WEB SURVEY DESIGN AND ADMINISTRATION

MICK P. COUPER
MICHAEL W. TRAUGOTT
MARK J. LAMIAS

Abstract Many claims are being made about the advantages of conducting surveys on the Web. However, there has been little research on the effects of format or design on the levels of unit and item response or on data quality. In a study conducted at the University of Michigan, a number of experiments were added to a survey of the student population to assess the impact of design features on resulting data quality. A sample of 1,602 students was sent an e-mail invitation to participate in a Web survey on attitudes toward affirmative action. Three experiments on design approaches were added to the survey application. One experiment varied whether respondents were reminded of their progress through the instrument. In a second experiment, one version presented several related items on one screen, while the other version presented one question per screen. In a third experiment, for one series of questions a random half of the sample clicked radio buttons to indicate their answers, while the other half entered a numeric response in a box. This article discusses the overall implementation and outcome of the survey, and it describes the results of the imbedded design experiments.

I. Introduction

Web surveys are proliferating at a rapid pace. Despite this increase, we have yet to see much empirical research aimed at exploring various features of Web survey design. With the graphic and multimedia capabilities of the World Wide Web, the survey researcher has an almost unlimited set of design choices in developing a survey for administration on the Web. As a result, the quality of design seen in Web surveys is highly variable. There generally appears to be

MICK P. COUPER and MICHAEL W. TRAUGOTT are at the Institute for Social Research, University of Michigan, and MARK J. LAMIAS is at Thinkologies, Inc. We thank Dave Dybicki, Robyn Rissell, and Qian Yang for programming the instrument; *The Michigan Daily* for its support of this project; and the reviewers for their helpful comments on earlier drafts.

Public Opinion Quarterly Volume 65:230–253 © 2001 by the American Association for Public Opinion Research
All rights reserved. 0033-362X/2001/6502-0004$02.50

little or no concern for the potential effect of design on the answers that people may give.

Research on self-administered surveys suggests that the design of the instrument may be extremely important in obtaining unbiased answers from respondents. The work of Schwarz and colleagues (e.g., Schwarz 1995, 1996; Schwarz, Strack, and Mai 1991) suggests that in the absence of an interviewer to motivate the respondent or to provide guidance on how to answer each question, the respondent seeks such information from the instrument itself, using both the verbal and visual elements of the interface (Ware 2000). There are several studies that show how the physical layout of a paper questionnaire may affect respondent answers. For example, Smith (1995) shows several examples of unintentional layout changes producing differences in both self-administered and interviewer-administered surveys (see Sanchez [1992], for another example), while Dillman, Redline, and Carley-Baxter (1999) show how routing or skip errors are affected by the design of a paper questionnaire.

In Web surveys, question text can be supplemented with a variety of visual elements, including color, graphics, and interactive features that provide immediate feedback on actions taken by the respondent. Given the greater range of visual design features and the interactive nature of Web surveys, attention to the visual elements of design is important. These auxiliary features (Redline and Dillman 1999) can facilitate or distract from the task of completing the survey. In general, we would expect that when the visual design elements complement or support the verbal features of the survey instrument, efficiency and data quality gains may be achieved. Despite the many Web surveys already being fielded, there appears to be little empirical research to date on the effect of various design decisions on either the willingness to complete a Web survey or the answers given to the survey questions (one exception is the study by Dillman et al. [1998]). With this in mind, the goal of this study was to explore systematically some of the factors that may affect completion of a Web survey and to examine possible effects of design choices on responses provided to the survey.

II. Background and Hypotheses

The experiments we describe here were included in a survey of University of Michigan students conducted for the student newspaper, the *Michigan Daily*. The newspaper was interested in generating news stories based on student attitudes. At the time of the survey, there were two lawsuits over admission policies at the university, one against the College of Literature, Science and the Arts (LS&A) and the other against the Law School.[1] The substantive focus of

[1] The sample of currently enrolled students was stratified with the intent of facilitating comparisons between relevant student populations. There were two strata of graduate students, defined by the Law School, the locus of one of the suits, and all other graduate students. There were

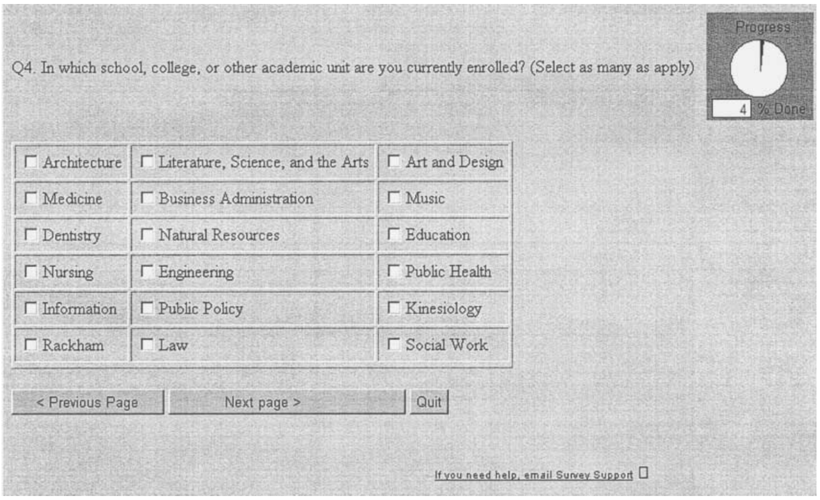
the study was to ascertain student attitudes about affirmative action generally, affirmative action in education, and admission policies at the university. The main correlates of such attitudes that were explored included student characteristics (demographics, class status, and course work), knowledge of university policies, and characteristics of social networks (see Edy and Traugott 1999).

The design experiments embedded in the survey were added after the content of the study had been determined. Given the content of the survey and the software at our disposal, we attempted to identify key design features that could readily be implemented within the short time frame available before the survey went into the field, without influencing too much the substantive information to be collected. An additional constraint was the design limitations of the software used. Given these constraints, we chose to focus on three design features in a 2 × 2 × 2 design.

1. One random half sample was given a progress indicator, while the other was not. In traditional mail surveys when a questionnaire is not returned, it is not known whether respondents refused to participate at all or began to complete the survey but gave up at some point, and, if so, when in the instrument they decide to stop and for what reason. One advantage of interactive Web surveys is that such abandonments are measurable. On the other hand, in interactive Web surveys (unlike in scrollable survey forms or mail surveys), respondents do not know how far they have progressed in the instrument and how much is left to complete. Thus, it is possible that surveys are abandoned close to the end, when respondents lose motivation. Abandonments appear to be a major concern, especially in self-selected Web surveys (see Jeavons 1998). The benefit of progress indicators is that they inform respondents of their progress through the instrument and should motivate them to complete the survey. At the same time, progress indicators may require additional download time. If so, they may slow the survey sufficiently to result in lower completion rates. In an interactive Web survey that does not permit a respondent to skip an item, the presence of a progress indicator should have no effect on item nonresponse because the effort of clicking a "don't know" (DK) is the same as selecting a substantive response. In a survey where skips are permissible, we expect that progress indicators may reduce item nonresponse by increasing respondent motivation.

