.074	.069
SRMR	.102	.079

Note—CFI, comparative fit index; RMSEA, root-mean square error of approximation; SRMR, standardized root-mean square residual. $^*p <$ .001 (*df* for χ^2 in parentheses).

When how well data fit a hypothesized factor structure is considered, many different fit indices are available; these indices have different advantages and disadvantages under different conditions, and it has become the convention to report more than one index. The rationale for inclusion of the specific indices reported in Table 2 was as follows. The χ^2 statistic was previously the most prominent goodness-of-fit statistic reported in studies using structural equation modeling and CFA. For these reasons and because the statistic is still commonly reported, χ^2 is reported here. A good fit is indicated by a nonsignificant value of χ^2. However, a number of problems have been identified with the χ^2 statistic. The nature of some of these problems depends on sample size; in particular, where sample sizes are relatively small (as in the present study), probability levels for evaluating the statistic can be inaccurate (Ullman, 2001). To some extent, problems with the χ^2 statistic are alleviated by considering the ratio of the value of χ^2 obtained to the degrees of freedom for the analysis, with a ratio lower than 2 suggesting a good fit (Ullman, 2001). Along with the above indices and those mentioned below, the comparative fit index (CFI) is reported in Table 2, since this gives a good estimate of fit for analyses with small sample sizes. This index takes values in the range 0 to 1, with values greater than .95 indicating a good fit (Ullman, 2001). Together with CFI, the root-mean square error of approximation (RMSEA) is currently the most commonly reported index (Ullman, 2001) and is considered to be the index of choice in the area of personality testing (Raykov, 1998). Values of RMSEA of .06 or lower are considered to be good, and values lower than .10 to be reasonable (Ullman, 2001). Finally, Ullman cites the 1999 work of Hu and Bentler as suggesting that the standardized root-mean square residual (SRMR) should be routinely reported. Again, this index takes values in the range 0 to 1, with values of .08 or lower indicating a good fit (Ullman, 2001).

Comparison of the fit indices for the Internet and the paper data in Table 2 gives a mixed message, as is often the case, and this provides another reason for reporting multiple fit indices (Ullman, 2001). For both analyses reported, χ^2 statistics were significant, indicating a poor fit, although the fit was marginally better for the Internet data. But given that the probabilities associated with these analyses may be problematic, it might be better to consider the ratio of χ^2 values to their degrees of freedom. This ratio was lower than 2 for both samples, indicating a good fit, with the ratio being lower (and therefore, better) for the Internet data. This observation obviously stems directly from the previously mentioned lower χ^2 for the Internet data. Although neither of the CFI indices indicated an adequate fit, again the fit for the Internet data was better than that for the paper data. Given the small sample sizes in both analyses, particular attention should be paid to this result. Both of the values of RMSEA were within the acceptable range, but in this case, the paper data exhibited a marginally better fit than did the Internet data. However, with small samples, RMSEA can be too large, and it is worth noting that the Internet sample was smaller than the non-Internet sample. This said, the value of SRMR was also better for the paper data than for the Internet data, with the Internet data exhibiting an inadequate fit in contrast to that for the paper data. To summarize, from overall consideration of the present fit indices, it seems that the very least that can be concluded is that the Internet data were no worse than the paper data at reproducing the factor structure of the MHLC Scale.

Comparability of MHLC Scores

Table 3 shows means and standard deviations for each subscale for the Internet and paper data.

Comparing the statistics in Table 3 with the norms reported in the literature (Wallston & Wallston, 1981), we find similarity. Wallston and Wallston reported norms based on analysis of various samples. Internality norms were 26.68 (student sample) and 25.55 (general adult sample), which is comparable to the values in Table 3. Powerful Others norms were 17.87 (students) and 19.16 (adults). The Internet means in the present study are somewhat lower than these norms; however, the studies reported in Wallston and Wallston displayed a range of scores, the lowest for the Powerful Others subscale being from a middle to upper class sample of parents with a mean score of 13.61 for females and 14.29 for males. Chance norms of 16.72 (students) and 16.21 (adults) were slightly lower than those obtained in our samples.

A 2 × 2 ANOVA was run for each subscale, with mode of administration (Internet or paper) and sex as factors and score as the dependent variable. Sex was included as a factor because it was confounded with mode of administration and, thus, was relevant in clarifying the nature of any differences observed between the Internet and the paper data. These analyses failed to show any effect of sample or sex on Internality and Chance scores. In these two analyses, for a medium effect size ($\eta^2 = .059$), power would be around .99 (in excess of the commonly recommended level of .80). However, the largest of the effect sizes in these analyses was an η^2 of .003 (as was the corresponding partial η^2). Post hoc analyses using the GPower package (Erdfelder, Faul, & Buchner, 1996) showed that such a small effect size would require a total sample size of in excess of 2,000 people to achieve .80 power. Powerful Others scores were found to be higher for the paper sample ($M = 16.38$) than for the Internet sample [$M = 14.95$; $F(1,355) = 9.940$, $p = .002$; η^2 and partial $\eta^2 = .027$] and higher for males ($M = 16.13$) than for females [$M = 15.45$; $F(1,355) = 4.26$, $p = .04$; η^2 and partial $\eta^2 = .012$]. No interactions were found.

