Research Paper ■

Randomized Testing of Alternative Survey Formats Using Anonymous Volunteers on the World Wide Web

DOUGLAS S. BELL, MD, PHD, CAROL M. MANGIONE, MD, MSPH, CHARLES E. KAHN, JR., MD

Abstract Consenting visitors to a health survey Web site were randomly assigned to a "matrix" presentation or an "expanded" presentation of survey response options. Among 4,208 visitors to the site over 3 months, 1,615 (38 percent) participated by giving consent and completing the survey. During a pre-trial period, when consent was not required, 914 of 1,667 visitors (55 percent) participated (odds ratio 1.9, $P < 0.0001$). Mean response times were 5.07 minutes for the matrix format and 5.22 minutes for the expanded format ($P = 0.16$). Neither health status scores nor alpha reliability coefficients were substantially influenced by the survey format, but health status scores varied with age and gender as expected from U.S. population norms. In conclusion, presenting response options in a matrix format may not substantially speed survey completion. This study demonstrates a method for rapidly evaluating interface design alternatives using anonymous Web volunteers who have provided informed consent.

■ **J Am Med Inform Assoc.** 2001;8:616–620.

Affiliations of the authors: UCLA School of Medicine, Los Angeles, California, and RAND Health, Santa Monica, California (DSB, CMM); Medical College of Wisconsin, Milwaukee, Wisconsin (CEK).

This work was supported in part by grant G08 LM05705 from the National Library of Medicine, for IAIMS planning at the Medical College of Wisconsin. Dr. Bell was supported by the Mary and Irving Lazar Program in Health Services Research and by National Research Service award T32 PE19001-09 from the Health Resources and Services Administration of the U.S. Department of Health and Human Services. Dr. Mangione was supported by Generalist Faculty Scholar Award 029250 from the Robert Wood Johnson Foundation.

Preliminary results of this study were presented at the Third Annual National Research Service Award (NRSA) Trainees Research Conference, cosponsored by Agency for Health Care Policy and Research (AHCPR) and Health Resources and Services Administration (HRSA), June 13-14, 1997, in Chicago, Illinois.

Correspondence and reprints: Douglas S. Bell, MD, PhD, UCLA Division of General Internal Medicine and Health Services Research, 911 Broxton Plaza, Room 201, Los Angeles, CA 90095-1736; e-mail: <dbell@ucla.edu>.

Received for publication: 2/28/01; accepted for publication: 5/31/01.

The World Wide Web provides a new medium for health survey research, offering lower costs,[1] greater accuracy, and faster study completion than traditional surveys, through the elimination of survey distribution and processing steps.[2] A few pioneering studies have used the Web to recruit and survey patients with particular diseases, including inflammatory bowel disease,[3–5] atopic eczema,[6] and benign prostatic hyperplasia.[7] Web-based surveying may introduce problems, however, if unfamiliar formats introduce usability problems or cause questions to be misinterpreted.[2] No studies have directly tested the effects of Web-based survey formats on the usability of health surveys or on the reliability and validity of the data collected.

From June 1995 to June 1996, we recruited 4,876 participants to take the SF-36 health survey[8] using Web-based forms.[9] Users could select from two survey formats. One format used HTML tables to present