There are many ways to indicate progress through a Web survey instrument. We implemented a graphic and text indicator in the upper right corner of every screen. The indicator was out of the main visual field, but it was likely

also two strata for the undergraduate student body, defined by LS&A, the locus of the other suit, and students in the Schools of Nursing and Engineering. The University of Michigan uses a selection index to assign points to applicants based on a wide variety of criteria, which was the cause of the suit. Gender is not a factor in the index except for underrepresented male students in the School of Nursing and female students in the School of Engineering. Within each stratum, separate samples of white and nonwhite students were drawn. Some undergraduates fell in neither stratum (e.g., students in the School of Natural Resources). Overall, the four strata covered 87 percent of the student body at the university.



Q4. In which school, college, or other academic unit are you currently enrolled? (Select as many as apply)

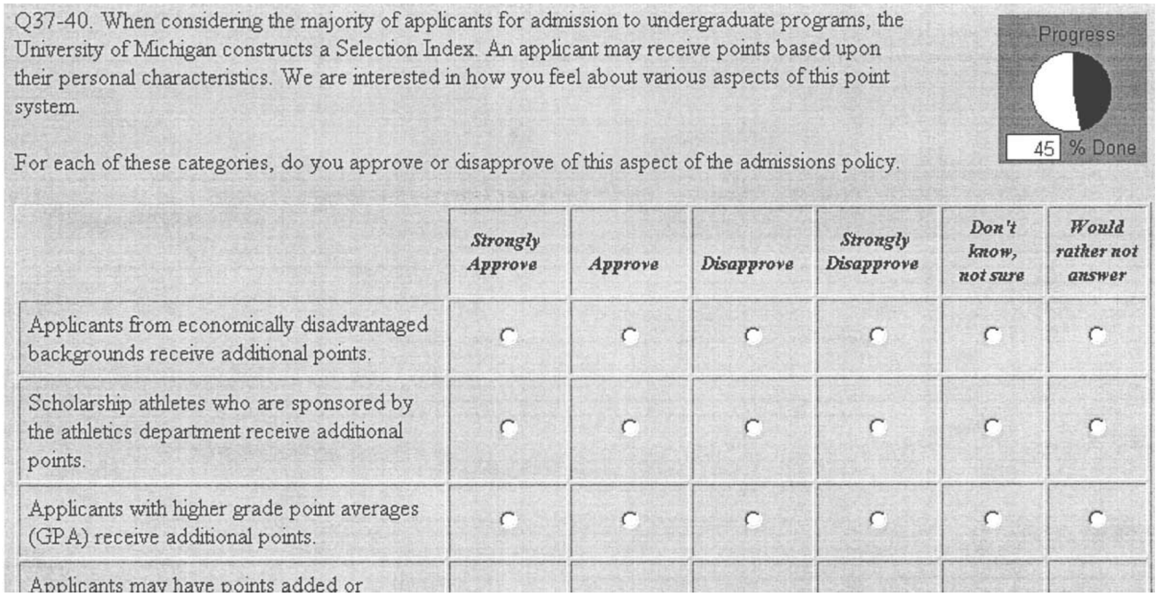
☐ Architecture	☐ Literature, Science, and the Arts	☐ Art and Design
☐ Medicine	☐ Business Administration	☐ Music
☐ Dentistry	☐ Natural Resources	☐ Education
☐ Nursing	☐ Engineering	☐ Public Health
☐ Information	☐ Public Policy	☐ Kinesiology
☐ Rackham	☐ Law	☐ Social Work

Figure 1. Example of progress indicator

to be noticed because of its size and brightness (see fig. 1). In addition, we added several motivational screens at key points in the instrument (e.g., "You are about one third done with the survey . . . please continue as your answers are important to us"). In summary, our key hypothesis is that indicating progress through the survey will increase completion of the survey and reduce abandonments.

2. One random half sample was shown sets of questions with common response categories all on the same screen, while the other saw the questions one per screen (see figs. 2a and 2b). A key difference between scrollable Web surveys and interactive Web surveys is that in the former respondents can browse the entire survey before answering a single question. In the former case, context or order effects should be minimized as the respondent can see several related questions at once. In the latter case, the respondent often sees only one question at a time, potentially enhancing order effects. In such surveys the grouping of related items on a single screen is likely to lead respondents to view the items as related entities, thereby increasing the correlation among them (Schwarz, Strack, and Mai 1991; but also see Metzner and Mann 1953).

Regardless of the substantive reasons for combining items on a single screen, there may also be efficiency reasons for doing so. Grouping related items is likely to reduce the time taken to complete the survey by requiring only one orientation to the question-and-response format, rather than reorienting on each screen in a single-item-per-screen design (see Fuchs, Couper, and

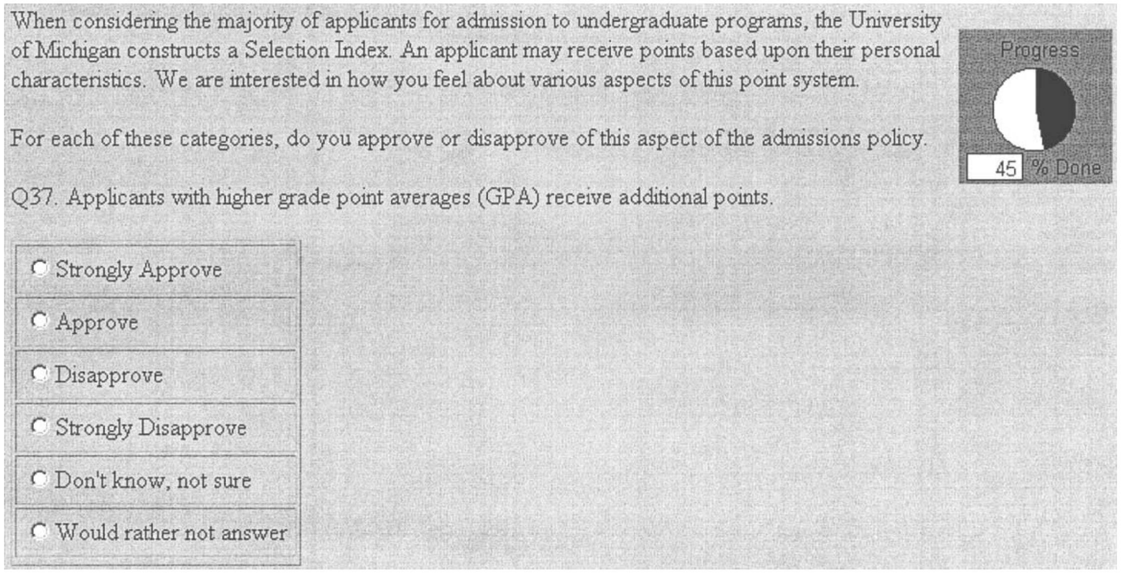


Q37-40. When considering the majority of applicants for admission to undergraduate programs, the University of Michigan constructs a Selection Index. An applicant may receive points based upon their personal characteristics. We are interested in how you feel about various aspects of this point system.