Table 3
Mean Scores and Standard Deviations for Each of the MHLC Subscales for the Internet and Paper Data

	Sample					
	Internet			Paper		
Subscale	*M*	*SD*	*n*	*M*	*SD*	*n*
Internality	25.20	5.09	162	25.10	4.03	195
Powerful others	14.95	5.16	163	16.38	5.21	196
Chance	18.26	5.45	161	18.94	4.86	194

Table 4
Correlation Matrix for the Three Subscales, for Internet and Paper Data Separately

Subscale	Sample	Internality	Powerful Others	Chance
Internality	Paper		.052[a]	−.127[b]
	Internet		−.111[c]	−.346*[d]
Powerful others	Paper			.292*[b]
	Internet			.404*[e]

*$p < .01$ (two-tailed). [a]$df = 193$. [b]$df = 192$. [c]$df = 162$. [d]$df = 160$. [e]$df = 161$.

Wallston and Wallston (1981) have reported (on the basis of their sample of 115 respondents recruited from Nashville's municipal airport) independence of the Internality and Powerful Others scales, slight negative correlations between the Internality and Chance scales, and modest positive correlations between the Chance and Powerful Others scales (.20), so this was examined for both the Internet and the paper data here (see Table 4).

Table 4 shows that although our paper data closely matched the pattern reported by Wallston and Wallston (1981), the correlations for the Internet data tended to be larger. However, tests for differences within each pair of correlations showed that the only case in which the correlation for the Internet data was significantly greater than that for the paper data was for the Internality–Chance relationship ($z = 2.17$, $p = .030$, two-tailed). The tests for both the Internality and Powerful Others ($z = 1.52$, $p = .129$, two-tailed) and the Chance and Powerful Others ($z = -1.19$, $p = .234$, two-tailed) relationships were nonsignificant. Reference to Cohen (1988) shows that under the a priori assumption of medium effect sizes ($q = .30$) and a two-tailed alpha of .05, the power of these z tests would be around .80 (the commonly recommended level). However, in the presence of the currently observed small effect sizes ($q = .13$, .16, and .23 for the powerful others–chance, powerful others–internality, and chance–internality analyses, respectively), post hoc analyses showed that power was lower (in the regions of .24, .36, and .58, respectively).

Comparability of Samples

The Internet and non-Internet samples were compared on sex, age, salary, occupation, nationality, and qualifications. Table 5 displays for each sample the percentage of responses falling in each category of each demographic variable. The Internet and non-Internet samples differed significantly on all the demographic variables measured.

The Internet sample was more balanced in terms of sex than was the non-Internet sample, which was predominantly female (72%), and showed a more even distribution over age categories, where the non-Internet sample was skewed toward younger respondents. Internet participants showed greater representation in the higher earning categories than did non-Internet participants, had higher levels of formal educational attainment, and had higher representation in professional and information technology occupational categories. Whereas nearly all the non-Internet participants were from the U.K. (98%) and all were actually in the U.K. at the time of responding, the Internet sample consisted of just under half U.K. respondents and roughly one-quarter North American and one-fifth European respondents. Using the GPower package (Erdfelder et al., 1996) and referring to Cohen (1988), post hoc power analysis showed that in the presence of the large effect sizes shown in Table 5, and given the large sample sizes, power was very high ($>.99$) for all the χ^2 analyses.

DISCUSSION

Comparability of Internet and Paper Data

Scale properties. The present study has shown that it is possible to administer a multidimensional scale to participants over the Internet and obtain data comparable to that acquired using pen-and-paper methods. When the internal consistency of the MHLC Scales was considered, values for the Internet data were better than those for the paper data and reached acceptable levels. Also, although the results were mixed, the fit of the Internet data to the assumed three-factor structure of the MHLC Scale was no worse than that for the paper data. The failure of the data from both sources to fit the factor structure of the MHLC Scale for some of the indices reported (most notably, CFI) should not be overemphasized, since the results of (unreported) exploratory factor analyses forcing three-factor solutions for both the Internet and the paper data showed that both data sets produced a very close approximation to the assumed factor structure. Using the Lagrange Multiplier and Wald test statistics in the EQS output, it would have been possible to modify the models tested to produce better fit indices. However, since the aim of the study was to use the previously established factor structure of the MHLC Scale as a reference point from which to evaluate the adequacy of the factor structure implied by data acquired using Internet and paper versions of the scale, these modifications were not performed.

