

Design Issues in Web-Based Electronic Business Surveys

Elizabeth Nichols and Elizabeth Murphy, US Census Bureau;

Kent Norman, Anna Rivadeneira, and Cyntria Eaton, University of Maryland

Elizabeth Nichols; U.S. Census Bureau, Washington D.C. 20233 (email: Elizabeth.May.Nichols@census.gov)

This report is released to inform interested parties of (ongoing) research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Introduction

Researchers are only beginning to create rules for the design of electronic forms (Couper et al. 2001). Some of these design rules originate from Web standards and “best practices.” Many of the initial thoughts about design come from layouts used successfully in paper forms. Visual restrictions and navigational differences between the two modes make one-to-one transformations between the paper layout and the electronic layout difficult.

For individual question types, the electronic mode also offers new design possibilities as compared to the paper mode. For instance, the response options for “choose one” questions can be designed with radio buttons, a drop-down box, or variations on those two designs. Little empirical research has been done on the effect of different design layouts in electronic forms on response accuracy, although we expect there to be these differences just as there are data quality differences when changes are made to visual elements in paper forms (Redline and Dillman 2002). Research on electronic form design (Couper et al. 2004) has focused primarily on design differences when questionnaires collect opinion data or fact-based data not requiring respondents to consult records for answers.

Many of the business survey questions developed by the U.S. Census Bureau require respondents to consult their records to complete the survey accurately. Many of these same business surveys offer respondents an electronic form mode. The record-lookup component might alter the way in which respondents interact with the design of the electronic form, as Norman and his colleagues (2000) found for navigation in their electronic survey study. Consequently, research is needed to determine the impact of different electronic form design decisions for situations mimicking the business survey process, even when some of those design decisions (e.g., radio buttons vs. drop-down lists) have been researched in opinion-oriented contexts.

In this paper we present research testing the effect of alternative design layouts for self-administered electronic business surveys requiring a record-lookup procedure similar to that used by Norman et al. (2000). This research was conducted in a laboratory setting at the University of Maryland using 65 student subjects as proxies for business respondents. We measured data accuracy and respondent burden for each design issue tested in order to identify usable design solutions that will allow establishment respondents to complete electronic forms quickly and accurately with little perceived cognitive burden.

Methodology

Design Issues Studied

Examining current business paper and electronic forms, Census Bureau staff identified eight design issues in electronic business forms.

1. Display and functionality of an automated summing feature in response fields
2. Presenting a matrix (e.g., grid) question/answer format
3. Presenting a response format for “Choose one” questions
4. Customizing question text and instruction display
5. Communicating reporting units
6. Placement of the response field in relation to the question
7. Formatting text entry fields
8. Presenting amount response fields with a separate field to capture “none”

We could find little or no published empirical research on the human performance effects of alternative design solutions to these issues in an electronic mode.

Mock-Business Questionnaire and Records

To test each of these design issues, we created a page-based electronic questionnaire largely in HTML. Pages varied in length based on the number of questions and the length of each question on the page. Therefore, some pages scrolled; others did not. All pages had ANext≡ and APrevious≡ buttons at the bottom left allowing for linear navigation between the pages. A menu available on the left-hand side of the page allowed the respondent to jump between non-consecutive pages. Each menu link reflected the subject matter on the corresponding page.

The questionnaire contained 29 fact-based questions and was modeled after several different U.S. Census Bureau business surveys (e.g., Company Organization Survey, the Quarterly Financial Survey) and censuses (e.g., primarily services and manufacturing). We used a mock questionnaire in this experiment instead of a real questionnaire, since no single, existing form contained questions which would permit the testing of all the design issues of interest.

For each question, we developed two separate designs (A and B), both with the same question wording and order, but differing in layouts. In our judgment, both the A and B designs were reasonable solutions to the design issues. Although we were not constrained by the paper form design, occasionally that design was tested. We used several other guiding principles to achieve our two designs:

- Σ Provide the respondent with a sense of control
- Σ Mimic designs which appeared to work well in the past, or common designs (“best practices”)
- Σ Minimize cognitive burden
- Σ Do not surprise the respondent
- Σ Use standard design principles or what feels natural (e.g., details summed to totals in a top-to-bottom and left-to-right fashion)

The form was titled, “2002 Economic Census Study, Trucking and Warehousing.” The subject matter of the questions was in a logical order, but the questions were not in the same order as the records provided. The electronic records were for a fictitious household moving company named “Move-It, Inc.” In order to complete the questions in the questionnaire accurately, the respondent needed to find the answers in the records. The records were stored in a separate file and displayed on a separate computer monitor from the questionnaire. Records were straightforward and clearly labeled. Although we wanted the respondent to locate the appropriate information easily, we did not want this to be a simple rote task of data entry; thus, we modified records trying to reflect real-life situations in which records do not always match the data requested exactly.

Experimental Design

In order to avoid a confounded design, we attempted to test only one design issue per question and per page. When a page contained more than one question, all of the questions either tested the same layout issue, or were questions that were identical between designs. We used a fractional factorial design to counterbalance any possible confounding effect of one design on a separate design issue. Using eight panels, we randomized whether a page presented the A design or the B design to the test user. When there was more than one page with questions testing the same design issue, each test user saw only one of the designs for the issue. For instance if they saw design A of the Item 1-3 page, then they also saw design A of the Item 4-5 page, since those two pages both had “choose one” questions. Thus, a test user never saw both A and B versions of any design issue.

During the late fall and winter of 2002, 69 undergraduate psychology students at the University of Maryland participated in our experiment to fulfill their course requirement. As test users, they were tasked with completing a business survey for the fictitious moving company. For role playing purposes, before beginning the experiment, each test user was given a “Move-It, Inc.” business card (Figure 1) which included their “new” employee name, “Jamie Doe,” a fictitious telephone number, FAX number, and email address. They were instructed to pretend to be this company employee and to fill out the form using the company’s records. They were also shown how to navigate (using the same mouse) between the questionnaire window and the records window (Figure 2).

After these preliminary instructions were given, test users started the experiment. They signed an online consent form agreeing to be videotaped. They completed a pre-questionnaire asking for their age, sex, experience with the computer (scale 1-10, where 1 is low), experience with the Web (scale 1-10, where 1 is low), and their SAT score. They were randomly assigned to one of the 8 survey panels. After the survey, an electronic post-test questionnaire was administered. It asked

about the ease of use of the different layouts they saw and contained some questions relating to what they remembered about the survey. The experiment took one hour on average to complete. Since we wanted to determine the better of two electronic layouts for each issue regardless of paper, we did not include a paper form.



Figure 1: Business card available to each test user in the experiment.

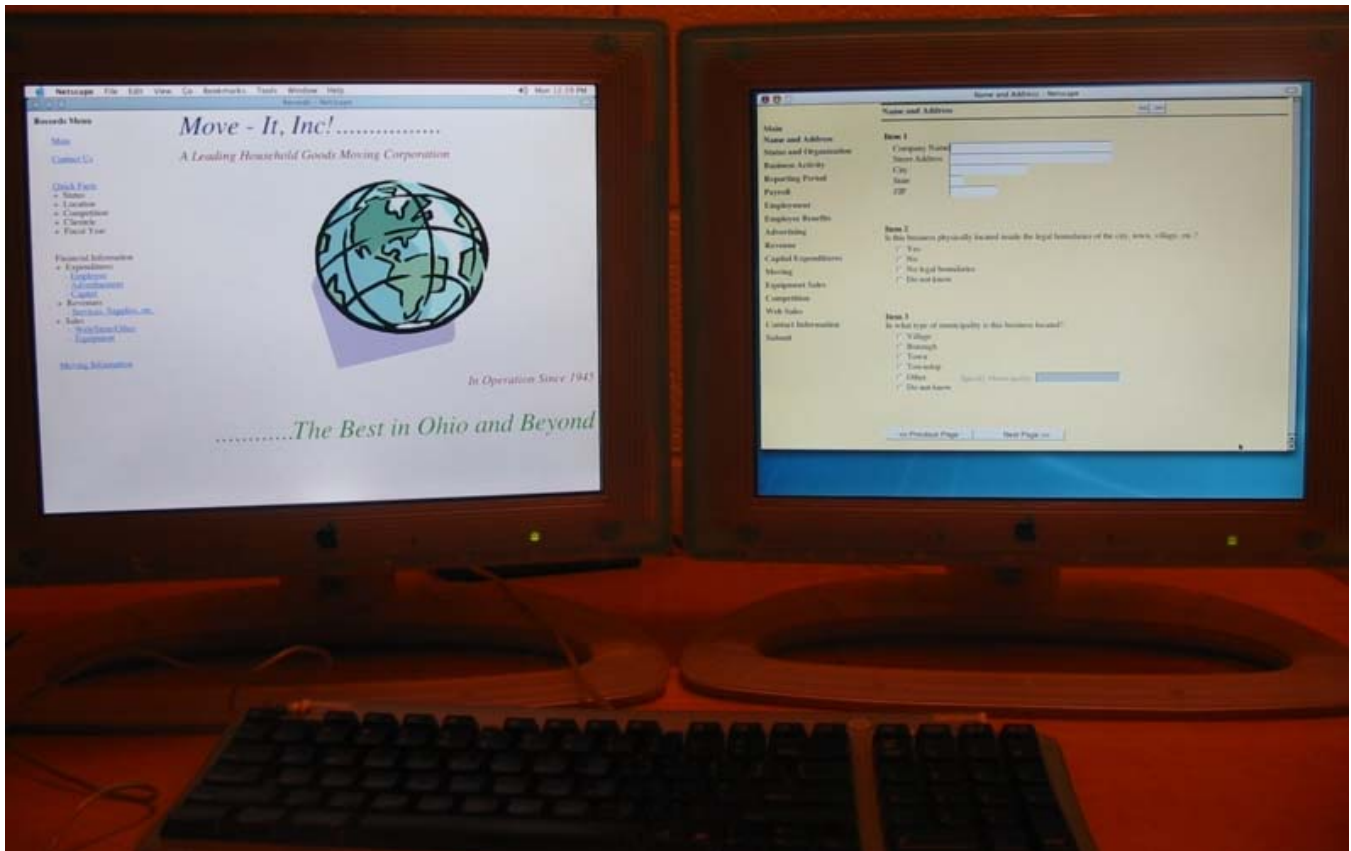


Figure 2: Picture of the two Macintosh monitors and keyboard used by test users at the University of Maryland. The left monitor contains the records, and the right monitor displays the questionnaire. Test users could navigate between the two monitors using a single mouse.