For each of these categories, do you approve or disapprove of this aspect of the admissions policy.

	Strongly Approve	*Approve*	*Disapprove*	*Strongly Disapprove*	*Don't know, not sure*	*Would rather not answer*
Applicants from economically disadvantaged backgrounds receive additional points.	○	○	○	○	○	○
Scholarship athletes who are sponsored by the athletics department receive additional points.	○	○	○	○	○	○
Applicants with higher grade point averages (GPA) receive additional points.	○	○	○	○	○	○
Applicants may have points added or						

Figure 2a. Example of multiple items per screen



When considering the majority of applicants for admission to undergraduate programs, the University of Michigan constructs a Selection Index. An applicant may receive points based upon their personal characteristics. We are interested in how you feel about various aspects of this point system.

For each of these categories, do you approve or disapprove of this aspect of the admissions policy.

Q37. Applicants with higher grade point averages (GPA) receive additional points.

○ Strongly Approve

○ Approve

○ Disapprove

○ Strongly Disapprove

○ Don't know, not sure

○ Would rather not answer

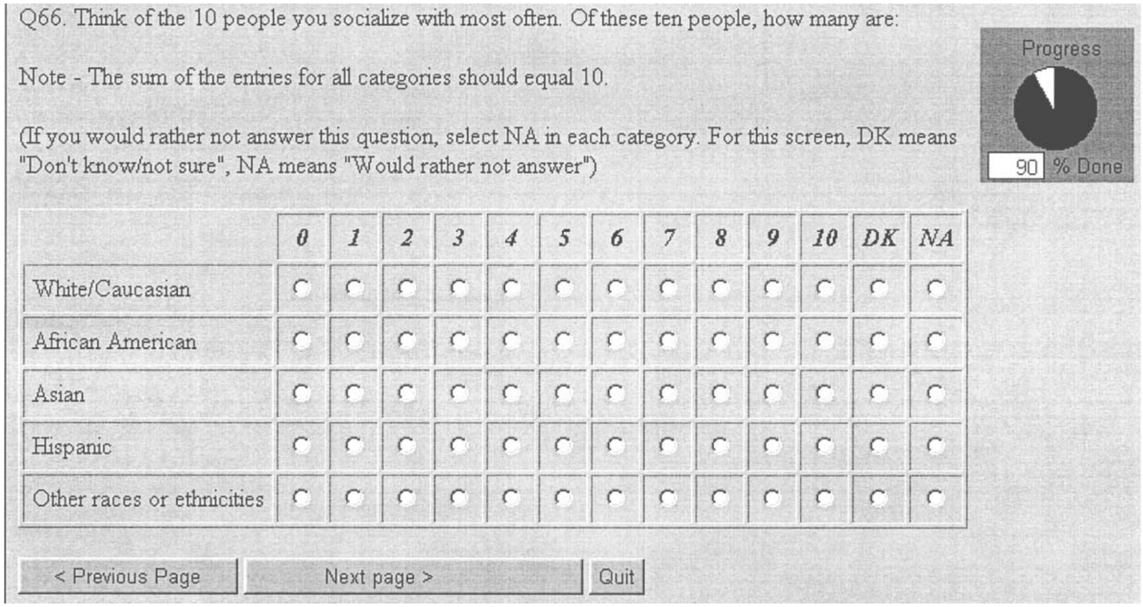
Figure 2b. Example of single item per screen

Hansen [2000] for a discussion of screen orientation effects in an interviewer-administered survey). This is analogous to the instance where an interviewer asks a series of questions with the same response categories, and the respondent, learning what is expected, answers more quickly, often before the entire question is read. In addition, download time may be reduced by having several questions on one screen, limiting the number of transfers from the Web server. In summary, our hypotheses are that grouping related items on a screen (1) would increase the correlation (internal consistency or scalability) among the items and (2) would reduce the time taken to complete the set of items, relative to presenting each item on a separate screen.

3. The survey contained a series of similar questions in which the students were asked to describe the various racial-ethnic characteristics of their social networks (i.e., 10 friends with whom you socialize most). Students were asked to indicate the number of friends in each category in a way that they totaled to 10. For one half of the sample, a radio button format was used, while those in the other half of the sample were required to type numbers in entry boxes (see figs. 3a, 3b, and 3c).

Our assumption here was that clicking a radio button requires less effort on the part of respondents than typing a response in a box. Furthermore, a mouse can be used for all input, rather than the keyboard. Not only should this take less time, but it should also lead to lower item-missing data because one could simply leave a box empty, while the radio button version required an entry for each field. Furthermore, radio buttons restrict the range of permissible answers, thereby preventing out-of-range answers.

On the other hand, the input area for the radio button version contains much more information (65 input locations as opposed to five for the entry box version). We expect this to require greater hand-eye coordination, in that the respondent has to line up both the row and column to identify the appropriate radio button. The vertical and horizontal lines may serve to make this visual alignment more difficult. In addition, the nature of the task (entering five numbers to add up to 10) is visually facilitated by entering these in a vertical format, as is the cultural norm for adding up a set of numbers. Thus, the radio button version is more complex, both in terms of screen density (Tullis 1983, 1988) and of alignment and grouping complexity (Parush, Nadir, and Shtub 1998). We thus expected that more of the entries would add to 10 in the entry box version than in the radio button version. In summary, the task involved two key elements: (1) entering each response and (2) determining that they added to 10. We expected that the radio buttons would facilitate the first element of the task but that the entry boxes would make the second element easier. With these experimental goals in mind, in Section III we describe the design and implementation of the survey. In the remaining sections we focus on the results of the three experiments.



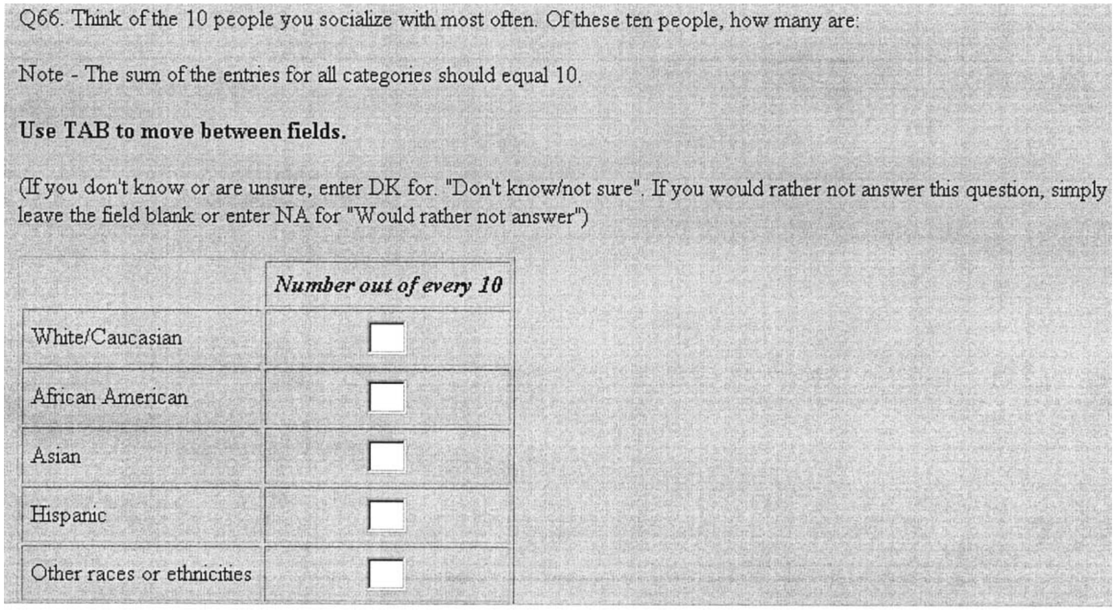
Q66. Think of the 10 people you socialize with most often. Of these ten people, how many are:

Note - The sum of the entries for all categories should equal 10.