In contrast to the present findings, previous studies in which multidimensional scales have been used have failed to show support for the validity of those scales when the mode of administration was the Internet. One possible reason for the difficulties encountered in replicating findings for multidimensional instruments with IMR methods may simply be that the greater complexity of such instruments makes findings involving features such as their factor structure inherently more difficult to replicate—an issue that is not associated with Internet data collection methods per se. Nevertheless, in this study, we did find good comparability of Internet and paper data for the MHLC Scale, thus showing that it is possible to successfully administer a multidimensional scale online.

MHLC scores. MHLC scores for the Internet and paper data were very similar for the Internality and Chance scales and showed good comparability with reported norms. They did differ significantly on the Powerful Others subscale, however, with the Internet scores being lower than the reported norms. However, the mean Powerful Others score for the Internet data did fall within the range of scores reported across different studies (Wallston & Wallston, 1981) and was still higher than the lowest scores, reported for a sample of middle to upper class parents. Given that the Internet sample displayed higher educational and income levels than did the paper sample, the most parsimonious account for this result is that the Internet sample contained a larger proportion of higher socioeconomic status respondents and, thus, obtained Powerful Others scores more in line with those of Wallston and Wallston's middle to upper class sample. This explanation would seem the most plausible; however, given that the Internet and the paper samples also differed on age and nationality, these factors cannot be ruled out as potentially having impacted upon Powerful Others scores, and neither can the possibility that mode effects (i.e., procedural differences between Internet and paper administrations) were instrumental in causing differences in the Powerful Others scores (e.g., by differentially affecting levels of socially desirable responding). However, the latter account is less plausible, given the specificity of the difference to the Powerful Others subscale.

The finding that our Internet sample obtained scores that, overall, are consistent with preestablished norms reported in the literature (derived from pen-and-paper administrations) is interesting with respect to the issue of whether established scale norms can be considered applicable to Internet-gathered data. Buchanan (2001) has highlighted this issue, pointing out that if the nature of the Internet medium can affect participants' responses—for example, by encouraging more candid answers—this has

Table 5
Percentage of Respondents in Each Category of Each Demographic Variable, for Internet and Paper Samples Separately

	Sample					
	Internet	Paper	χ^2	*df*	*p*	Effect Size (*w*)
Sex						
N	164	199	33.99	3	<.0005	.306
Male	55.5	27.6				
Female	44.5	72.4				
Age						
N	167	200	116.10	12	<.0005	.562
18–25	24.6	50.0				
26–33	22.8	16.0				
34–41	16.8	13.5				
42–49	23.4	9.0				
50+	12.6	11.5				
Salary (£)						
N	146	143	61.19	9	<.0005	.460
<15k	42.5	75.5				
16–25k	26.7	14.7				
26–35k	11.6	8.4				
>35k	19.2	1.4				
Occupation						
N	162	198	268.48	18	<.0005	.864
Professional	29.0	14.1				
IT	13.6	0.5				
Skilled	13.0	18.2				
Unskilled	7.4	8.6				
Student	24.7	53.5				
Retired	3.7	1.0				
Nationality						
N	157	198	196.72	9	<.0005	.744
U.K.	43.0	98.0				
Europe	21.7	1.5				
North America	26.1	0.0				
Other	8.9	0.5				
Qualifications						
N	159	184	105.14	9	<.0005	.554
Postgraduate	21.4	2.7				
Degree	25.2	13.0				
Further education	38.4	75.5				
School	15.1	8.7				

implications for the extent to which Internet data can be interpreted in relation to such norms. In the present study, Internet participants' scores were found to be comparable to established norms (even though these norms are based on data from several decades ago).

The observed higher intercorrelations between the MHLC subscales for the Internet data were also of interest; the paper data very closely matched the pattern reported by Wallston and Wallston (1981). There appear to be two possible interpretations of this finding. If one considers the higher reliabilities of the Internet data, it may be suggested that the Internet participants in this study were, in general, more committed and genuine in their responses or, at least, understood the questions better. If this was the case, we may consider the Internet data to be more reliable (hence, higher levels of internal consistency) and valid than the paper data and the higher correlations between subscales observed for this sample to be an indication that these subscales are not as independent as has been previously thought on the basis of results from non-Internet samples. On the other hand, we may interpret this result as indicating that the non-Internet data have greater validity because they fit more closely with preestablished results in the literature. It is not clear to us which of these explanations is most plausible, and further studies will be required to help clarify this. In any event, although differences in effect size were observed in the present study, this pattern should not be overemphasized, given that the observed difference was significant for only one of the three subscales.

Generalizability of Findings

The implications of the present findings for Internet-based research more generally will depend on the extent to which our results can be generalized to future IMR studies. Along with previous authors (e.g., Buchanan, 2001; Krantz et al., 1997), we consider it unwarranted to generalize from any IMR study beyond the particular test instruments (or experimental procedures) used therein. However, it is reasonable to predict that our results will generalize to future administrations of the MHLC Scale via the Internet, using similar presentation formats and sampling procedures, as will now be discussed.