Analysis Method

To determine if one of the designs was better than the other for any of the issues, we compared test users’ performance with design A and B of each page. We measured A better in terms of three different factors:

- Σ the accuracy of the reported data for the question
- Σ the amount of time it took to complete a page in the form

Σ the test user's satisfaction with a particular design for a question

Since we created the records and knew the correct answers, we could calculate accuracy. To compare the response accuracy of design A to design B, we created a score for each question for each test user. Each question included one or more data fields, which were individually scored. If the answer was accurate, the data field was assigned a score of 1; otherwise it was assigned a 0 score. We summed the scores of the appropriate data fields to determine the total score for each question. Using an analysis of covariance model in SAS, we determined whether the difference in total score for each question depended on design (A or B), while controlling for self-reported SAT score (1000-1470), age (17-23), sex (M/F) and an indexed computer/Web usage score (1-20). We also calculated the mean accuracy score per design (A and B). We report the design which elicits the more accurate score based on the mean and based on the coefficient in the model (if it differs from the mean).

For each test user, we calculated the time spent on each page in the form from trace file data. For pages with only one question, our calculation translated into the time per question. For pages containing more than one question, it was impossible to accurately determine when the test user finished one question and started on the next question; thus our most accurate time measure was time per page. For each page visited, we calculated time by subtracting the time a test user entered the page from the time s/he hit a navigational button to exit the page. Since test users could visit pages more than once, we added together all the time spent on all the visits. To adjust for differing record-lookup speeds we subtracted the time that the test user's cursor was in the records. We then performed *t*-tests in SAS to compare the time spent per page for each design, where we assume the lower the mean time per page, the less burdensome the page was for the test users.

We used post-questionnaire data to determine test user satisfaction with the designs. For each of the layout issues, test users were shown a picture of the design they saw. Using a 9- point scale, with the endpoints labeled disagree and agree, test users rated each design they used for the level of their agreement or disagreement with the statement, "The question-and-answer layout was easy to use."

Testing of independence assumptions and data cleaning

Of the 69 test users, we eliminated the four who had the lowest overall percent correct (49-65%). Presumably, these "outlier" test users did not use the records, did not understand the instructions, or did not take the task seriously. No other tester had a percent correct below 70%; the highest percent correct was 93%. We use the remaining 65 students in the analysis.

Each test user was assigned to complete one of the eight panels. Since the designs were assigned equally across the eight panels, over 30 test users used each design (A or B) tested. The percent of correct answers overall was independent of the panel to which the test user was assigned. This means that no particular combination of designs in a particular panel led to higher data accuracy than the combination of designs in another panel. Likewise the mean self-reported SAT score was independent of panel, which suggests a successful random assignment and removes (as much as possible) a potential confounding factor of student ability as measured by the SATs.

To see if there were any potential effects on accuracy other than the designs, we modeled the students' overall correct score against the demographic variables of sex, age, computer/www usage, and SAT score. At $p < 0.01$, the self-reported SAT score was a significant predictor of accuracy. The SAT score coefficient in the model was positive, which is interpreted as follows: The higher the SAT score of the student, the greater the number of correct responses the student provided in the questionnaire. This suggests that the competencies of comprehension, record lookup, and decision making needed to complete the questionnaire are similar to the competencies needed to perform the tasks in the SAT.

Limitations

As is true for other laboratory experiments, the extent to which we can generalize from these results is limited by the sample size, the business respondent proxies we used (students), and the constraints we placed on the experiment (e.g., no paper form). We strongly believe results might differ in situations where an available paper form deviates from the look-and-feel of the electronic form, as is the case in many current Census Bureau surveys and censuses. Results might have also differed if we had not used the record-lookup component and had instead relied on memory or respondent opinion when answering questions.

We are less concerned by our use of students as proxies for business respondents. Students have many similarities to business respondents including similar educational experience (some college), similar motivation (both groups are required to perform the task), and potentially similar work experience (sometimes a new employee is given the Census-form task in

businesses). Secondly, we tested 8 non-students with some administrative and/or business experience. Their means were within 2 standard deviations of the student means for 26 of the 29 questions, suggesting that the students and the nonstudents performed similarly with the designs.

Our burden measures (time and satisfaction) are limited for several reasons: We objectively measured time spent per page, yet perceived time is probably a better indicator of burden for actual respondents. We do not know how strongly perceived time is related to objectively measured time. As measured by the ease-of-use questions, subjective satisfaction might be perceived differently by actual respondents in real circumstances.

Finally, the questionnaire and records underwent three rounds of review and correction prior to the experiment. This included debugging, cognitive interviewing, and timing the experiment. We did not want any computer or cognitive aspects of the experiment to confound the design comparisons. Unfortunately, after the experiment was run, we found a few wording and visual confounding differences between the A and B designs. We either mention those differences in the results section, or we eliminated the affected data fields from the analysis.

Results

In all of the analysis and tables provided in this section, we present both significant and nonsignificant differences in accuracy, speed and satisfaction. The design with the higher accuracy score is listed as the more accurate design, even if significance testing did not find a statistically significant difference. We calculated the mean accuracy score, and we modeled accuracy with the design (A or B) as an independent variable. Occasionally, the mean score and the coefficient associated with the design variable in the model point in opposite directions. For example, sometimes the mean will suggest that design A elicits the more accurate response; while the coefficient (adjusting for other demographic variables: SAT score, sex, age and computer/Web usage) suggests that design B elicits the more accurate response. In the tables, we make note of any such discrepancy between the mean and the coefficient. The design with the lower mean time per page is listed as the faster design, even if significance testing did not find a statistically significant difference between the means. The design with the higher mean for ease of use is listed as the more satisfying design, again, even if the difference between the means was nonsignificant. Significant differences are indicated with a * and the corresponding p value. We present all these data so the reader has a complete picture of what we found, but we caution against drawing conclusions based on the nonsignificant results.

Display and Functionality of an Automated Summing Feature in Response Fields

Often in business survey forms, respondents are asked to sum a set of numerical responses to a total. Sometimes the total of these sub-totals is also requested. When translating this paper-based approach into an electronic form, automating the manual summation procedure is a logical step. If programmed, an electronic form can automatically sum entered data into sub-totals and totals. Some record-keeping practices in businesses hinder what would be a strict automated summing procedure, however. Situations arise where respondents don't know the specific amounts requested by a form, but know the total, or they might know only some of the specific amounts. Other situations arise where the breakdown of amounts in a respondent's records might not match the specific amounts requested on a form.

We tested two different automated summing designs to try to account for these situations, as well as the situation where the breakdown of amounts in a respondent's records matches the specific amounts requested by the form. In both the A and B designs, any entry into a specific field (e.g., the individual lines for kinds of revenue in Figure 3a and Figure 3b) was automatically summed into the appropriate sub-total and/or total fields. How the respondent overwrote the sub-total and totals differed between the two designs.

In design A, any entry into a specific field was automatically summed into the appropriate total field as shown in Figure 3a. Since the total and sub-total fields were disabled as the default, the test users could not click into the field to change the total value. To change the value in the total field, test users first needed to manipulate (click on) a calculator icon next to the sum field. (We refer to this calculator icon as a switch.) Clicking on the switch disabled it and enabled the sum field, which the test user then could modify as shown in Figure 3b. If the switch was later enabled (clicked on again), the automated sum would overwrite whatever the test user typed in that field.

Revenue from Truck Rentals (excluding moving staff)	
26 ' mover.....	\$ 1,234
24 ' mover.....	\$
17 ' mover.....	\$
14 ' mover.....	\$
10 ' mover.....	\$
Cargo van.....	\$
Pickup.....	\$
Total Truck Rental Revenue.....	\$ 1,234




Figure 3a: Design A,
Default automated summing (the control switch is on)

Revenue from Truck Rentals (excluding moving staff)	
26 ' mover.....	\$ 1,234
24 ' mover.....	\$
17 ' mover.....	\$
14 ' mover.....	\$
10 ' mover.....	\$
Cargo van.....	\$
Pickup.....	\$
Total Truck Rental Revenue.....	\$ 4,321




Figure 3b: Design A,
Sum modified after turning off the control switch

In design B, shown in Figures 4a and 4b, the total and sub-total fields summed automatically based on data typed into the specific fields. These total fields were not disabled, so at any time, test users could enter into the sub-total fields and change data. If, however, the test user reentered one of the detail fields that summed to the sub-total, the sub-total and total recalculated automatically.