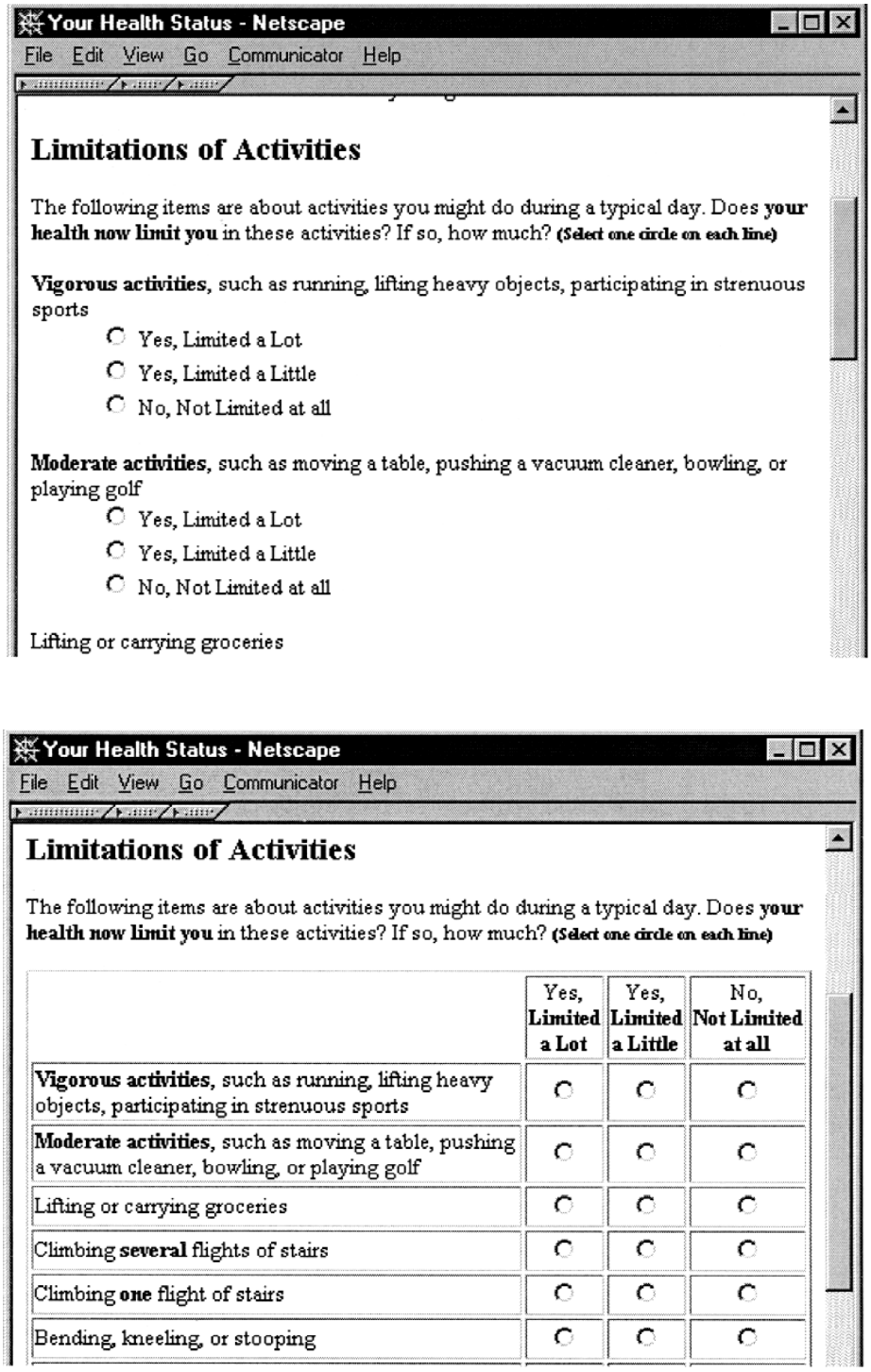

Figure 1 Survey formats compared in the experiment. *Top*, The "expanded" survey format. *Bottom*, The "matrix" survey format, which was approximately half as long in the vertical dimension.

response options in a matrix, and the other, intended for browsers that could not handle HTML tables, repeated the response options below each question (Figure 1). We found that users selecting the matrix format completed the SF-36 in 13 percent less time.[9] That difference, however, could have been due to self-selection rather than to the format itself. By randomizing participants to alternative formats, the current study aimed to determine the true effect of a matrix presentation of response options on users' speed and also on the internal consistency reliability and the known-groups validity of their responses.

Table 1 ■

Effects of Age, Gender, and WebTV Use on Mean Response Times and Health Status Scores

Subgroup	Response Time (min.)	PCS		MCS	
		Web Users	Norms	Web Users	Norms
Age (years):					
18–24	4.75	52.6	53.4	40.8	49.1
25–34	4.99	52.3	53.7	43.2	48.6
35–44	5.02	51.1	52.2	47.0	49.9
45–54	5.42	49.6	49.6	48.7	50.5
55–64	6.10	48.6	45.9	48.2	51.1
65+	7.60	44.6	41.3	52.1	49.1
Gender:					
Male	5.35	52.1	51.1	46.0	50.7
Female	4.98	50.6	49.1	44.2	49.3
WebTV user:	6.03				
Non-WebTV browser	5.06				

NOTES: Values shown are unadjusted means for each population subgroup. Web users are the 1,464 trial participants who met inclusion criteria. Norms are from administration of the SF-36 to a nationally representative sample.[8,12] The physical component summary (PCS) and mental component summary (MCS) health status summary scores are by definition normalized, so that a score of 50 is equivalent to the general population mean and a difference of 10 points represents 1 standard deviation in the general population.

Methods

We constructed a Web site for testing alternative formats for the SF-36 health survey.[8] Software for the Web site was written in Perl, using the Common Gateway Interface (CGI) to an HTTP server. Users first viewed a page that explained potential risks and asked for informed consent. Those who accepted the terms of participation were randomly assigned, using Perl's random number generator, to receive the SF-36 survey in one of two formats, which are shown in Figure 1. In the "expanded" format, each question was followed by a list of its response options, even if questions in a series all used the same response options. In the "matrix" format, when a series of questions shared the same response options, they were grouped together in a table that had a row for each question and a column for each response option.