Earlier, we outlined several sources of difficult-to-control (or in most cases, impossible-to-control) variation in IMR studies: presentation format, participation context, and participant behavior. Although all these sources of variation no doubt occurred in our study, at least to some extent, the data we obtained were shown to be valid and reliable. This suggests that in the present research context at least, these variations do not significantly affect the quality of the data obtainable. This adds to the many previous studies showing similar results across other research contexts. Some authors have suggested that the potential for increased variability in IMR studies may cause IMR data to be more *noisy* than traditional data (e.g., Reips, 2002). However, we found little support for this. The greatest difference between the standard deviations of MHLC scores across our two samples occurred for the Internality subscale, with Internet participants showing slightly greater variability in scores ($SD = 5.09$) than did non-Internet participants ($SD = 4.03$). The IMR literature is inconsistent on this issue, with some studies reporting greater variability for Internet data (e.g., Shavit et al., 2001) or for non-Internet data (e.g., Davis, 1999), and other studies showing no difference (e.g., Metzger et al., 2003). No doubt, a number of factors will interact to influence such results, including procedural variations, sample characteristic variations, and the sensitivity of any particular measurement of interest to these sources of variation.

A crucial consideration with respect to the implications of the present study for future IMR administrations of the MHLC Scale is the extent to which our results will generalize to future Internet (and indeed, non-Internet) samples. We will not address the issue of the generalizability of results from traditional psychology student samples here, although this issue has been raised by previous authors (e.g., Smart, 1966). We *are* concerned, however, with the extent to which our data may feasibly generalize to other Internet samples. As has been discussed, Internet samples accessed via advertisements posted online tend to display similar characteristics, at least in the way they compare with traditional undergraduate student samples. One aim of the present study was to explore this further by comparing the characteristics of Internet and traditional samples. Our results confirm previous findings: The Internet sample was more balanced in gender, more diverse in age and nationality, and more highly educated than was the non-Internet sample. Furthermore, Internet participants had higher earnings and were more likely to be in professional occupations. Nevertheless, the psychometric properties of the two data sets were largely equivalent, and both showed comparability with established norms. It thus seems reasonable to suggest that the present results will most likely generalize to future Internet samples acquired using similar sampling procedures—first, because it is likely that these procedures will generate samples similar to ours, and second, because the MHLC Scale has been shown (here and elsewhere; see Wallston & Wallston, 1981) to be relatively robust in response to variations in sample composition. This said, one can never be sure of the type of sample that will be generated by any particular IMR sampling procedure, and researchers should be aware of this. Fortunately, for many areas of psychological research, this will not prove to be a major problem: Psychological research is often concerned with measures that are presumed to remain relatively invariant across different demographic populations (hence, the general lack of concern about the widespread development of hypotheses and theories based on undergraduate student populations). For other disciplines, however, obtaining more broadly *representative* samples will be more important. The issue of how to obtain large representative samples on the Internet is an ongoing research topic. Also, although the present scale may reasonably be expected to generate valid and reliable data when administered online to future Internet samples, it is less clear that score distributions will remain comparable across samples that differ on key relevant variables.

The finding that almost half of our Internet sample came from the U.K. (and around a quarter from the U.S.) contrasts with previous studies, which have typically reported Internet samples as overwhelmingly from the U.S. (Krantz & Dalal, 2000). An obvious explanation for this is that the newsgroups we targeted had a higher U.K. readership than did those in previous studies. Recent estimates of worldwide Internet access (NUA, 2002) suggest that around 28% of the entire 600 million Internet users come from the U.S. and around 6% from the U.K. (in each case, approximately 50% of the national population). Although these statistics are at best approximations, it is quite apparent that Internet-accessed samples that consist almost entirely of U.S. or U.K. respondents must reflect biased sampling procedures with respect to the entire Internet user population, a major contributing factor no doubt being that most IMR studies are conducted in English. A further possible explanation for geographical differences between our samples and previous samples concerns changing trends in the availability of cheap Internet access. Unlimited access for a fixed fee has been widely available in the U.S. for a while, whereas in the U.K. users have until fairly recently had to pay for online time by the minute. With *broadband* now more widely available, a larger number of U.K. users have high-speed permanent online access for a fixed monthly fee and, hence, may be more likely to spend time online to complete a study. Thus, although there is emerging evidence that similar sampling procedures in IMR will often generate samples with similar characteristics, small variations in these procedures, as well as changes in Internet usage patterns, can clearly influence the types of samples obtained, and researchers should be aware of these influences.