Revenue from Truck Rentals (excluding moving staff)	
26 ' mover.....	\$ 1,234
24 ' mover.....	\$
17 ' mover.....	\$
14 ' mover.....	\$
10 ' mover.....	\$
Cargo van.....	\$
Pickup.....	\$
Total Truck Rental Revenue.....	\$ 1,234.00

Figure 4a: Design B, Default automated summing

Revenue from Truck Rentals (excluding moving staff)	
26 ' mover.....	\$ 1,234
24 ' mover.....	\$
17 ' mover.....	\$
14 ' mover.....	\$
10 ' mover.....	\$
Cargo van.....	\$
Pickup.....	\$
Total Truck Rental Revenue.....	\$ 4,312

Figure 4b: Design B, Modified sum

Two questions in our business survey tested the automated summing layouts. In one question (Item 10 in the questionnaire), the records matched the form data fields exactly. Test users had to simply transcribe the amount from the records into the appropriate data fields. If done correctly, the automated sum would be correct. There was no need to manipulate the switch or the sum.

The other question (Item 16 in the questionnaire) was modeled after a question in the 2002 Economic Census. This question asked the test user to report the revenue for five different moving categories. After each category a subtotal was requested, and at the end of the page a total of the subtotals was requested. Figures 3 and 4 show one of the moving categories for that question. In the records, the answers were in the same order as the responses requested by the form, but the records did not match the data fields on the form exactly. For example, the records corresponding to the truck rental revenue category in Figures 3 and 4 contained only a total truck rental revenue. This total was not broken down into any detail in the records. For other moving categories a total with a partial breakdown of the total into specific detailed amounts was available, but the detail provided did not match the specific data fields on the form exactly.

No instructions were provided in either design informing test users what to do if the records did not match the information requested in the survey. Test users were left on their own to figure out how to report. Our rules for assigning accuracy were relatively simple: (1) If a breakdown of amounts was available in the records that matched the specific fields on the form, the test user needed to enter the detailed amount in the appropriate specific field on the form. (2) Any detail in the records not matching exactly to a specific data field on the form should have been excluded from the individual data fields on the form, but included in any corresponding subtotal amount. (3) We had one exception to rule (2). If the test user explained in a note any non-matching specific field, we treated the value in the data field as a correct response. Only a few test users did this, since there was not a separate field to write the explanation, and instead the test users put the explanation in the same field as

the value. (4) When only a total was provided in the records, the test user needed to place the total in the appropriate sub-total field, but not in any of the specific fields which summed to the subtotal. Using these accuracy rules, one could assume the each specific data field accurately reflected what was in the records and that the totals were accurate. In a real-world situation, a clerical follow-up operation would be needed for the Census Bureau to determine why subtotal amounts did not equal the specific detail reported in the form.

As shown in Table 1, no significant differences were found in either accuracy or speed between the two designs for the two questions that differed by the type of automated summing design used. \

Table 1: Results of Automated Summing Design
 Design A=With the switch (n=32); Design B=No switch (n=33)
 N.S.=not significant

Question	More accurate**	Faster
Question where records matched form data fields exactly (Item 10, Employment question)	No switch (N.S.)	No switch (N.S.)
Question where records did not match form data fields exactly (Item 16, Revenue question)	Switch (N.S.)	No switch (N.S.)

**The analysis of mean score and the coefficient associated with the analysis of covariance model point in the same direction.

In addition to capturing information to compare the speed and accuracy of the two designs, the instrument also captured how, when, and where test users manipulated the control switch in design A. We worried that test users might not realize the functionality of the control switch and instead assume it was only an icon. (As previously mentioned, the control switch was a calculator button located to the right of sub-total fields, as shown in Figures 3A and 3B.) We found that some test users never manipulated the switch. None of the 32 test users who received design A turned off the switch in the employment question (Item 10). We expected this since test users did not need to manipulate the sum field. Interestingly, 12 of the 32 or 38% of the test users never turned off any of the five switches in the revenue question (Item 16). To answer the revenue question accurately, at least a few of the switches should have been turned off. Based on two other pieces of information, we suspect that these test users did not know that the switch was functional.

Our first clue that some test users might not have been aware of the switch functionality was the amount of time those 12 test users spent on the first page of the survey, which contained instructions explaining the switch. The first page of the survey, which we refer to as the Welcome page, provided an explanation of the switch for those test users receiving the design with the switch (Figure 5). A reminder was also present at each of the two questions containing the switch (Figure 6). (There was no explanation of the automated summing feature for the test users receiving design B.) The average time spent on the Welcome page by the twelve test users who never used the switch was 11 seconds, while the average time spent on the page by 18 test users who used the switch was 17 seconds. (We eliminated two test users who used the switch from this calculation. They appeared to be outliers in the amount of time they stayed on the Welcome page, staying approximately 9 minutes and 2 minutes respectively.)

Although the difference between 11 and 17 seconds is not significantly different, it is interesting that those who spent an average of 17 seconds on the Welcome page were more likely to use the switch. Ignoring the banner information, this page has approximately 123 words. Rayner (1998) would estimate that a normal reader would spend 29 seconds on this page, a skilled reader would spend 24 seconds on the page and a skimmer would spend 10.6 seconds on the page, which is very close to the mean score of 11 seconds found by those who didn't use the switch function. Those who are skimmers are less likely to retain information than those who read thoroughly. We hypothesize then that the test users who skimmed the Welcome page were less likely to understand the functionality of the switch.

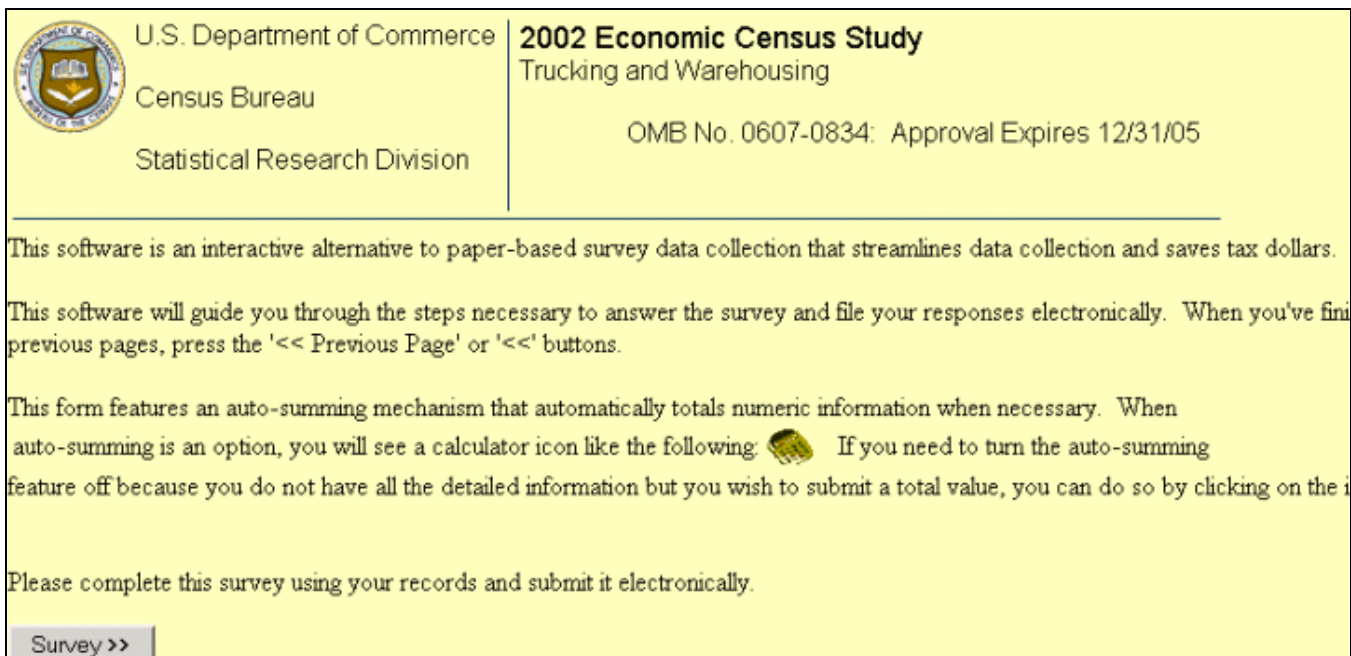


Figure 5: Introductory screen for survey containing design A of the automated summing design.

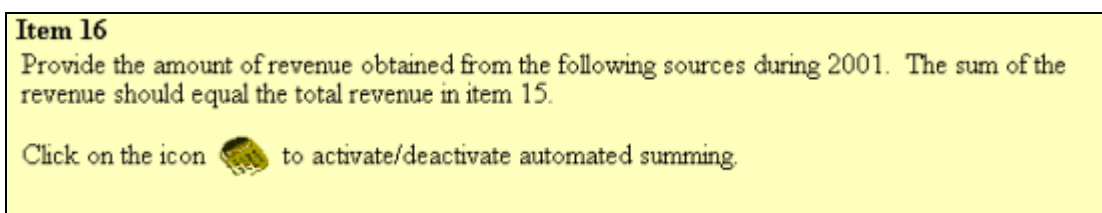


Figure 6: Example of the reminder note on one of the two questions containing the automated summing feature in design A.