(If you would rather not answer this question, select NA in each category. For this screen, DK means "Don't know/not sure", NA means "Would rather not answer")

	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*DK*	*NA*
White/Caucasian	○	○	○	○	○	○	○	○	○	○	○	○	○
African American	○	○	○	○	○	○	○	○	○	○	○	○	○
Asian	○	○	○	○	○	○	○	○	○	○	○	○	○
Hispanic	○	○	○	○	○	○	○	○	○	○	○	○	○
Other races or ethnicities	○	○	○	○	○	○	○	○	○	○	○	○	○

Figure 3a. Radio button version



Q66. Think of the 10 people you socialize with most often. Of these ten people, how many are:

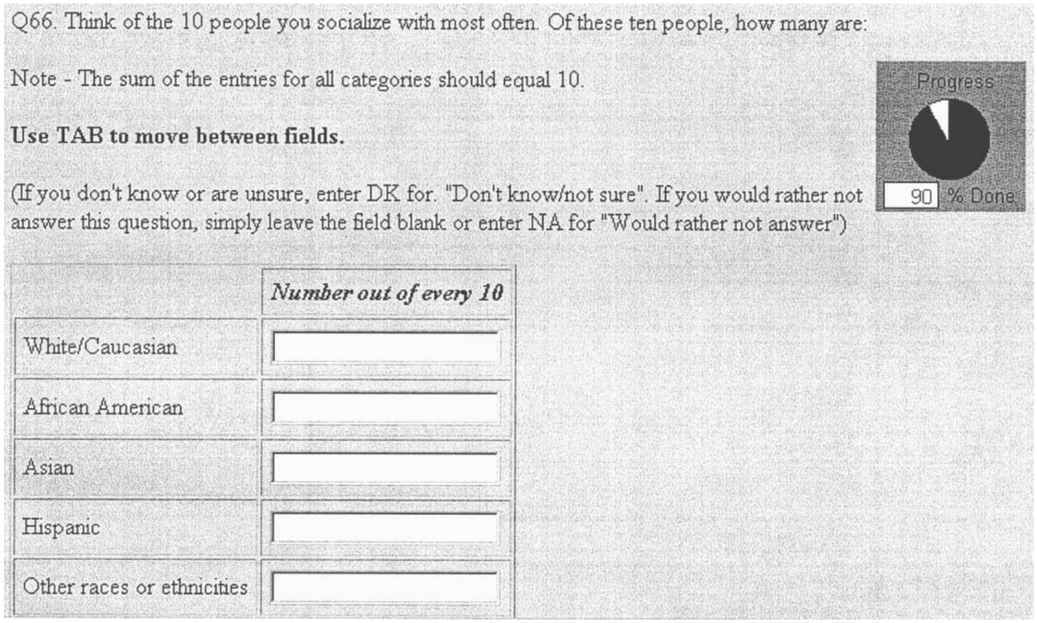
Note - The sum of the entries for all categories should equal 10.

Use TAB to move between fields.

(If you don't know or are unsure, enter DK for. "Don't know/not sure". If you would rather not answer this question, simply leave the field blank or enter NA for "Would rather not answer")

	Number out of every 10
White/Caucasian	□
African American	□
Asian	□
Hispanic	□
Other races or ethnicities	□

Figure 3b. Short entry box version



Q66. Think of the 10 people you socialize with most often. Of these ten people, how many are:

Note - The sum of the entries for all categories should equal 10.

Use TAB to move between fields.

(If you don't know or are unsure, enter DK for. "Don't know/not sure". If you would rather not answer this question, simply leave the field blank or enter NA for "Would rather not answer")

	Number out of every 10
White/Caucasian	
African American	
Asian	
Hispanic	
Other races or ethnicities	

Figure 3c. Long entry box version

III. Design and Implementation

The fact that every student at the university has an e-mail account and free access to the Web makes the use of the Web for surveying students attractive. Cooperation from the Registrar made it possible to design a stratified sample of students to be drawn by that office. These names were then matched to the Information Technology Division's (ITD) list of unique names to create e-mail addresses for all students in the sample.

A main sample of 1,602 students and a second replicate of 540 students was selected and distributed. Our focus in this article is on the main sample only; the second replicate was released later and was found to contain some duplicate names from the first release. The Web survey was implemented using Surveycraft's ScyWeb program, which is an interactive system in that the answer to one item or screen is submitted before delivery of the next screen. This permits control of skips and branching and embedded edit checks. The software provided flexibility in the design of screen formats, and it permits suspension and resumption of the survey by respondents. Furthermore, the ScyWeb software provided timing data by screen to track respondent flow through the instrument.

On the other hand, ScyWeb suffered from a number of limitations, exacerbated by the fact that the software was new to the programming staff. A beta version of the system was used that was not well documented. The system had no facility for progress indicators, but it was flexible enough to permit the inclusion of these as embedded objects in the instrument. ScyWeb had a facility to randomize the order of items across screens (when presented one screen at a time), but not within screens (when several items appeared on one screen).

The survey was designed as an interactive instrument. Respondents completed one screen at a time, at which point the answer was transmitted to the server and the following item displayed on the screen. This is contrasted with a scrolling approach where the entire survey is in one long HTML page and the respondent clicks "submit" after completing all items to transmit the data to the server (Dillman [2000] argues for this latter design approach). One advantage of the interactive approach is that it offers greater control over skips and branching. Another benefit is that information from partial completes (abandonments) is available. With the single-form approach (as with mail surveys), one cannot distinguish between a sample person who opens the survey but decides not to complete it, and one who completes almost all questions but fails to return the questionnaire to the survey organization. A disadvantage of the interactive approach (as implemented in ScyWeb) is that respondents are required to provide an answer to each question (hence, a DK or "choose not to answer" response should be provided) in order to proceed to the next; they also cannot ascertain the length of the instrument or how

Table 1. Response Rate Results

	Percent	*N*
Nonrespondents:		
Checked e-mail after first mailing	47.7	764
Did not check e-mail after first mailing	5.2	84
Total nonrespondents	52.9	848
Partial interviews (abandonments)	5.6	89
Completed interviews	41.5	665

far they have to go to completion (one reason progress indicators are deemed important in such surveys).