A possible criticism of the present study is that because mode of administration and sample type are confounded, we cannot unequivocally attribute any differences found to either one of these factors (Epstein et al., 2001). However, given the overall comparability of our Internet and paper data, this issue does not pose a major problem in interpreting our results. Indeed, it allows generalization beyond what would have been possible had we used equivalent samples. However, the Internet and the paper data were found to differ on Powerful Others scores. Although a plausible account, consistent with previous research, has been offered that explains this result as due to sample composition, a possible mode effect cannot be ruled out. Other authors have eliminated such potential ambiguity by conducting IMR validation studies that control for sample variation (e.g., Cronk & West, 2002; Epstein et al., 2001) and even for both sample variation and participation context: Salgado and Moscoso (2003) reported that in their "Internet" condition (in a repeated measures design), participants completed the study in groups of 19 (presumably in the laboratory). It is not clear to us that such approaches, which maintain tight control over the types of variables (stimulus display, participation context, sample composition, etc.) that have been highlighted as most problematic in IMR, have much relevance for Internet-mediated research procedures. In the most extreme cases, these studies would appear to be generalizable to procedures that use undergraduates in a laboratory setting.

However, the appeal of Internet-mediated research is surely to open up possibilities for moving beyond these traditional contexts. Indeed, this point has been recognized by researchers who have used designs that maintain sample equivalence; thus, Epstein et al. (2001) state the following:

> Although the chosen procedure of recruiting participants from a single location and randomly assigning them to experimental conditions allowed us to control for selection biases, this method of completing a survey over the Internet is not necessarily representative of how the Internet is used on a day-to-day basis. (p. 345)

Similarly, Metzger et al. (2003), who found no mode effect in a face recognition task using an equivalent samples design, comment that

> future studies should focus on populations other than college students. It would be advantageous to compare a group of non-student participants who complete these experiments online and compare them to a group of college students participating in the traditional laboratory setting. This will allow one to determine if WWW data (since it will come from a variety of individuals) is truly comparable to data collected in the college laboratory. (p. 620)

It is our view that using Internet-accessed samples in IMR validation studies will, in most cases, lead to results that are more generalizable and relevant to the types of IMR procedures that social and behavioral researchers would want to use.

In the discussion above, we have argued that the results of our study can reasonably be considered generalizable to IMR studies in which similar sampling procedures and presentation formats are used. Given that a key appeal of Internet-mediated research is the sampling opportunities it affords (Musch & Reips, 2000, actually found this to be one of the two most important reasons for conducting an IMR study, along with enhanced statistical power, in a survey of Internet researchers), the finding of equivalence despite differences in both mode of administration and sample composition is an important one.

In the present study, a scale that was originally developed and validated using U.S. samples was shown to display comparable psychometric properties and norms when administered to a more nationally diverse Internet sample. The finding that our non-Internet data produced reliabilities that were barely acceptable could potentially be attributed to the non-Internet sample's consisting almost entirely of U.K. respondents; however, this is a tentative suggestion that requires further investigation.

Rate of Response and Sample Size

Given the number of newsgroups to which the participation request was posted (15 in total, each with a follow-up posting within 1 week) and the period of data collection (approximately 100 days), the number of responses obtained was lower than expected. A number of reasons why responses were lower than those reported in previous studies may be offered. First, the prevalence of Internet studies is now much higher than previously, and therefore, obtaining participants for any one of the many studies being advertised may be more difficult, due to both the greater competition for people's time and the reduced novelty value of completing an online questionnaire. Second, newsgroups with a large number of postings were targeted; Buckley and Vogel (2003) have pointed out that this may be an ineffective strategy for generating large samples, since participation requests are more likely to go unnoticed among large volumes of postings, and posting to newsgroups with a smaller number of daily postings may be more effective, since these can have a very large number of "silent readers." Another possibility is that issue salience was low in the present study. Studies that have reported very high numbers of responses have often selectively posted to newsgroups and mailing lists for which the research topic is likely to be of particular interest (e.g., Birnbaum, 2001).

Although the size of the Internet sample in the present study was smaller than that in many previous studies, this is not necessarily a problem. In fact, the great disparity of the Internet and the paper samples in some previous studies in which CFA has been used might pose a problem. For example, Buchanan and Smith (1999) obtained sample sizes of 963 and 224 for their Internet and paper samples, respectively. In addition to comparing the adequacy of fit indices with those obtained by other authors, as in the present study, Buchanan and Smith compared the adequacy of fit indices for CFAs involving the two data sources and concluded that values of the goodness-of-fit index (GFI), the adjusted goodness-of-fit index (AGFI), and the Bentler–Bonnett normed fit index (NFI) for their Internet data were superior to those for their paper data. However, values of the GFI and AGFI tend to increase with sample size and might underestimate fit for small sample sizes (Bollen, 1990), and small sample sizes might also lead to underestimation of fit for the NFI (Ullman, 2001). Therefore, conclusions with respect to these indices were confounded with sample size in Buchanan and Smith's study. Of the other indices that Buchanan and Smith reported, two showed better fit for the paper data (χ^2 and the Tucker–Lewis coefficient), and one (the root-mean square residual) showed a marginally better fit for the Internet data. Although the Internet sample was smaller in the present study than that of Buchanan and Smith, the greater equivalence of sample sizes (a ratio of 1.19:1 in favor of the non-Internet sample) was less likely to lead to erroneous conclusions based on sample size effects. It is also worth noting that, if anything, the slightly greater non-Internet sample size in the present study would lead to more conservative conclusions concerning the adequacy of the Internet data for fit indices that can be influenced by sample size, such as RMSEA.