Our second clue to possible lack of awareness of switch functionality came from post-questionnaire survey data. Test users receiving the design with the switch were asked to rate the clarity of the operation of the automated summing and the helpfulness of the presence of the automated summing (see Figure 7). With a 9-point scale for each, both scores were extremely positive with 8.63 for operation of the automated summing and 8.36 for presence of the automated summing. The scores of the 12 test users who never used the switch did not differ significantly (at 8.75 and 8.18) from the scores of the test users who used the switch at least once. Since the scores are approximately the same as those scores given by test users who used the switch, we conclude that the 12 test users did not include use of the switch into their rating. If they had, we would have expected a lower score or no score for the rating of the automating summing. We suspect that these 12 test users did not factor in the use of the switch when rating the automated summing feature of the questionnaire, presumably because they didn't know the switch functioned.

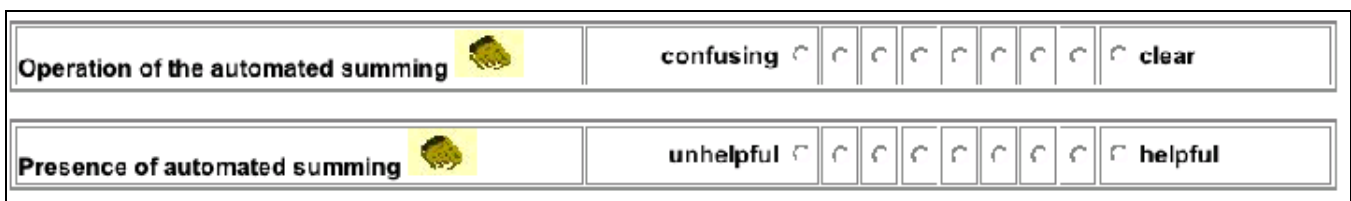


Figure 7: The two post-questionnaire satisfaction survey items pertaining to the automated summing in design A.

To understand how the designs could be further improved, Table 2 presents the distribution of types of errors test users made on the five moving categories in Item 16. The records did not match the data fields exactly for Item 16. We defined four major types of errors as follows:

- Error type 1= The total entered on the form was correct, but at least one specific field contained an incorrect amount. This happened when the record amounts were forced into an incorrect field on the form or when test users did not enter any specific amounts since the categories on the form did not match the description of the amounts in the records exactly.
- Error type 2= All the specific amounts entered on the form were correct, but the total on the form was less than the true total in the records. This happened when the respondent records contained information not specifically requested by the form. Ideally, respondents should add any additional amounts to a total.
- Error type 3= We believe test users made a reading error since the detail was placed into data fields in ways suggesting test users did not compare labels from the records to the form. For example, a test user entered the first five values from the records into the first five fields on the form, not realizing that they entered the total into the wrong field. In this situation the form automatically adds the five values together to create another total. This total is higher than the true total in the records.
- Error type 4= Item nonresponse occurred in either the detail and/or the total, presumably due to test user frustration with mismatches between the records and the form.

Table 2: Percent of errors distributed by error type for each of the five separate categories in Item 16. Item 16 provided the instructions, “Provide the amount of revenue obtained from the following sources during 2001. The sum of the revenue should equal the total revenue in item 15.”

5 Moving Categories	Records Contained	Error type 1	Error type 2	Error type 3	Error type 4	No Error
1 st category**	Data for 3 fields on the form, an amount which did not match the form exactly, and a total	48%	9%	3%	0%	40%
2nd category	Total only	22%	0%	3%	0%	75%
(7 fields with a total)						
(See figures 3 and 4)						
3rd category	No records	0%	0%	3%	0%	97%
(4 fields with a total)						
4 th category	Total only	18%	0%	0%	9%	73%
(4 fields with a total)						
5 th category	Data for 2 fields on the form, an “other” amount (not on the form), and a total	57%	12%	0%	3%	28%
(4 fields with a total)						

**Only design B results are analyzed since there was a slight error in the display of design A. Data (design A and B) from the first category were also excluded from speed and accuracy results in Table 1.

We conclude from this analysis that test users were very likely to make an error when the records contained some specific fields present on the survey, but also contained other amounts which did not fit into the specific fields, as shown by the first and fifth moving categories. When presented with such records, test users most often added any extra detail from records into a left-over specific field on the form. When only a total was in the records, about 20% of test users put the total into a specific field as shown by the error types generated by the second and fourth moving categories. Most often they put the total in the first field. In each of these cases, although the correct subtotal was created, there were incorrect amounts in some of the specific fields. Any post-data-collection editing would most likely not discover the error since the individual fields summed correctly to the total. An improved design would allow test users to enter all the detail they had, but into appropriate data fields. Adding an Aother field≅ to the end of each sub-total section seems like an appropriate place. Adding space for

the test user to explain their Aother≡ entry would provide information in order to update the forms in the future. More importantly, it would allow for more accurate data analysis.

In conclusion, the addition of the automated summing switch did not seem to help test users answer faster or more accurately. It was not a detriment except for those test users who did not spend time reading the welcome page containing the explanation of the functionality of the switch. For that reason alone, we refrain from recommending additional steps (such as the switch) that need explanation. If we assume our test users represent actual respondents, this experiment confirmed the assumption that a portion of respondents will not take the time to read and comprehend instructions. We also learned from the automated summation exercise that errors are likely when respondents have relevant information that does not fit perfectly into the form. Many of the test users forced this relevant information into the form and, by doing so, contaminated some data fields. Our data found that between 18 to 58% of test users made this type of error. The solution is to have no-fault data fields allowing the respondent to communicate their information without contamination of other data fields.

Presentation of response fields in a matrix

Many paper forms use a matrix or grid layout to convey the information needed. Matrices save paper real estate, and data relationships are easily seen since respondents can scan column entries. On the negative side, Jenkins and Dillman (1993) provide some examples suggesting respondents have more trouble answering matrix-question formats than single-question formats. In transitioning to the computer, form designers should be aware that large matrices cause even more potential problems for users. Many matrices used on a paper form do not fit on the screen without requiring the respondent to use the horizontal scroll to see all the columns. User-interface design guidelines (e.g., Koyani et al. 2003; Galitz 1993) recommend against the use of horizontal scrolling. There is the potential for item nonresponse if respondents do not see the need to horizontally scroll and do not complete the columns to the right of what is immediately in view. Depending upon the size of the matrix, if the column and row headings do not move, we hypothesize that respondents might lose their place toward the lower right cell in the matrix where headings are not visible after scrolling.

We tested a matrix response layout (Figure 8) against what we called a “stacked design” (Figure 9). For the stacked design, each column or row (whichever was more appropriate) in the matrix was vertically oriented, and the headings were stacked on top of one another to create one long vertical column, with repeats of headings where necessary.

Two questions tested this layout difference. In one of the questions, no scrolling (either vertically or horizontally) was needed in either design. The records contained an answer for each of the data fields in this question. The answers were not in the same order as the questions in either the matrix or the stacked design. In the other question (see Figures 8 and 9), the matrix design scrolled horizontally, and the stacked design scrolled vertically. For this question, the records contained information for three of the data cells. Here, too, the answers were not in the order of either form design.

Item 23 What was the revenue received from the sale of these types of equipment?						
	January	February	March	April	May	June
Trucks						
Trailers						
Other Equipment						
Total						

Figure 8: Design B, Matrix design with horizontal scrolling (horizontal scroll bar is not shown)

Item 23
What was the revenue received from the sale of these types of equipment?

Trucks

January..... \$

February..... \$

March..... \$

April..... \$

May..... \$

June..... \$

July..... \$

August..... \$

September..... \$

October..... \$

November..... \$

December..... \$

Trailers

January..... \$

February..... \$

March..... \$

April..... \$

May..... \$

June..... \$

Figure 9: Design A, Stacked design with vertical scrolling (vertical scroll bar is not shown)

The data confirmed the difficulty that we suspected test users would have in using the horizontally scrolling matrix layout. As shown in Table 3, when the matrix scrolled horizontally, as it did in the question shown in Figure 8, test users performed significantly more accurately with the stacked design ($p < 0.10$). There was no significant difference in speed between designs, nor was there any difference in the observed ease-of-use (satisfaction) rating between the designs.

Table 3: Results for Stacked and Matrix Designs
Design A = Stacked design (n=33); Design B = Matrix design (n=32)
* = significant; N.S. = not significant

Question	More accurate**	Faster	More satisfied
No scrolling necessary in either design (Item 19)	Stacked (N.S.)	Stacked (N.S.)	Not measured
Horizontal scrolling needed in matrix design; vertical scrolling needed in stacked design (Item 23)	Stacked (* $p < 0.10$)	Matrix (N.S.)	Stacked (N.S.)

**The analysis of mean score and the coefficient associated with the analysis of covariance model point in the same direction.

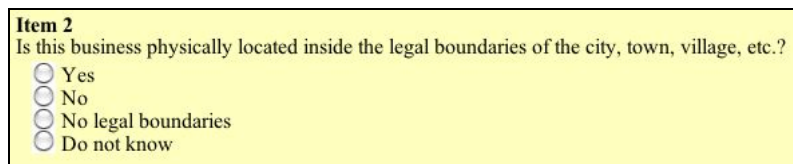
Most of the inaccuracies occurred when test users placed data in the far bottom right cell of the matrix for the question where the matrix scrolled horizontally off the screen. There were two instances of nonresponse, and three test users placed the entry into an incorrect field. In this design the column and row headings did not move as the test user scrolled vertically and horizontally. The incorrect cell placement could be attributed to not seeing the row headings. We hypothesize that most test users probably realized there were more columns in the matrix than just those visible on the computer screen initially. The columns were the months of the year, a well-known sequence. A better test would be to have column headings that are not sequential or known. Even so, these results are interesting since they show nonresponse even with familiar, sequential column headings and a horizontally scrollable matrix.