Eight different versions of the instrument were created reflecting the 2 × 2 × 2 experimental design. The versions (and their associated URLs) were randomly assigned to the 1,602 subjects in the sample, with 200 cases in 6 treatments and 201 in the remaining 2. Anonymous identification (ID) numbers and passwords were generated to limit access to the site to those in the sample and to prevent multiple completions of the survey. A series of articles in the *Michigan Daily* were used to announce and promote the survey.

Sampled students were sent an advance e-mail message notifying them of their selection. The message contained the URL and the study-generated ID and password for completing the survey. An incentive was offered to those who completed the survey—a copy of the book *We're Number One: The National Championship Season.* This was done by providing a password at the end of the survey, which students could use to pick up their copy of the book at the *Michigan Daily* offices on campus. Students were provided e-mail addresses to contact if they had questions or concerns about the survey. (See the appendix for an example of the e-mail invitation.)

E-mail messages inviting students to participate in the survey were sent on March 30, 1999. Three days after the initial e-mail was sent, a reminder was sent to those who had not completed the survey. A final reminder was sent on April 8, and the survey was closed on April 18.

IV. Results

Table 1 shows the distribution of sample cases. A Unix-based script was written to check the last login time and was run against the sample list on April 27, 1999. From this we determined that 39 students (2.4 percent of the sample) had not accessed the university e-mail server since the beginning of the semester (January 1). An additional 45 students had checked e-mail during the semester, but not since the time of the initial invitation to do the survey.

Thus, we find no evidence that a total of 84 persons (only 5.2 percent of the sample) had used e-mail during the survey period. This suggests that noncoverage (lack of access to or use of e-mail) is not a great cause of concern in a student population such as this.

The overall response rate for the study was just over 41 percent, or 47 percent if we include the partial interviews. This is in line with expectations based on a number of other Internet-based studies of similar populations (see, e.g., Couper, Blair, and Triplett 1999; Schaefer and Dillman 1998; Sheehan and Hoy 1999).

As others have noted (Comley 1997; Terhanian 1999), a major advantage of Internet surveys, conducted via e-mail or the Web, is the speed with which completed questionnaires are returned. This study was no exception: by the end of the first day, a total of 201 completed surveys were received, making up 30 percent of the total number of completes. By the third day, 50 percent of all completes had been returned.

Within each of the schools or strata, minority students had lower response rates than their white counterparts. Overall, the response rate for white students was 46.1 percent (369/801), compared with 37.0 percent for nonwhite (minority) students (296/801), a statistically significant difference ($p < .01$). We do not know how much of this difference results from the topic of the survey or from differential use of or familiarity with the Web.

As noted earlier, the e-mail invitation to sample persons provided them with a unique user ID and password that they would use to gain access to the survey on the Web site. This approach is designed to limit “ballot-stuffing” and ensure that only the sampled respondent completes the survey, as well as to provide an assurance of confidentiality. One unintended consequence of this approach may be sample persons who fail to complete the survey because they cannot enter the correct information. The five-character passwords automatically generated by the system consisted of three lowercase letters alternated with two single digit numbers (e.g., y2n6p). We found that sample persons who were sent passwords containing ambiguous characters (the letters l [el] and o [oh] and the numbers 1 [one] and 0 [zero]) were significantly ($p < .05$) less likely to start the survey than those whose passwords did not contain such ambiguities. Specifically, 43.7 percent of those with ambiguous passwords successfully started the survey, compared with 50.4 percent of those having passwords with no such ambiguities ($p < .01$).

As further evidence of the problem caused by ambiguous passwords, we found that 27 of the 57 messages sent by sample persons in response to the initial invitation were in regard to problems with passwords. Of these, all but one had ambiguous passwords as defined above. After being sent clarification of the password specification, 17 of the 26 subsequently completed the survey.

A. PROGRESS INDICATOR VERSUS NO PROGRESS INDICATOR

As noted earlier, our expectation was that the progress indicator would increase completion of the survey among those who began it. We thus need to contrast completed versus partial interviews (abandonments). Among those who started the survey, 89.9 percent of the 378 who received a progress indicator completed the survey, compared with 86.4 percent of the 376 who did not receive a progress indicator. While in the expected direction, this difference does not reach statistical significance ($p = .13$). We examined the 89 partial interviews in detail and found no apparent peak or trend in the point at which abandonments occurred. This suggests that there was no single question that led to many abandonments, nor were respondents dropping out toward the end of the survey. In fact, 53 of the 89 partials apparently failed to complete any questions (we have evidence that they logged in, but no data). This may partly result from problems with the software, as we received messages from some of these persons saying that they had completed the survey.

One possibility we considered is that the graphic image used for the progress indicator may have increased download time sufficiently to counteract the positive effect of the graphic. A similar problem may have confounded the study by Dillman et al. (1998) in their comparison of "plain" versus "fancy" design on Web survey completion. The color features of the fancy design resulted in longer download times, which could have explained the higher abandonment rates they found.

We compared the average time to complete the survey with and without the progress indicator. The mean time in minutes to complete the survey was significantly higher ($p < .01$) for the progress indicator version (22.7 minutes) than for the version with no progress indicator (19.8 minutes), providing some support for this speculation. Furthermore, the effect of the progress indicator remains unchanged when comparing surveys completed at home or on campus. Surveys done at home (over modem) took longer (23.1 minutes for progress indicator, 19.9 minutes for no progress indicator) than those completed in a computer lab or library (high-speed line; 21.8 vs. 18.7 minutes). Thus, the added download time associated with the graphic progress indicator on each screen may have mitigated the positive effect of having such an indicator. We do not know for sure whether the additional time resulted from the added file size of the progress indicator; an alternative explanation is that those respondents who received the progress indicator took more care over their answers. An obvious next research step is to explore whether a progress indicator that does not add to download time (e.g., a simple text indicator) has a greater positive effect on likelihood of completion.

We further hypothesized that the presence of a progress indicator should have no effect on item-missing data in a survey such as this where respondents must provide some response to every question and where the effort of selecting a DK response is no less than choosing a substantive response. In other words, satisficers gain nothing by repeatedly selecting the DK option. In contrast, in surveys where a response is not required (i.e., where skipping is permitted), progress indicators may indeed reduce item-missing data through motivating respondents to provide an answer and continue with the survey.

To examine this question, we summed all “DK/not sure” and “would rather not say” (NA) responses across all 69 nondemographic items in the survey. The mean number of such responses is virtually identical in both versions: 7.91 for the progress indicator version and 7.92 for the version with no progress indicator.

B. MULTIPLE-ITEM SCREENS VERSUS SINGLE-ITEM SCREENS

Based on Sudman, Bradburn, and Schwarz (1996, p. 123), we hypothesized that grouping related items on one screen would increase the correlations among them. We are not arguing that one format necessarily yields better data quality than another but merely wish to examine whether response differences occur. For substantive reasons, there are likely to be times that separating items across several screens may be desirable, especially if one was interested in the effect of item order.