The observations above illustrate the point that the large sample sizes obtainable using the Internet as a medium for data collection can have drawbacks in CFA studies if these data sets are compared with the typically smaller sample sizes obtained from paper administrations. To avoid these problems, where Internet sample sizes are far greater than non-Internet sample sizes, it is reasonable to recommend either that studies making these types of comparison should randomly select a smaller subsample of the Internet data obtained or that fit indices that are not prone to sample size effects should be emphasized in reaching conclusions.

Further Issues

A potential threat to data validity emerged during the pilot study. A few days after the participation request was posted, a thread ensued in which newsgroup subscribers commented on the validity of the study, offered suggestions about its aims, criticized the quality of the MHLC Scale, and posted information contained in the debrief page. In this case, these posts appeared after the majority of responses had been received and did not contain any crucial information likely to have invalidated the participants' responses. However, this may often not be the case, and in addition to monitoring newsgroup postings subsequent to advertising a study, it may also be worthwhile to take measures to include, in the initial posting, a request to participants not to discuss the study within the newsgroup until data collection is complete. Interestingly, this problem did not emerge for any of the other 15 newsgroups that advertisements were sent to.

Although we found no major differences between Internet and paper administrations in terms of response completeness, other authors have reported such effects, with Internet data typically being superior to pen-and-paper data in this respect (e.g., Stanton, 1998; Truell, Bartlett, & Alexander, 2002). In our study, overall, the number of missing responses to questionnaire items was low. The only case in which a large difference was observed was for salary, with 28.5% of non-Internet participants failing to respond to this item, as compared with 12.6% of Internet participants. This can be attributed to the larger number of students in the non-Internet sample. Of the 18 MHLC Scale items, the mean numbers of items completed were 17.96 and 17.88 for the Internet and non-Internet samples, respectively. Incorporating a response completeness check that is activated when a respondent attempts to send data is relatively easy (e.g., using Javascript) and can help ensure that all questions are answered by prompting participants to go back and complete unanswered questions. This presents one advantage of Internet over pen-and-paper administrations. If ethical worries are raised by this procedure, prompts can inform the user that although they are under no obligation to provide the missing data, they are being prompted to do so in case they have omitted the response in error. Of course, respondents should be free both to submit partial data and to withdraw at any point.

Conclusions

To summarize, the present study showed comparable results when the MHLC Scale was administered via traditional and Internet modes, thus lending further support to the growing body of literature that demonstrates comparability of Internet and non-Internet data, despite the differences in sample composition typically found using these approaches. The results suggest that using the Internet to recruit participants and administer the MHLC Scale is likely to produce as useful data as do methods more commonly used in psychology. This opens the way for using the scale in this manner (a copy of the HTML version of the scale can be obtained by e-mailing the first author). Although the growing number of successful validation studies to date makes it tempting to predict that comparability of Internet and traditional methods will pertain across a number of other psychological domains, it would be premature to make generalizations about the robustness of psychological measures in response to the types of variations found in Internet and non-Internet samples and procedures beyond those for which this has been empirically verified. Still, the results to date are very encouraging. Although previous studies have tended to show positive support for the use of unidimensional scales, but not multidimensional scales, in IMR (Buchanan, 2001), the present study supports administration of the MHLC Scale and thus makes a useful contribution to this literature.

REFERENCES

Anderson, G., Kaldo-Sandström, V., Ström, L., & Strömgren, T. (2003). Internet administration of the Hospital Anxiety and Depression Scale in a sample of tinnitus patients. *Journal of Psychosomatic Research*, **55**, 259-262.

Bailey, R., Foote, W., & Throckmorton, B. (2000). Human sexual behavior: A comparison of college and Internet surveys. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 146-168). San Diego: Academic Press.

Barbeite, F. G., & Weiss, E. M. (2004). Computer self-efficacy and anxiety scales for an Internet sample: Testing measurement equivalence of existing measures and development of new scales. *Computers in Human Behavior*, **20**, 1-15.

Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.

Best, S. J., Krueger, B., Hubbard, C., & Smith, A. (2001). An assessment of the generalizability of Internet surveys. *Social Science Computer Review*, **19**, 131-145.

Birnbaum, M. H. (1999). Testing critical properties of decision-making on the Internet. *Psychological Science*, **10**, 399-407.