Presenting a response format for “Choose-one” questions

Survey respondents are often offered a set of possible responses and asked to choose one of those responses. Sometimes the list of responses is relatively short and sometimes it is not. Using the paper mode, it is very easy for the respondent to choose more than one response, either intentionally or unintentionally. The electronic questionnaire can be designed to allow only one response, thus avoiding this post-collection data analysis problem.

A Choose-one questions with a short set of response choices

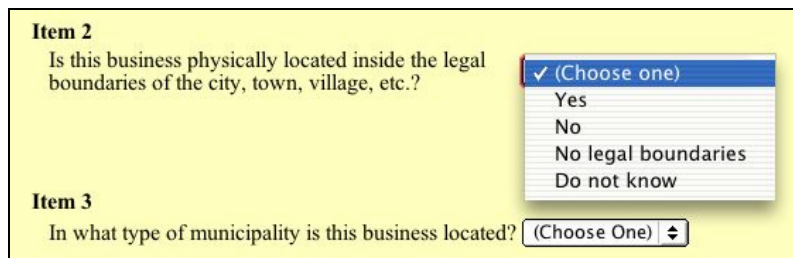
In design A (Figure 10), radio buttons were used for the choose-one questions with fewer than six response choices. One advantage of radio buttons is that all response options are visible. With one click an answer is selected. Respondents can easily change their answer by clicking another response. The disadvantage is that respondents still have to visually scan the list of response options to see which one they chose if verifying their selection, and a longer list of options could require a vertically scrollable page. If a selected option has scrolled off the page, some respondents may not realize that their previously selected answer will be deselected when they select another radio button. A design goal is to minimize scrolling to reduce respondent burden.



Item 2
Is this business physically located inside the legal boundaries of the city, town, village, etc.?
 Yes
 No
 No legal boundaries
 Do not know

Figure 10: Design A, Radio Buttons

In design B (Figure 11), drop-down lists were used for those same choose-one questions. With this design, the response options are not immediately visible. Respondents have to click on the small arrow to open the list and then click on the list of possible responses to make a selection. (Item 2 in Figure 11 shows an open list, while Item 3 shows a default display.) Thus, two clicks are necessary to make a selection. Once a selection is made, the list closes up, and the selection is the only text visible. Unlike a set of radio buttons, a drop-down list quickly communicates the selected answer to the respondent since that is the only text visible. One common reason for using a drop-down list is to save screen real estate.



Item 2
Is this business physically located inside the legal boundaries of the city, town, village, etc.?

Yes
No
No legal boundaries
Do not know

Item 3
In what type of municipality is this business located?

Figure 11: Design B, Drop-down list

Four questions tested these two designs. The answers to these four questions were in text format and on a single page of the records. Unintentionally on our part, test users had difficulty locating the answer to one of the questions. Even so, we see in Table 4 that there were no significant differences in accuracy between the two designs. Test users appeared to spend less time per page with the radio buttons than they spent on the page with the drop-down list. A significant difference in time was found between radio buttons and drop-down lists, in favor of radio buttons. There was no significant difference in the satisfaction rating given to radio buttons as compared to the rating given to drop-down lists.

Table 4: Choose-One Designs for a Short List of Response Options
 Design A= Radio button (n=33); Design B=Drop-down list (n=32)
 *=significant; N.S.=not significant

Question	More accurate**	Faster	More satisfied
Physical location question 4 response options (Item 2)	Neither (N.S.) (all test users answered correctly)	Radio button (* p<0.05) (page included both Items 2 and 3)	Drop-down list (N.S.)
Municipality question 6 response options and a write-in (Item 3)	Radio button (N.S.) Mean score and coefficient both point in the same direction		Not measured
Status question 4 response options and a write-in (Item 4)	Mean points to radio button (N.S.) Coefficient points to drop-down (N.S)	Radio button (N.S.) (page included both Items 4 and 5)	Not measured
Organization question 5 response options and a write-in (Item 5)	Mean points to radio button (N.S.) Coefficient points to drop-down (N.S)		Not measured

**The analysis of mean score and the coefficient associated with the analysis of covariance model DO NOT point in the same direction for two of the items.

A Choose-one ≅ questions with a large number of response choices

Although not as common as choose-one questions with a few response options, business questions occasionally ask respondents to choose an answer from a long list of response options. A long list of response options poses different problems for respondents. In general, navigating a list to make a response selection or change a selection, without losing one's place, is difficult. This problem is potentially compounded if the response options themselves are lengthy, if the long list is not organized, or if the list is so long that it spans more than one paper page. Because these issues differ from the issues for a short list of response options, we tested different designs, especially since user-interface design experts do not recommend using drop-down list for more than 8 response options (e.g., Galitz, 1993).

In design A, an expandable list design was used because the response options tested were organized into larger topic headings and then sub-categories. The default display was the five larger topic headings alone. If the test user clicked on the box next to a topic heading, the subcategories were displayed under that main heading as shown in Figure 12. From there, the test user could select a subcategory using the adjacent radio button. The test user could also change their selection at any time and open and/or close any of the larger topic headings in any order. Their selection was displayed in a field at the bottom of the page.

Item 6
Which specific sub-category best reflects your company's principal kind of business or activity in 2001?

To find your answer, click on one of the categories listed below to access the appropriate sub-category:

Warehousing and storage facilities

Courier and messenger services

Truck transportation

Household goods moving.....	<input type="radio"/> 4219101
General freight, truckload (TL).....	<input type="radio"/> 4219211
General freight, less-than-truckload (LTL).....	<input type="radio"/> 4219221
Solid waste collection.....	<input type="radio"/> 4212611
Hazardous waste collection.....	<input type="radio"/> 4212621
Other waste collection.....	<input type="radio"/> 4212691
Dump trucking.....	<input type="radio"/> 4212701
Hazardous materials trucking (except waste).....	<input type="radio"/> 4219301
Agricultural products trucking.....	<input type="radio"/> 4219401
Log hauling (NO cutting).....	<input type="radio"/> 4219402
Specialized trucking.....	<input type="radio"/> 4219901

Other transportation-related activities

Other kind of business or activity

You have chosen: _____

Figure 12: Design A, Expandable box. The default display contains only the five main categories. The figure shown has the third category expanded for the selection of one of the 11 options shown.

Design B used a long scrollable list of all the choices with topic headings (Figure 13). The headings and response options were identical between the two designs. There were five links at the top of the page, which corresponded to the five main categories. When a hyperlink was clicked, the screen jumped to that topic heading and all of its subcategories. The test user could also scroll the entire page without using the links. After each group (topic heading and subcategories), there was a top and bottom link to aid in navigating the page. As in design A, the test user's selection was displayed in a field at the bottom of the page (not shown in Figure 13).

Item 6
Which specific sub-category best reflects your company's principal kind of business or activity in 2001?

To quickly find your answer, use one of the links below to jump to the appropriate category:

- [Warehousing and storage facilities](#)
- [Courier and messenger services](#)
- [Truck transportation](#)
- [Other transportation-related activities](#)
- [Other kind of business or activity](#)

Warehousing and storage facilities

Cotton and linters.....	<input type="radio"/> 4221001
Grain elevators, storage only.....	<input type="radio"/> 4221002
Other farm products (except cold storage).....	<input type="radio"/> 4221003
Refrigerated products (except fur storage).....	<input type="radio"/> 4222001
Fur storage.....	<input type="radio"/> 4226221
Self-service or mini-warehousing.....	<input type="radio"/> 4225201

Figure 13: Design B, Long radio button list (partial display)

This question was modeled after the Kind-of-Business question on the Economic Census. The numbers to the right of the radio buttons were not meaningful to the test users in our experiment, but are associated with an industrial classification code and might aid real respondents filling out actual economic census forms. In our experiment, we provided a text answer to this question in the first sentence of the first page of the records. We also thought the name of the company, AMove-It, Inc. might help test users identify the business as a household moving company. Unfortunately, this proved to be a difficult question to answer accurately. Results in Table 5 show that no design was significantly better in terms of accuracy. With all the demographic characteristics in the model, SAT score was significant: The higher the test user's SAT score, the more accurately the test user performed on this question. The model with an interaction between design and SAT score suggests a possible tendency for those with both high and low SAT scores to perform more accurately with the long scrollable list, but this tendency was not significant (not shown in table). There was no significant difference in speed either. Interestingly, test users gave the expandable list a significantly easier-to-use score. This combination of results highlights the tension between

designs that take less time and seem easier to use, but potentially produce less accurate data. A goal for designers of electronic forms is to improve the ease of use of the scrollable design or improve the accuracy of the expandable list design.

Table 5: Designs for a long list of choose-one response options
 Design A= Expandable list (n=32); Design B= Long scrollable list (n=33)
 *=significant; N.S.=not significant

Question	More accurate**	Faster	More satisfied
Kind-of-business question (Item 6)	Long scrollable list (N.S.)	Expandable list (N.S.)	Expandable list * (p<0.10)

**The analysis of mean score and the coefficient associated with the analysis of covariance model point in the same direction.