In order to examine the effect of multiple-item screens, we examined two types of items. One was a five-item knowledge measure in which the items were either all on one screen or on five separate screens. The other set consisted of 11 attitude items, grouped into three screens (4, 4, and 3 items, respectively) versus items presented on 11 separate screens. Each item was measured on a 5-point Likert-type scale.

In table 2 we present the item-total correlations (measured by Cronbach’s alpha coefficient) for the sets of attitude items that could appear together on each screen in the multiple-item version. As expected, the correlations are consistently higher among items appearing together on a screen than items separated across several screens. However, the overall effect is not large, and none of the differences between each pair of correlations reach statistical significance. We also conducted factor analyses of the set of attitude items and find similar factor structures across the two versions. Thus, we find modest support for the grouping hypothesis. We expect any such effects to be relatively modest, but only replication can resolve this for other types of items.

We also hypothesized that there would be efficiency gains of putting related items on the same screen. The results in the lower part of table 2 support this expectation: the multiple-item-per-screen version took significantly less time ($p < .01$) to complete the 16 items than the single-item-per-screen version. The differences for the five knowledge items and 11 attitude items measured separately were also statistically significant ($p < .05$).

We further speculated that the effect for the multiple-item screens would be larger after the first item on each screen. A multiple-item screen contained more information than a single-item screen, so the initial orientation to the

Table 2. Effect of Multiple- versus Single-Item Screens on Item-Total Correlations and Mean Time to Complete

	Multiple-Item Screen	Single-Item Screen	Difference
Item-total correlations:			
Knowledge scale (5 items)	.460	.456	+.004
Attitude scale (11 items):			
First set (4 items)	.318	.245	+.073
Second set (4 items)	.301	.175	+.126
Third set (3 items)	.311	.295	+.006
Total (11 items)	.640	.610	+.030
Mean time to complete scales (in seconds):			
Knowledge scale (5 items)*	54.2	65.9	−11.7
Attitude scale:			
First set (4 items)	55.1	57.0	−1.9
Second set (4 items)	31.6	39.0	−7.4
Third set (3 items)	26.9	32.1	−5.2
Total (11 items)*	113.6	128.1	−14.5
Total (all 16 items)**	167.8	194.0	−26.2
N	(338)	(327)	

* $p < .05$.
** $p < .01$.

screen should take slightly longer. However, once the initial orientation is done, the remaining items on the screen should take less time. We cannot distinguish the item-by-item times in the multiple-item version, only the screen-by-screen times. But generally we would expect to see an increasing advantage of the multiple-item approach after the initial screen or set of items (requiring the greatest amount of orientation). The results in the lower part of table 2 provide some modest support for this. While the time difference between the multiple-item and single-item versions for the first four items is small (1.9 seconds), the difference increases to 7.4 seconds and 5.2 seconds for the remaining two sets of items; however, none of these differences reach traditional levels of statistical significance ($p > .10$).

Finally, we examined the effect of multi- versus single-item screens on item-missing data (DK/not sure and NA responses). On the one hand, if multiple-item screens are less burdensome to respondents, they may produce more substantive answers. On the other hand, presenting a series of responses to several questions on one screen may encourage the use of response sets such as clicking the DK choice down the column for all items. In order to examine the effect of design on item nonresponse, we summed the DK/NA responses across all five knowledge items and 11 attitude items. The mean number of such responses for these 16 items in the multiple-item version was significantly lower ($p < .01$) than for the single-item version (1.2 vs. 1.7). We found similar results when examining the four sets of items separately. Not only does the multiple-item version appear to take less time, but it also appears to result in fewer nonsubstantive answers. We examined other forms of response set (e.g., answering all 1s or all 5s) but found no significant differences between versions.

C. RADIO BUTTONS VERSUS ENTRY BOXES

In one section of the survey, we were interested in obtaining information about the racial and ethnic composition of various domains of the respondents' social networks. We wanted the respondents to indicate what percentage of their acquaintances they would classify within different racial and ethnic groups. They could indicate the number by clicking on a radio button next to a number or by typing a number in a box (see figs. 3a and 3b).

The choice of radio buttons versus entry boxes depends on the type of task. We believe that for most Web survey items, radio buttons are preferred because this allows mouse-only entry. Indeed, most Web surveys we have reviewed adopt this approach, using entry boxes only for open-ended text responses. However, we felt that the particular response task in this series of questions lent itself more to an entry box format.

We hypothesized that the radio button version would take less time to complete than the entry box version, given the added burden of typing in numbers versus clicking a button. We also hypothesized that the radio button version would produce lower item nonresponse than the entry box version, given that a DK and NA response had to be explicitly selected in the former version, whereas respondents had the option of simply leaving an entry box blank in the latter version. Finally, we expected that the radio button version would also make the task of selecting five numbers adding up to 10 more difficult, as it is harder to visualize and total five numbers in this version than when they are vertically aligned.

In addition to the planned comparison between these two versions, a programming slip led to an additional inadvertent test of short entry boxes versus long entry boxes. In three of the four random groups assigned to the entry box version, the boxes were significantly longer (wider) than in the fourth version. We hypothesized that providing more space than was necessary for the response may induce respondents to provide more information. In other words, the task communicated to the respondent by the long entry box may appear to contradict the task communicated in the question.

The task consisted of four sets of items similar to those presented in figures

Table 3. Effect of Entry Format on Item-Missing Data and Invalid Responses

	Radio Button	Short Entry Box	Long Entry Box
Cases with one or more DK/NA/ blank responses (%):			
First set***	21.3	57.7	51.2
Second set***	15.9	49.3	53.7
Third set***	11.6	60.6	57.4
Fourth set***	10.8	66.2	56.6
Mean number of DKs/NAs***	1.66	5.20	5.29
N	352	71	242
Cases with one or more invalid responses (%)***	.0	11.3	20.7
N	348	71	242

NOTE.—DK = don't know; NA = would rather not answer.
*** $p < .001$.

3a, 3b, and 3c.[2] In terms of completion time, we did not find significant differences by version ($p > .10$). On average, the radio button version took 183 seconds to complete, compared with 168 seconds for the short entry box version and 180 seconds for the long entry box version.

Our next question was whether differences on item nonresponse occurred between versions. In the radio button version, DK and NA options were provided, whereas in the entry box versions respondents were instructed to type "DK" or "NA" in the box or to leave the field blank. We expected that this approach, giving the option of not entering anything in the entry box version, would yield more such responses. Table 3 shows the percent of cases with one or more of the five items missing for each of the four series of items, as well as the mean item-missing data for the full set of all 20 items. In each case, including the overall mean, the radio button version differs significantly from either entry box version ($p < .05$), but the entry box versions do not differ from each other ($p > .1$). The results support our expectations that more missing data would result from the entry box versions.