All users were instructed to complete the form without interruption. Enrollment data were collected on the consent page. Users were required to provide their age and gender, and they could provide their race or ethnicity if they chose. Users were also required to indicate whether they planned to provide an honest self-report, to answer on behalf of another person, or to "just test" the system with different combinations of answers. On submitting a completed survey, users received their SF-36 scores.

The randomized trial was conducted from Feb 2, 1997 to Apr 30, 1997. To promote the study, we submitted our URL to five Web indexing sites (Yahoo, AltaVista, Lycos, Excite, and Infoseek). Participants were not directly recruited or contacted by the investigators. The Human Research Review Committee of the Medical College of Wisconsin approved the study protocol.

Participation rates were calculated for the 3-month randomized trial and also for the period from Dec 1, 1996 to Feb 1, 1997, when the same Web site administered the SF-36 with an introduction page that was one third as long and did not involve informed consent.[9] Submitted surveys were excluded from analysis if SF-36 questions were unanswered, if more than 20 min. were taken to complete the survey (suggesting a substantial interruption), or if the user did not check the "honest self-report" option on the enrollment form. Group comparisons used chi-square tests for categorical variables and *t*-tests for continuous variables. Internal-consistency reliability of responses was tested using Cronbach's coefficient alpha.[10] Known-groups validity of responses was tested by using least squares regression to examine whether SF-36 scores varied with the age and gender of the respondent in the direction expected from U.S. population norms.

Results

During the 2-month pre-trial period, 1,667 visitors accessed the introductory page. Of these, 1,117 (67 percent) requested a survey, and 914 (82 percent) of those who requested a survey submitted their answers. During the randomized trial, 4,208 visitors accessed the introductory page; of these, 1,938 (46 percent) gave informed consent and received a survey (odds ratio 0.4, compared with the pre-trial period, $P < 0.0001$), and of these 1,615 (83 percent) submitted their answers (odds ratio 1.1, compared with the pre-trial period, $P = 0.33$). Among the 1,615 who submitted a survey, 80 (5.0 percent) were "just testing" the system, 34 (2.1 percent) were answering for someone else, 29 (1.8 percent) left one or more items blank, and 8 (0.5 percent) had completion times ranging from 21 to 105 minutes, leaving 1,464 final participants who met inclusion criteria.

Among the final participants, 745 were randomized to the expanded format and 719 to the matrix format. Their mean age was 36 years; 5.7 percent were 55 to 64 years of age and 2.3 percent were 65 years of age or older. Among participants, 57 percent were female; 3.9 percent were Asian/Pacific Islander, 2.1 percent Black, 2.4 percent Hispanic, 0.9 percent Native American, and 90 percent Non-Hispanic White. These characteristics did not differ significantly by study group ($P > 0.15$). In comparison, the 1997 U.S. adult population was 52 percent female; 11 percent age 55 to 64 years of age, 17 percent 65 years of age or older; 3.7 were percent Asian/Pacific Islander, 13 percent Black, 11 percent Hispanic, 0.9 percent Native American, and 72 percent Non-Hispanic White.[11]

Mean response times were 5.22 min for the expanded format and 5.07 min for the matrix format, a 9-sec difference in means that was not statistically significant ($P = 0.16$). Completion times were longer for older participants, for male participants, and for those using WebTV, for which a remote control device is typically used in lieu of a mouse and keyboard (Table 1, Response Time column). Multivariate regression showed that longer response times were also associated with poorer scores on the SF-36 Physical Component Summary (PCS) scale. Partial F tests showed that age, gender, WebTV use, and PCS scores were each independently correlated with response time ($P < 0.05$).

Physical health (PCS) scores averaged 51.6 for the expanded format and 51.0 for the matrix format ($P = 0.20$), compared with an expected mean of 51.9 based on U.S. norms, adjusted using published coefficients[8] to the age and gender of the Web participants. Mental Component Summary (MCS) scores averaged 45.4 in the "expanded" group and 44.5 in the "matrix" group ($P = 0.14$), compared with an expected mean of 49.4 for the age- and gender-adjusted U.S. norms. Cronbach's coefficient alpha was greater than 0.80 for each of the SF-36 subscales, indicating reliability adequate for intergroup comparisons; alpha scores did not differ significantly by survey format.

Physical health scores decreased with advancing age, mental health scores increased with advancing age, and men reported better health than women on both scales (Table 1). The national norms[8] show age and gender differences in the same direction, but older Web participants had better physical health than the norms and younger Web participants had worse mental health than the norms. In multivariate regression, partial F tests showed that age and gender were each independently correlated with PCS and MCS scores ($P < 0.05$).