In addition to the overall measure of accuracy, we also wanted to determine if any of the specific navigational aids within the designs led to accurate reporting. In design A, test users had to open at least one of the five topic headings to select an answer. We suspected this design might quicken the survey response by allowing the test user to navigate quickly to the appropriate topic heading. We didn't know if the additional step of opening up the topic headings by clicking on the box would improve data accuracy, however. In design B, the test user could select links at the top of the page to navigate down the page, or the test user could scroll down the page. There were also "top" and "bottom" links after each topic and subcategory group within the page. We thought that the links at the top could help the test user navigate the page easily and quickly, but we didn't know if they would use these links because the test user could also just scroll down the entire screen. We added the "top" and "bottom" links within the page since we had seen that design used in other Web sites. We also didn't know whether use of any of these links would aid data accuracy.

We collected data from the test sessions on the number of times each of the boxes in design A was selected and the number of times the links in design B were used. In design A, we found that on average test users selected between 2 to 3 of the boxes. Thus most test users did not open each topic heading to see the individual answer categories. In design B, we found that on average test users used one of the links at the top of the page. Only 4 of the 33 test users did not use a link, while 22 used a link only once, so most test users used a link at the top to navigate down the page. On the contrary, we found few test users used the links within the page. No one used any of the "bottom" links, and only two test users used a "top" link. Thus, these additional navigational links within the page did not seem to assist test users, and we do not recommend them for future long scrollable lists.

To determine if the checkbox or link use had an effect on data accuracy, for each design, we modeled accuracy by the number of checkboxes/links used and whether the test user selected the correct checkbox/link initially (models are not shown). One fear we had with each of these designs was that test users would select a single checkbox or link and then not deviate from their initial selection. In design A, we did not find a significant relationship between the number of checkboxes used and accuracy. Thus, test users who opened only one topic heading were just as likely to answer correctly as those who opened four topic headings, but if test users initially selected the correct topic heading box, they were more likely to answer correctly. Thus, it seems as if test users were partial to the answer categories in the topic heading they initially opened.

In design B, a similar conclusion was reached. There was a significant inverse relationship between the number of links used and accuracy: Use of fewer links is associated with better data quality. And, similar to design A findings, if test users initially selected the correct link, they were more likely to answer correctly. These data suggest that if either of the designs is used, survey designers should make sure that respondents are selecting the correct topic initially. This translates into making sure that there is a clear and mutually exclusive relationship between the answer categories and their assigned topic headings. Now that we know test users use the short-cut navigational features, we recommend comparing these designs to a design without short-cuts. It could be that any reduction in respondent burden with the short-cut results in lower data quality.

Customizing question text and instruction display

Unlike paper forms, electronic forms can use automatic fills to personalize text on the form. For example in demographic-interviewer-assisted forms, once the sex of the person is known, additional questions can be phrased using the appropriate pronoun - his or hers, he or she, and so forth.

Using this same technique, instructions, which might be overlooked otherwise, can be built into the question. For example, many Census Bureau business surveys request calendar year information, and there is an instruction stating so on the forms. If a business keeps its books in fiscal-year terms, the respondent should transform the data into calendar year figures for reporting purposes. If respondents, however, do not see, read, or remember the instruction, they very likely will complete each question using their fiscal year data. Instead of relying upon the instruction at the top of the form, survey designers might find that building this instruction into the question improves data accuracy.


We tested a similar concept. In one question early in the form (Item 7), **all** test users received an instruction requesting that fiscal year data be reported. The instruction, “Please answer remaining questions based on your Fiscal Year” popped up after test users answered a question asking them whether their company uses a fiscal or calendar year record-keeping system as shown in Figure 14. Another question (Item 17) later in the questionnaire asked test users to report capital expenditures. In the design A version of this question (Figure 15), the words Afiscal year≅ were added to the question text, AWhat capital expenditures did your business have in fiscal year 2001?≅ In design B (Figure 16) those words were not included and instead that question read, AWhat were the capital expenditures your business had in 2001?≅

Item 7
 What reporting period does your business use?

Calendar Year
 Fiscal Year From: / To: /

Please answer remaining questions based on your Fiscal Year

Figure 14: All test users received this question. Popup instruction “Please answer remaining questions based on your Fiscal Year.” appears after respondent selects Fiscal Year from the radio button choices. (The records indicated that the company kept its books on a fiscal year.)

Item 17
 What capital expenditures did your business have in fiscal year 2001?
 Click on the icon  to activate/deactivate automated summing.

	New		Used		Total
Trucks	\$	<input type="text"/>	\$	<input type="text"/>	\$ <input type="text"/>
Trailers	\$	<input type="text"/>	\$	<input type="text"/>	\$ <input type="text"/>
Computer Equipment	\$	<input type="text"/>	\$	<input type="text"/>	\$ <input type="text"/>
Other Equipment	\$	<input type="text"/>	\$	<input type="text"/>	\$ <input type="text"/>
Total	\$	<input type="text"/>	\$	<input type="text"/>	\$ <input type="text"/>



ON

Figure 15: Design A: Question text contains the reminder to report for fiscal year. This was not dynamic text. Instead, all test users saw the same question text.

Item 17

What were the capital expenditures your business had in 2001?

	New	Used	Total
Trucks	\$	\$	\$
Trailers	\$	\$	\$
Computer Equipment	\$	\$	\$
Other Equipment	\$	\$	\$
Total	\$	\$	\$

Figure 16: Design B: Question text does not contain the reminder to report for fiscal year. All test users saw the same question text.

The records for Move-It, Inc. provided fiscal year information for all appropriate topics. Data for the capital expenditures of Move-It, Inc. included both fiscal year and calendar year data. The title of the first screen (see Figure 17a) containing capital expenditures states that the records are for the calendar year, "Capital Expenditures for Calendar Year 2001". To see the fiscal year capital expenditures data, test users had to click to another link on that page, which took them one level deeper in the records (see Figure 17b).

2001 Financial Information - Expenditures

Capital Expenditures for Calendar Year 2001
The following lists our purchasing schedule:

- We essentially purchased eight new trucks and so far we have spent \$75,000 on the loans for those trucks.
- We have also purchased other factory-new equipment like carpet cleaners and hand carts at a total rate of \$13,500.
- Finally, our used capital purchases solely consisted of computer equipment at a total of \$3,500.

[For Fiscal Year 2001](#)

Figure 17a: Initial screen containing capital expenditures for the calendar year. The link for fiscal year capital expenditure data is at the bottom of the screen.

Capital Expenditures for Fiscal Year 2001
The following lists our purchasing schedule:

- We essentially purchased eight new trucks and so far we have spent \$12,000 on the loans for those trucks.
- We have also purchased other factory-new equipment like carpet cleaners and hand carts at a total rate of \$9,800.
- Finally, our used capital purchases solely consisted of computer equipment at a total of \$5,000.

[For Calendar Year 2001](#)

Figure 17b: Capital expenditures screen containing fiscal year data.

Results comparing the designs showed that, even though all test users received the same pop-up instruction early in the form asking them to complete the form with fiscal year data, those test users in design A, who received the customized fill reminding them to provide fiscal year data, provided more accurate data than the test users who did not receive the customized fill, as shown in Table 6. Results also show there was no significant difference in the time spent per the page; thus, burden did not increase with the customized fill. For completeness, we examined speed and accuracy for the question containing the pop-up instruction. The design layout used for this question was identical in the two designs and as one

would expect, there was no significant difference in either time spent on the page or the accuracy of response between the two designs. We did not ask test users to rate the ease of use of the design they saw.

Table 6: Results of customizing the question text
 Design A= Customized fill of “fiscal year” (n=30); Design B= No fill (n=30)**
 *=significant; N.S.=not significant

Question	More accurate***	Faster
Popup instruction question (no difference between designs) (Item 7)	No fill (N.S.)	No fill (N.S.)
Capital expenditures question (Item 17)	Customized fill* (p<0.05)	Customized fill (N.S.)

**We included in the analysis only the 60 test users who answered Item 7 correctly. Our conclusions did not differ when we ran the analysis with the full 65 respondents, but we felt the more accurate analysis excluded the test users who incorrectly answered “Calendar Year” on Item 7.

***The analysis of mean score and the coefficient associated with the analysis of covariance model point in the same direction.

Even though this experiment shows the positive impact of including instructions within the question text, there are some criticisms of the experimental design we used. The question is perhaps more tricky than the situations respondents would normally encounter. One would not have expected calendar year data to be the first set of data encountered in the records, if encountered at all, since the records for the other topics did not include calendar year data. It is also not clear whether our test users understood the difference between fiscal and calendar year data. Secondly, the popup instruction earlier in Item 7 is not a proven method for conveying important information. Respondents might not see, read, or remember the instruction. In fact, another possible conclusion of the experiment is that the initial pop-up instruction did not work well. Interestingly, during the post-questionnaire test, 72% of the test users claimed to remember the popup instruction “Please answer remaining questions based on your Fiscal Year.” even though some of those very test users who claimed to remember the instruction, did not follow the instruction on Item 17. Thirdly, the capital expenditure question (Figures 15 and 16) contained two confounding designs. The automated summing switch, used in design A, was not used in design B. Fortunately, modification of sums was not necessary in this question (there were only three data fields to enter) and thus, we do not suspect the switch or lack of switch contributed to the difference in accuracy. Also, aside from the inclusion/exclusion of the words “fiscal year,” the versions of the questions in design A and B differed slightly. We did not spot this inconsistency until after the testing was completed. Again, we do not think this difference caused the difference in accuracy.

Communicating reporting units

Two different types of questions in Census Bureau economic censuses and surveys require the respondent to report in a particular unit. One request asks respondents to report figures in thousands of dollars. Another request asks respondents to report figures in either dollars or percents (but not both). We will treat each of these design issues separately.