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

Birnbaum, M. H. (2001). A Web-based program of research and decision making. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 23-55). Lengerich, Germany: Pabst.

Birnbaum, M. H. (2002). Wahrscheinlichkeitslehren. In D. Janetzko, M. Hildebrandt, & H. A. Meyer (Eds.), *Das experimentalpsychologische Praktikum im Labor und WWW* (pp. 141-151). Göttingen: Hogrefe. English translation retrieved August 24, 2004, from http://psych.fullerton.edu/mbirnbaum/papers/probLearn5.doc.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology*, **55**, 803-832.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin*, **107**, 256-259.

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

Browndyke, J. N., Santa Maria, M. P., Pinkston, J., & Gouvier, W. (1998). *A survey of general head injury and prevention knowledge between professionals and non-professionals*. Retrieved August 12, 2004, from www.premier.net/%7Ecogito/project/onp1_poster.html.

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

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

Buckley, M., & Vogel, C. (2003, November). *Improving Internet research methods: A Web laboratory*. Paper presented at IADIS International Conference WWW/Internet, Algarve, Portugal.

Chaplin, W. F., Davidson, K., Sparrow, V., Stuhr, J., van Roosmalen, E., & Wallston, K. A. (2001). A structural evaluation of the expanded Multidimensional Health Locus of Control Scale with a diverse sample of Caucasian/European, native and black Canadian women. *Journal of Health Psychology*, **6**, 447-455.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Coomber, R. (1997). Using the Internet for survey research. *Sociological Research Online*, **2**(2). Retrieved April 1, 2004, from http://www.socresonline.org.uk/2/2/2.htm.

Corley, M., & Scheepers, C. (2002). Syntactic priming in English sentence production: Categorical and latency evidence from an Internet-based study. *Psychonomic Bulletin & Review*, **9**, 126-131.

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.

Eichstaedt, J. (2002). Measuring differences in preactivation on the Internet: The content category superiority effect. *Experimental Psychology*, **49**, 283-291.

Epstein, J., Klinkenberg, W. D., Wiley, D., & McKinley, L. (2001). Insuring sample equivalence across Internet and paper-and-pencil assessments. *Computers in Human Behavior*, **17**, 339-346.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, **28**, 1-11.

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

Herrero, J., & Meneses, J. (in press). Short Web-based versions of the perceived stress (PSS) and Center for Epidemiological Studies–Depression (CESD) Scales: A comparison to pencil and paper responses among Internet users. *Computers in Human Behavior.*

Hewson, C. [M.] (2003). Conducting psychological research on the Internet. *The Psychologist*, **16**, 290-292.

Hewson, C. M., Laurent, D., & Vogel, C. M. (1996). Proper methodologies for psychological and sociological studies conducted via the Internet. *Behavior Research Methods, Instruments, & Computers*, **32**, 186-191.

Hewson, C. M., Yule, P., Laurent, D., & Vogel, C. M. (2003). *Internet research methods: A practical guide for the social and behavioural sciences*. London: Sage.

Huang, H.-M. (2006). Do print and Web surveys provide the same results? *Computers in Human Behavior*, **22**, 334-350.

Im, E.-O., & Chee, W. (2004). Issues in an Internet survey among midlife Asian women. *Health Care for Women International*, **25**, 150-164.

ISC (2004). *Internet software consortium, Internet domain survey*. Retrieved August 10, 2004, from http://www.isc.org/index.pl?/ops/ds/.

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

Joinson, A. N. (2001). Knowing me, knowing you: Reciprocal self-disclosure in Internet-based surveys. *CyberPsychology & Behavior*, **4**, 587-591.

Kaye, B. K., & Johnson, T. J. (1999). Taming the cyber frontier: Techniques for improving online surveys. *Social Science Computer Review*, **17**, 323-337.

Kline, P. (1993). *Personality*. London: Routledge.

Knapp, H., & Kirk, S. A. (2003). Using pencil and paper, Internet and touch-tone phones for self-administration surveys: Does methodology matter? *Computers in Human Behavior*, **19**, 117-134.

Krantz, J. H. (2001). Stimulus delivery on the Web: What can be presented when calibration isn't possible. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 113-130). Lengerich, Germany: Pabst.

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

Krantz, J. H., & Dalal, R. (2000). Validity of Web-based psychological research. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 35-60). San Diego: Academic Press.

Laugwitz, B. (2001). A Web experiment on colour harmony principles applied to computer user interface design. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 131-145). Lengerich, Germany: Pabst.

Linnman, C., Carlbring, P., Åhman, Å., Andersson, H., & Andersson, G. (2006). The Stroop effect on the Internet. *Computers in Human Behavior*, **22**, 448-455.

Mann, C., & Stewart, F. (2000). *Internet communication and qualitative research: A handbook for researching online*. London: Sage.