Furthermore, respondents receiving the short entry box version were much

[2] These figures show only one of the four questions in this series. The other three questions were (1) "Thinking of the building, house or residence hall where you currently live, on average, how many out of every ten people who live there are . . . "; (2) "In the classes you are currently taking, on average, how many out of every ten people are . . . "; and (3) "Think of the 10 people you socialize with most often. Of these ten people, how many are . . . ".

more likely to leave the box blank (as opposed to typing a DK/NA response) than those receiving the long entry box version. For example, in the first set, 52.1 percent of respondents (out of 57.7 percent) left one or more of the four items blank in the short entry box version, compared with only 17.7 percent in the long entry box version (similar results are found for the remaining three sets of items). Conversely, respondents were more likely to type in an explicit DK/NA response in the long entry box version than in the short entry box version.

In addition to item-missing data differences, we expected that the entry box version would facilitate the task of entering five numbers that add to 10. There were two ways we could evaluate task completion: whether out-of-range or invalid responses were entered on any items and whether the total for each screen added to 10 (as instructed). We did not build in an edit check for this set of items, and the system accepted all responses entered, in order to reduce respondent burden. We therefore defined an invalid response as anything other than a number between zero and 10, DK/NA, or blank. Table 3 also shows the percent of cases with at least one invalid response on any of the four sets of items.

By definition, the radio box version should have no invalid entries, as the respondents could only select the items on the screen. However, we retain this column in table 3 for purposes of contrast. We see that in the long entry box version, respondents were significantly ($p < .01$) more likely to enter an invalid response than in the short entry box version. Further inspection revealed that respondents were being guided by the size of the entry box and providing more information than was required. For example, instead of simply providing a number, respondents would enter "about 3," "between 4 and 5," and so on.

Finally, what proportion of subjects entered or selected a set of numbers that added up to 10 for each screen? Given the large number of out-of-range responses and missing data, we made the following simplifying assumptions: (1) we treated all DK/NA and blank responses as zeros, and (2) we assigned all ranges the midpoint of the range (e.g., an answer of "between 3 and 4" was assigned a value of 3.5). These results are presented in table 4, separately for each question and then across the full set of items. The data show that, in all cases, the level of correct summation tends to increase with respondent exposure to the question form, suggesting the impact of increased familiarity with the format. For each of the four sets of items (as well as for the total), the differences between the three formats reach statistical significance ($p < .01$). Given that the items were answered, our expectation that the entry box version would facilitate the completion of the specified task (summing the responses correctly) is confirmed. We note further that the short entry box version has a greater percentage of respondents completing the task than the long entry box version, again suggesting that the longer box encouraged respondents to provide more information than was required.

Table 4. Effect of Entry Format on Completion of Task as Requested

	Radio Button	Short Entry Box	Long Entry Box
Cases (%) with responses adding to 10:			
First set***	67.1	85.9	79.8
Second set**	75.2	91.5	83.1
Third set**	83.2	97.2	87.6
Fourth set**	81.0	91.5	88.8
Total (%)***	50.9	76.1	67.4
N	351	71	242

** $p < .01$.
*** $p < .001$.

In summary, we found marginal support for the hypothesis that a progress indicator reduces abandonments, and we speculate that the advantages may have been counteracted by the longer download time associated with the use of the progress indicator. Furthermore, the overall level of abandonment was relatively low for this survey, and even some of that may have resulted from software problems. We thus speculate that the effect of a progress indicator would be larger in a survey with higher burden, where the progress indicator does not add to download time. We are currently exploring text-based alternatives to the graphic progress indicator used in this study.

We found faster completion times and less missing data for multiple-item screens and saw some suggestion of stronger correlations among items on the same screen than in the version with a single item on each screen. This does not necessarily imply that combining items on one screen leads to improved data quality. The use of this approach should also take into account the type of items being considered and their relationship to each other. Furthermore, combining items on a screen should take into account screen size limitations of different browsers. If using multiple-item screens means that some respondents do not see all items on the screen and need to scroll, the possible benefits of this approach may be offset.

Finally, we examined radio button versus long- and short-text box entry for a particular task. Here we found mixed results. The entry box versions made it easier for respondents to avoid answering the items, which produced more missing data. Those who did complete the task, however, were more likely to do so successfully by making the items total correctly, as we expected. A further test of these alternatives could require an entry in the entry box versions to make them equivalent to the radio button version. Furthermore, the long entry box version resulted in different responses than the short entry box version, suggesting that relatively minor formatting changes could have an impact on the responses to a survey question (see also Couper 2000).

Rather than arguing for one approach over another for all applications (i.e., generic design principles), these results suggest that Web survey design should reflect the particular task at hard. Along with question wording, the presentation of the items in a Web survey can and does provide guidance to respondents on what kinds of answers are being sought, as they often do in other interviewing contexts. Design also affects the efficiency with which respondents complete a Web survey, which may be an important consideration in reducing burden and minimizing incompletes and nonresponse.

V. Conclusions

Web surveys provide many more options for the designer, far exceeding the relatively limited design features of traditional mail surveys. They may be used as powerful tools to guide the respondents through the survey, to motivate them to complete the task, and to provide a rich variety of audio and visual stimuli to enhance the survey questions. Yet, our results suggest that such choices clearly also come with a responsibility. We need to learn how to use these design features judiciously to maximize survey data quality and minimize error. To do so, we need a greater understanding of how such design features influence respondents.

We designed a series of experiments imbedded in a survey with a specific substantive focus; the manipulations we employed were quite modest and designed to represent reasonable options for Web surveys. Our results generally support the hypotheses we began with, suggesting that there are systematic effects of design on the behavior of respondents in Web surveys. Additional research explicitly intended to evaluate the effects observed here and employing other alternatives could produce more information about the impact of Web survey design on response patterns. In particular, something as simple as changing the length of an entry box for a series of otherwise identical items changed the distribution of responses. Future research can also address how much of this can be attributed to differences among individual respondents as opposed to cultural factors.

Self-administered surveys, whether on paper or the Web, rely on both verbal and visual information to communicate with respondents (Redline and Dillman 1999; Ware 2000). While most attention has been paid to the verbal elements of survey instruments (question wording), it has long been recognized that the visual elements may have important effects on respondents' answers. With the rich multimedia capabilities of Web-based surveys, the range of design elements has expanded considerably. Features such as graphics, color, typography, animation, and so on are potentially powerful tools for maintaining respondent interest in the survey and for encouraging completion of the instrument. On the other hand, the presence of these graphic elements may well have an effect on the answers that respondents provide. The fact that self-administered surveys such as these rely primarily on visual means of communication is an important consideration for design. And the effects of such design on respondents are clearly worthy of additional research attention. Web surveys also permit the introduction of sound and video into an instrument, and these possibilities also require research.

These findings also hint at the possible design trade-offs that should be explored. For example, the graphic progress indicator may help to keep respondents motivated to complete a Web survey, but the additional download time associated with such images may have a counteracting effect. This may be true for other visual features designed to maintain respondent interest in the survey. In doing so, these features may distract the respondent from the task of answering the questions accurately and honestly, or they may have other potentially deleterious effects on data quality. Similarly, other design features such as multiple-item screens may be beneficial, as in our study. But they may also require more careful design than single-item screens, given screen limitations and browser variations.