Reporting in thousands of dollars

On the economic survey and census forms, amounts are often requested in thousands of dollars. The layout on the paper form separates the amount into millions, thousands and dollars. The dollars column is the same color as the background of the form indicating to the respondent that a response is not necessary (see Figure 18). The paper forms have this visual display, and they also have written instructions to report figures in thousands of dollars, although in the example given in Figure 18 the instruction is not next to the question, but earlier in the form.

B. Payroll before deductions (Exclude employer's cost for fringe benefits.)		Mark 'X' if None	2002		
			\$ Mil.	Thou.	Dol.
1. Annual payroll	0300	<input type="checkbox"/>			
2. First quarter payroll (January-March, 2002).	0310	<input type="checkbox"/>			

Figure 18: Example of paper-based display of reporting in thousands from the 2002 Economic Census Form #TW-48460

Electronically, there are several ways to communicate dollar amounts. One way is to mimic the layout used in the paper form, by simply graying out appropriate fields. It is also possible to add three zeros to the right of the response field, or inside the response field, with the instruction to report in thousands of dollars. It is also possible for the respondent to report data in their preferred unit. The survey organization could easily convert amounts into the desired unit once the data are received at headquarters.

We compared a design allowing the test user to report in any unit against a design requesting that test users report in thousands of dollars. Two questions tested this design difference. In the records, the answer to the first of the two questions (Item 8) was given down to the cents, and the answer to the second of these two questions (Item 9, not shown) was given in thousands of dollars. These questions were on the same page of the form and mimicked the quarterly and annual payroll questions found in many Census Bureau business data collections.

In design A (Figure 19) the test user was allowed to report in dollars, thousands of dollars, or millions of dollars. As the test user typed the number, it was right justified and commas marking the thousands and millions place automatically appeared. In this design, test users needed to designate their reporting unit. They did this by selecting one of the choices from the drop-down list to the right of each data field. The choices were “Dollars,” “Thousands of Dollars,” and “Millions of Dollars.” Test users were allowed to select the unit or enter the number in any order. The first of the two questions containing design A is shown in Figure 19. They could answer in one unit for Item 8 and in another unit for Item 9, but the test user had to specify the unit for each item.

Item 8

What was the payroll for the pay period including March 12, 2001? \$ (Select Unit) (Select Unit)

Figure 19: Design A, Select-a-Unit design.

In design B, (Figure 20) instructions below the question directed test users to report in thousands of dollars. A comma and three zeroes (e.g., “,000”) were placed to the right of the text entry field. Any numbers entered in the data field were right justified so the three zeroes looked like a part of the entire number. Just as in design A, commas appeared as the test user typed the amount.

Item 8

What was the payroll for the pay period including March 12, 2001? \$,000
(Report in thousands of dollars)

Figure 20: Design B, Report in thousands of dollars

Results in Table 7 indicate there were no significant differences in either accuracy or time spent per page between the two designs. When the records were in dollars, as they were for Item 8, we hypothesize that it was easier to report in dollars with the select-a-unit design. Only one test user correctly reported Item 8 in thousands of dollars when given the select-a-unit design. Everyone else who reported correctly reported in dollars down to the cents (\$29,321.60). The significantly different satisfaction rating suggests our hypothesis might be true. Test users who received the select-a-unit design gave it a higher satisfaction rating than did those who received and rated the design asking them to report in thousands of dollars. We

speculate that conversion from dollars to thousands of dollars is difficult. Even though the select-a-unit design might be preferred, it did not significantly improve accuracy.

When data were in thousands of dollars, however, as they were in the records for Item 9, there was no trend in accuracy favoring either the design. In Item 9, three test users who completed design A reported correctly in dollars and selected the appropriate unit, while the remaining test users reported in thousands of dollars and selected that unit. These data suggest that most respondents will report their data in the same unit used in their records. This trend supports one of the general principles of user-interface design: Avoid making the user perform mental conversion of other arithmetic operations that can be done more efficiently and accurately by the computer. Mental arithmetic is notoriously burdensome, inaccurate, and unreliable.

Table 7: Results of communicating the reporting unit
 Design A= Select-a-Unit (n=32); Design B= Report in Thousands of Dollars (n=33)
 *=significant; N.S.=not significant

Question	More accurate**	Faster	More satisfied
Records contained answer in dollars with cents (Item 8)	Select-a-unit (N.S.)	Report in thousands of dollars (page included both Item 8 and 9) (N.S.)	Select-a-unit *(p<.01)
Records contained answer in thousands of dollars (Item 9)	mean => Report in thousands of dollars (N.S.) coeff =>Select-a-unit (N.S.)		Not measured

**The analysis of mean score and the coefficient associated with the analysis of covariance model DO NOT point in the same direction for item 9, but do point in the same direction for the select-a-unit.
 => Read as “points to”

There are pros and cons to each of these designs. With the select-a-unit design, the accuracy of the response depends on two entries: the figure provided and the unit provided. We designed the unit designation to be a drop-down list so the entry could be read on one line, for example, 762 Thousand Dollars. Half of the test users in design A had seen and worked with drop-down lists on choose-one questions earlier in the questionnaire. One test user, who had not seen the drop-down list earlier, failed to select a unit for the two questions. Another test user selected the incorrect unit in Item 8 for the amount they entered, but selected the correct unit in Item 9. Thus, 2/32 or approximately 6% of the test users made a mistake with regard to specifying the unit provided. Another potential drawback of this design is the extra burden of selecting the unit. In this experiment there were only two drop-down select unit lists to manipulate. We can imagine that the design might become very tedious if there were a lot of them in a long list. A less burdensome design might have respondents defining their preferred reporting unit once and then reporting all of their answers in that unit.

The difficulty with reporting in thousands of dollars is that some respondents will not understand how to round; they won't see or read the instruction telling them to round, or they will misinterpret the three zeros in the design layout. In this experiment, when the records provided a dollar amount, four test users (about 12%) entered dollars instead of thousands of dollars in design B, which asked them to report in thousands of dollars. This experiment did not provide an example of how to round to thousands of dollars.

Reporting in dollars or percents

Some Census Bureau economic paper forms ask the respondent to report in either dollars or percents. The typical paper form layout (see Figure 21) includes two columns (one for the dollar figures and one for percent figures) side-by-side. The respondent only needs to complete one of the two columns. However, sometimes the respondent reports in both. Although reporting in both columns is not detrimental to accuracy, we suspect it takes respondents longer to report in two columns than it would to report data in only one column. It also potentially requires more time in post-collection editing and analysis. Data-accuracy problems potentially arise if the respondent alternates between reporting some amounts and some percents. It might be difficult to get the true distribution of the item of interest.

Description of sales, shipments, receipts, or revenue	Cen- sus use	2002				
		Estimates are acceptable. Report dollars OR percents.				
		\$ Bil.	Mil.	Thou.	Dol.	Percent
0722	0720	0721				0722
1. Motor carrier revenue						
a. Local motor carrier revenue	42000					
b. Long-distance motor carrier revenue	42010					

Figure 21: Example of paper-based display of reporting in dollars or percents from 2002 Economic Census Form #TW-48460 (partial display)

There are several challenges for the electronic design of this type of question and response fields. The form needs to communicate to respondents that they need to report only in dollars or percents, not both. Additionally, the electronic form should be as flexible as the paper form. For example, respondents should be able to change their reporting unit easily, even if they had begun to report in a particular unit.

Two questions tested this design variation. The records for the first of the two questions (Item 11 asking about business expenses) contained percents only, while the records for the second of the two questions (Item 28 asking about the distribution of sales) contained enough information to report in either dollars or percents. In design A (Figure 22) the test user saw two columns similar to the paper form, whereas in design B (Figure 23), only one column was visible. In both designs, the test user needed to select their preferred reporting unit (either dollars or percents). In design A, once the selection was made, the appropriate column became enabled. In design B, once the selection was made the single column became enabled, and a column heading corresponding to the unit chosen appeared. Design A required test users to perform an additional step. Test users had to enter a total amount initially in Item 11A (see figure 22). Then as the test user completed one of the columns, the other column would fill automatically with the appropriate number based on the computed calculation from the total. We thought this additional feature might aid data accuracy since test users might easily see gross errors.

Item 11

A. Provide the total expense that your business incurred as a result of employee benefits. \$

B. Provide the distribution of employee benefits your business incurs as an expense in either percents or dollars.
Select your preferred reporting unit:

Dollars
 Percents

How much of your employee expenses were generated through the following sources:

	Dollars	Percents
Health.....	\$ <input type="text"/>	<input type="text"/> %
Unemployment.....	\$ <input type="text"/>	<input type="text"/> %
Retirement.....	\$ <input type="text"/>	<input type="text"/> %
Life Insurance.....	\$ <input type="text"/>	<input type="text"/> %
Disability.....	\$ <input type="text"/>	<input type="text"/> %
Other.....	\$ <input type="text"/>	<input type="text"/> %

Figure 22: Design A, Dollars and percents

Item 11
Provide the distribution of employee benefits your business incurs as an expense in either percents or dollars.

Choose your preferred unit:

Dollars
 Percents

How much of your employee expenses was generated through the following sources:

Health.....	<input type="text"/>
Unemployment.....	<input type="text"/>
Retirement.....	<input type="text"/>
Life Insurance.....	<input type="text"/>
Disability.....	<input type="text"/>
Other.....	<input type="text"/>

Figure 23: Design B, Dollars or percents

We found, however, that data accuracy did not improve with design A. The additional information did not seem to help test users report accurately and in fact, slowed them down significantly as compared to design B. The accuracy and speed trends shown in Table 8 suggest the simpler design with only one column (design B) is potentially the better design in this test.