Metzger, M. M., Kristof, V. L., & Yoest, D. J., Jr. (2003). The world wide web and the laboratory: A comparison using face recognition. *CyberPsychology & Behavior*, **6**, 613-621.

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.

Musch, J., & Reips, U.-D. (2000). A brief history of Web experimenting. In M. H. Birnbaum (Ed.), *Psychological experiments on the Internet* (pp. 61-87). San Diego: Academic Press.

NUA (2002). *NUA Internet surveys: How many online?* Retrieved August 10, 2004, from www.nua.com/surveys/how-many-online/index.html.

Pohl, R. F., Bender, M., & Lachmann, G. (2002). Hindsight bias around the world. *Experimental Psychology*, **49**, 270-282.

Raykov, T. (1998). On the use of confirmatory factor analysis in personality research. *Personality & Individual Differences*, **24**, 291-293.

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

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology*, **49**, 243-256.

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of the Internet in psychology research: Comparison of online and offline questionnaires. *CyberPsychology & Behavior*, **6**, 73-80.

Salgado, J. F., & Moscoso, S. (2003). Internet-based personality testing: Equivalence of measures and assessees' perceptions and reactions. *International Journal of Selection & Assessment*, **11**, 194-205.

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

Senior, C., Barnes, J., Jenkins, R., Landau, S., Phillips, M. L., & David, A. S. (1999). Attribution of social dominance and maleness to schematic faces. *Social Behavior & Personality*, **27**, 331-338.

Shavit, T., Sonsino, D., & Benzion, U. (2001). A comparative study of lotteries: Evaluation in class and on the Web. *Journal of Economic Psychology*, **22**, 483-491.

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, & Computers*, **29**, 496-505.

Smither, J. W., Walker, A. G., & Yap, M. K. T. (2004). An examination of the equivalence of Web-based versus paper-and-pencil upward feedback ratings: Rater- and ratee-level analyses. *Educational & Psychological Measurement*, **64**, 40-61.

Stanton, J. M. (1998). An empirical assessment of data collection using the Internet. *Personnel Psychology*, **51**, 709-725.

Szabo, A., Frenkl, R., & Caputo, A. (1996). Deprivation feelings, anxiety, and commitments in various forms of physical activity: A cross-sectional study on the Internet. *Psychologia*, **39**, 223-230.

Truell, A. D., Bartlett, J. E., II, & Alexander, M. W. (2002). Response rate, speed, and completeness: A comparison of Internet-based and mail surveys. *Behavior Research Methods, Instruments, & Computers*, **34**, 46-49.

Ullman, J. B. (2001). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (4th ed.). Boston: Allyn & Bacon.

Voracek, M., Steiger, S., & Gindl, A. (2001). Online replication of evolutionary psychology evidence: Sex differences in sexual jealousy in imagined scenarios of mates' sexual vs. emotional infidelity. In U.-D. Reips & M. Bosnjak (Eds.), *Dimensions of Internet science* (pp. 91-112). Lengerich, Germany: Pabst.

Wallston, K. A., & Wallston, B. S. (1981). Health locus of control scales. In H. M. Lefcourt (Ed.), *Research with the locus of control construct*: *Vol. 1. Assessment methods* (pp. 189-243). New York: Academic Press.

NOTES

1. We do not include studies in which university contacts, such as heads of departments, have been used to forward e-mail participation requests to students at their institution (e.g., Metzger, Kristof, & Yost, 2003; Pohl, Bender, & Lachmann, 2002). This method, not surprisingly, tends to generate samples equivalent to traditional undergraduate student samples (e.g., Metzger et al., 2003).

2. It is worth noting that in both these studies, the authors conclude, overall, that their data suggested that the Internet is a viable data collection tool within that area, pointing out either that observed effect sizes were very small and/or that, on the whole, their results replicated established psychological measures and effects.

3. We recognize that this approach could be criticized as confounding mode of administration and sample composition. However, although the practicable alternative of administering the Internet and pen-and-paper versions of the scale to two equivalent (e.g., undergraduate) samples would have allowed a direct test of mode effects, it would not have succeeded in meeting the present aim of examining the effects of IMR and traditional sampling procedures on sample composition. Furthermore, such a design would have limited implications for the validity of IMR procedures (as elaborated in the Discussion section).

4. Student and general population groups did not differ in their composition with respect to sex and nationality. Neither did scores on the MHLC subscales differ for these two groups. However, relative to the general population, as would be expected, the students were younger, earned less money, and were more likely to have experienced further education but, since they were still in higher education, were less likely to have obtained a degree.

5. The authors thank Mark Williamson, who provided the original CGI script, which we adapted for use in this study.

6. Ullman (2001) notes that studies showing that other methods of estimation, such as generalized least squares (GLS), often perform better with low sample sizes. However, for the present data, GLS solutions led to worse fit statistics than did maximum likelihood solutions.

(Manuscript received November 5, 2003;
revision accepted for publication January 5, 2005.)