Web surveys bring new challenges for minimizing nonresponse and measurement error. Yet they also provide an enormous opportunity for methodological research and for understanding the role of the instrument in self-administered surveys. As the survey industry embraces the Web for data collection, many important questions remain to be addressed. One of the most important—and interesting—is whether Web surveys are a technological advance in the mail survey format or whether they are an entirely new format with multimedia capabilities that challenge the survey designer and present exciting new opportunities for questionnaire design.

Appendix

Example of E-Mail Invitation

From: Michigan Daily Survey
Sent: Tuesday, March 30, 1999 2:52 PM
To: [e-mail address]
Subject: Michigan Daily Student Survey
Dear University of Michigan Student:

You have been selected at random as part of a sample of 1,600 currently enrolled students at The University of Michigan to participate in a survey organized by The Michigan Daily. We are interested in campus life and in your opinions about some University of Michigan policies that affect students, including affirmative action. We need your answers to make our survey a more accurate representation of student views.

Your participation in the study is completely voluntary, and your answers will not be linked to you by name in anyway. The overall results from the survey will be presented in series of articles in The Michigan Daily.

You can complete the survey whenever it is convenient for you at any computer where you can get access to the World Wide Web. It will take you about 15 minutes.

The information you provide us is very important to the accuracy of our survey. No one from the study staff will be able to connect you to your answers.

As a token of our appreciation, we would like to offer you a gift to thank you for your participation in the survey. When you complete the survey, you will receive instructions about how to claim a copy of The Daily's color book "We're No. 1: The 1997 National Championship Season." Thank you in advance for completing the survey. You are contributing to the discussion of important campus issues that will begin when the survey results are published in The Daily.

Please use MS Explorer or Netscape versions 4 or higher.

Enter this unique CASE IDENTIFICATION NUMBER: [ID]

Use this PASSWORD to verify your eligibility as a study participant: [password]

To start the survey, just go to: http://survey.isr.umich.edu/michsurvey/welcome.htm

If you have questions about the survey or are having trouble signing on to the web site, please contact either one of us.

Michael W. Traugott
Department of Communication Studies
xxxxx@umich.edu
Jennifer Yachnin
The Michigan Daily
xxxxx@umich.edu

References

Comley, Pete. 1997. "The Use of the Internet as a Data Collection Tool." Paper presented at the ESOMAR annual conference, Edinburgh, Scotland.

Couper, Mick P. 2000. "Usability Evaluation of Computer Assisted Survey Instruments." *Social Science Computer Review* 18(4):384–96.

Couper, Mick P., Johnny Blair, and Timothy Triplett. 1999. "A Comparison of Mail and E-Mail for a Survey of Employees in Federal Statistical Agencies." *Journal of Official Statistics* 15(1): 39–56.

Dillman, Don A. 2000. *Mail and Internet Surveys: The Tailored Design Method*. New York: Wiley.

Dillman, Don A., Cleo D. Redline, and Lisa R. Carley-Baxter. 1999. "Influence on Type of Question on Skip Pattern Compliance in Self-Administered Questionnaires." Paper presented at the Joint Statistical Meetings of the American Statistical Association, Indianapolis.

Dillman, Don A., Robert D. Tortora, Jon Conradt, and Dennis Bowker. 1998. "Influence of Plain versus Fancy Design on Response Rates for Web Surveys." Paper presented at the Joint Statistical Meetings of the American Statistical Association, Dallas.

Edy, Jill A., and Michael W. Traugott. 1999. "Affirmative Action in Higher Education: The Students' Perspective." Paper presented at the annual conference of the Midwest Association for Public Opinion Research, Chicago.

Fuchs, Marek, Mick P. Couper, and Sue Ellen Hansen. 2000. "Technology Effects: Do CAPI or PAPI Interviews Take Longer?" *Journal of Official Statistics* 16(3):273–86.

Jeavons, Andrew. 1998. "Ethology and the Web: Observing Respondent Behaviour in Web Surveys." In *Proceedings of the Worldwide Internet Conference* (CD). London: ESOMAR.

Metzner, Helen, and Floyd Mann. 1953. "Effects of Grouping Related Questions in Questionnaires." *Public Opinion Quarterly* 17 (Spring): 136–41.

Parush, Avraham, Ronen Nadir, and Avraham Shtub. 1998. "Evaluating the Layout of Graphical

User Interface Screens: Validation of a Numerical Computerized Model." *International Journal of Human-Computer Interaction* 10(4):343–60.

Redline, Cleo D., and Don A. Dillman. 1999. "The Influence of Auxiliary, Symbolic, Numeric, and Verbal Languages on Navigational Compliance in Self-Administered Questionnaires." Paper presented at the International Conference on Survey Nonresponse, Portland, OR.

Sanchez, Maria Elena. 1992. "Effect of Questionnaire Design on the Quality of Survey Data." *Public Opinion Quarterly* 56(2):206–17.

Schaefer, David R., and Don A. Dillman. 1998. "Development of a Standard E-Mail Methodology: Results of an Experiment." *Public Opinion Quarterly* 62(3):378–97.

Schwarz, Norbert. 1995. "What Respondents Learn from Questionnaires: The Survey Interview and the Logic of Conversation." *International Statistical Review* 63(2):153–68.

———. 1996. *Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation*. Mahwah, NJ: Erlbaum.

Schwarz, Norbert, Fritz Strack, and Hans-Peter Mai. 1991. "Assimilation and Contrast Effects in Part-Whole Question Sequences: A Conversational Logic Analysis." *Public Opinion Quarterly* 55(1):3–23.

Sheehan, Kim B., and Mariea G. Hoy. 1999. "Using E-Mail to Survey Internet Users in the United States: Methodology and Assessment." *Journal of Computer Mediated Communication*, vol. 4(3) (online at http://www.ascusc.org/jcmc).

Smith, Tom W. 1995. "Little Things Matter: A Sampler of How Differences in Questionnaire Format Can Affect Survey Responses." In *Proceedings of the American Statistical Association, Survey Research Methods Section*, pp. 1046–51. Alexandria, VA: American Statistical Association.

Sudman, Seymour, Norman M. Bradburn, and Norbert Schwarz. 1996. *Thinking about Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass.

Terhanian, George. 1999. "Lessons from the Harris Poll Online." Paper presented at the annual meeting of the American Association for Public Opinion Research, St. Petersburg Beach, FL.

Tullis, Thomas S. 1983. "The Formatting of Alphanumeric Displays: A Review and Analysis." *Human Factors* 25(6):657–82.

———. 1988. "A System for Evaluating Screen Formats: Research and Application." In *Advances in Human-Computer Interaction*, ed. H. Rex Hartson and Deborah Hix, vol. 2, pp. 214–86. Norwood, NJ: Ablex.

Ware, Colin. 2000. *Information Visualization: Perception for Design*. San Francisco: Morgan Kaufman.