Discussion

We hypothesized that a matrix arrangement of response options would speed survey completion, but we found evidence that favors no significant improvement. The expanded-format survey was twice as long vertically, but this less efficient layout apparently did not distract users significantly as they considered and answered each question. Since users who self-selected the expanded format were significantly slower in our earlier study,[9] the current study underscores the need for randomization to limit bias when comparing alternative Web designs.

Our hypothesis that the matrix format would not affect the reliability or the validity of users' responses was generally supported. Furthermore, the internal consistency and known-groups validity of Web users' responses indicate that they gave honest answers despite their anonymity. We may have improved honesty by asking users whether they were "just testing" the system, answering for someone else, or answering honestly for themselves. This feature allowed us to eliminate 7 percent of respondents who might otherwise have provided misleading data.

This study shows that alternative features for a health-related Web site may be evaluated rapidly and inexpensively by randomizing anonymous volunteers. Almost 500 subjects per month gave consent and participated even though the study's only promotion was through Web indexing sites and despite

the fact that the informed consent requirement appeared to reduce overall participation by 31 percent. Other Web-based health surveys[3–7] have recruited smaller numbers, probably because they were targeting specific diseases. We estimate that this project cost about $2,000 in labor and resources, an amount similar to the $1,916 cost estimated by Schleyer and Forrest.[1] Our cost per participant was therefore about $1.37, substantially lower than the per-subject costs of administering the SF-36 by mail ($27) or telephone interview ($48) among the U.S. population sample (2,474 participants).[12]

The chief limitation of conducting evaluations using anonymous Web volunteers is the under-representation of demographic subgroups that have poor Internet access. Although we achieved some representation of persons older than 65 years and those of Black and Hispanic race/ethnicity, the results of our study must be generalized with caution to these subgroups. We also found evidence that older Web respondents were healthier than expected for their age. We did not measure respondents' education or socioeconomic status, but we could also expect differences in these attributes. Disparities in Internet access are starting to lessen,[13] however, so new studies might address some of the demographic limitations by targeting under-represented subgroups for special recruitment.

Our finding of poorer mental health than expected, particularly among younger Web users, raises an additional potential limitation. Because our site provided users with their health status scores, we may have attracted a disproportionate number of persons with greater health concerns and poorer mental health. The potential biases induced by providing health status feedback deserve further investigation.

In conclusion, Web-based randomized studies may provide a method for testing a variety of hypotheses about survey content, and about the usability of interfaces for consumer health informatics. Over time, we expect that this method will contribute to a growing base of evidence and theory for guiding user-interface and survey design. Ultimately, the efficiency of online surveying should amplify our ability to study health care and the determinants of health.

The authors thank Drs. Ron Hays, Thomas Rice, Martin Shapiro, and Neil Wenger for helpful discussions and comments on the manuscript.

References ■

1. Schleyer TK, Forrest JL. Methods for the design and administration of Web-based surveys. J Am Med Inform Assoc. 2000;7:416–25.
2. Wyatt JC. When to use Web-based surveys. J Am Med Inform Assoc. 2000;7:426–9.
3. Soetikno RM, Provenzale D, Lenert LA. Studying ulcerative colitis over the World Wide Web. Am J Gastroenterol. 1997;92(3):457–60.
4. Soetikno RM, Mrad R, Pao V, Lenert LA. Quality-of-life research on the Internet: feasibility and potential biases in patients with ulcerative colitis. J Am Med Inform Assoc. 1997;4:426–35.
5. Hilsden RJ, Meddings JB, Verhoef MJ. Complementary and alternative medicine use by patients with inflammatory bowel disease: an Internet survey. Can J Gastroenterol. 1999;13(4):327–32.
6. Eysenbach G, Diepgen TL. Epidemiological data can be gathered with World Wide Web. BMJ. 1998;316(7124):72.
7. Lenert LA, Cher DJ. Use of meta-analytic results to facilitate shared decision making. J Am Med Inform Assoc. 1999;6(5):412–9.
8. Ware JE Jr, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey: Manual and Interpretation Guide. Boston, Mass.: Nimrod, 1993.
9. Bell DS, Kahn CE Jr. Health status assessment via the World Wide Web. Proc AMIA Annu Fall Symp. 1996:338–42.
10. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika. 1951;16:297.
11. U.S. Census Bureau. Statistical Abstract of the United States. 1998. Available at: http://www.census.gov/prod/www/statistical-abstract-us.html. Accessed Aug 20, 2000.
12. McHorney CA, Kosinski M, Ware JE Jr. Comparisons of the costs and quality of norms for the SF-36 health survey collected by mail versus telephone interview: results from a national survey. Med Care. 1994;32(6):551–67.
13. U.S. Department of Commerce. Falling Through the Net: Toward Digital Inclusion. Oct 18, 2000. Available at http://osecnt13.osec.doc.gov/public.nsf/docs/fttn-tdi-executive-summary. Accessed Dec 14, 2000.