There was one problem with the design of the question displayed in Figures 22 and 23. We originally intended to have the amounts in each column sum automatically to a total. Programming difficulties prevented us from having a total on this question in both designs. We wonder if the lack of a total slowed down test users in the A design where there were two columns. Although adding the columns to a total was not necessary in either design, perhaps users wanted to make sure their data summed correctly. We speculate that the design with two columns simply took longer to sum mentally. The summing total was displayed for the second of the two questions that tested this design difference (Item 28).

Table 8: Results of communicating how to report dollars or percents
Design A= Dollars and percents (n=32); Design B= Dollars or percents (n=33)
*=significant; N.S.=not significant

Question	More accurate**	Faster	More satisfied
Distribution of employee benefits question Six data fields to answer, records provided primarily in percents (Item 11)	Dollars or percents (N.S.)	Dollars or percents (p<.01*)	Neither design
Web/Store/Other sales question Three data fields to answer, records contained 2 percents and 2 dollar amounts (Item 28)	Dollars or percents (N.S.)	Dollars or percents (N.S.)	Not measured

**The analysis of mean score and the coefficient associated with the analysis of covariance model point in the same direction.

Placement of the response field in relation to the question

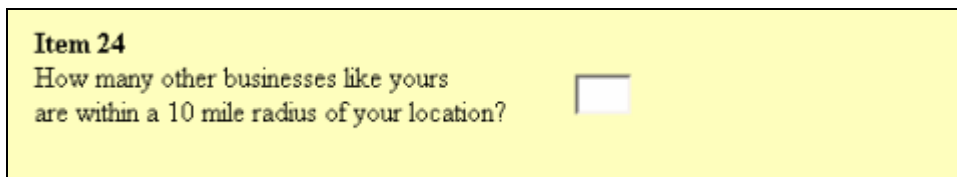
Another design issue is how best to position the answer field(s) in relationship to the question text. In some paper forms, the question text spans the width of the paper (usually 8 ½ inches minus the margins), and the answer field is below the question text, creating a “vertical flow.” With this design the respondent uses a top-to-bottom scanning pattern to read and answer the question. In other paper forms, the question text is on the left, and the answer field is to the right of the question text. When

the question text is long, it wraps within an invisible column. The answer field is maintained to the right of the text, typically centered between the top and bottom lines of the question text.

Space considerations on the form often dictate which of these designs is used. Sometimes the type of question influences the layout decision. For instance, if there are several similar questions and the answers need to sum to a total, then the left question text block and right answer field block make the most sense for summation purposes. Within a form there is generally a consistent pattern of question and answer layouts for the different types of questions whether it be top to bottom or left to right. This consistency provides visual cues to help the respondent find the appropriate answer field for the question. Research has not yet shown whether one of these layouts is better at reducing burden and increasing accuracy for a simple question and answer sequence, although the typical recommendation is to use the “vertical flow” approach on paper forms (Jenkins and Dillman, 1993).

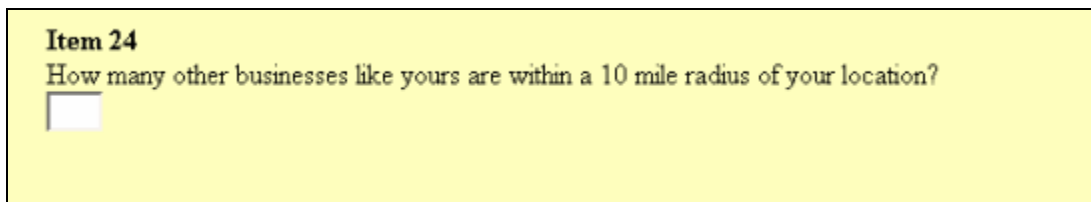
The Web guidelines at usability.gov (<http://usability.gov/guidelines/readscan.html> or Koyani et al. 2003) state that if reading speed is important, then a line length around 100 characters is ideal; yet, readers tend to prefer line lengths around 55 characters per line or less, which is about 11 words. If 11 words are used per line, a large chunk of white space to the right of the question is available, which might be ideal for the answer category.

We embedded one question within our electronic form, to test the issue of where to place the response field. In design A, (see Figure 24) we placed the question text to the left and the answer field to the right. In design B (see Figure 25), the question text spanned the width of the screen with the answer field below the question text. In both cases, the question was fact-oriented, and the answer was easily found in the records.



Item 24
How many other businesses like yours
are within a 10 mile radius of your location?

Figure 24: Design A, Question text to the left and answer field to the right



Item 24
How many other businesses like yours are within a 10 mile radius of your location?

Figure 25: Design B, Question text spans width of the page and answer field is left-justified below text

Results are found in Table 9. We found no difference in accuracy between the two designs since everyone answered correctly. The satisfaction rating did not differ significantly between designs either. Interestingly there was a significant difference in the amount of time test users spent on the page. Test users spent significantly less time on the page with the text expanded across the page and the answer category below the question text (Design B). The comparison of time also included test users reading and completing Item 25 in addition to this question (Item 24). Since Item 25 was identical across designs, we attribute the time difference to Item 24. This result is in keeping with previous speed-related findings, although we were surprised at the significant result since the question was so easy to understand that everyone answered correctly.

Table 9: Results of modifying the question and answer positions

Design A = Question to the left and answer field to the right of the question text (n=32); Design B = Question spans width of screen and answer field is below question text (n=33)

*=significant; N.S.=not significant

Question	More accurate	Faster	More satisfied
Competition Question (Item 24)	Neither (N.S.) (all test users answered correctly)	Question spans width Answer field below (p<.10*)	Question to the left Answer field to the right (N.S.)

To test this issue thoroughly, numerous questions, of differing lengths and comprehension difficulties should be tested. Testing these kinds of questions with eye-tracking equipment might also help highlight how respondents handle different layouts. These measures might help determine if one layout is more visually burdensome than another. Post-test questions might also try to determine whether respondents read and comprehended the questions.

Formatting text-entry fields

When writing numbers, a standard set of characters is used in the United States to help make the numbers readable and more understandable. For instance, we use the dollar sign “\$” and the percent sign “%” to signify different types of amounts. We use commas “,” to delineate the place for numbers greater than 999. We use the decimal point “.” as another place holder to identify values less than one. We use “:” to separate hours from minutes from seconds. There is no standard format for the characters and spacing used to delineate the numbers within telephone and FAX numbers. Some examples include (411) 555-1212, 411-555-1212, or 411.555.1212. When there is an extension, often it is identified by the abbreviation (ext.).

Another design issue is whether the electronic form should embed these characters into text fields automatically. There are two alternatives to not embedding characters automatically: 1) provide format instructions, or 2) have a free-format, letting the respondent chose whether or not to add characters to the figures entered. In this experiment, the form did not request any time estimates. We used both the “\$” and the “%” outside of the text fields where appropriate, as shown in several of the preceding figures. We also programmed the form to embed “,” automatically into numbers as test users typed values over 999. We allowed the test user to type in a decimal point “.” if appropriate. All numeric entries were right justified once the test user lost focus on (i.e., exited) the response field. We used this format set for two reasons: it seemed be the natural way the numbers would be written out long hand, and it minimized the amount of extraneous information the test user would have to type. We were not sure however, if the automated formats we chose improved data accuracy or efficiency, since we did not test this set against another set.

We were not sure how to evaluate the format set we chose. We decided to see whether test users implicitly learned the set of formats we used. This might provide some insight into how important or useful numeric formats are to a respondent’s response task. To do this we asked test users in the post-test questionnaire, “If you typed a numerical entry of 12345 for revenue, how would the questionnaire display it?” The response choices are shown in Figure 26. Over 62% of the 65 test users correctly picked the 5th response choice in Figure 26, (the questionnaire displayed revenue figures right justified, with commas, and the dollar sign outside the text field). Only two test users chose the first response choice, which did not contain commas. Eleven test users (17%) chose the second choice and another 11 chose the third choice. These data do not confirm that adding commas aided data accuracy or reduced burden. We also cannot confirm with certainty that most test users remembered or guessed correctly what format the questionnaire used. It could be that this format is the most used and, thus, the most familiar. Perhaps the format with addition of commas, justification and use of the dollar sign corresponded to test users’ mental models of how such information *should* be displayed. Perhaps, the automatic addition of commas should be the first of any future format issues studied.

\$	12345
	\$12,345
\$	12,345
	12,345
\$	12,345
	\$12,345

Figure 26: Response options in the post-questionnaire survey for the question “If you typed a numerical entry of 12345 for revenue, how would the questionnaire display it?”

In the experiment, we also compared two different designs for the telephone and FAX numbers. In the last item within the questionnaire, we requested that test users enter their name, telephone and FAX numbers as provided in the records and on the business card (Figure 1). In the default display for design A as shown in Figure 27, both the telephone and FAX fields contained the characters “() -”. All the test user had to do was type in the numbers. The numbers would automatically be placed into the correct position based on the format of (411) 555-1212. The test user did not have to try to place the cursor between the characters to enter the number. Design B, shown in Figure 28, contained a free-form text entry box for the telephone and FAX numbers. There were no written instructions on either design for the requested format.

Item 29
Provide the contact information of the person submitting this report

Filer's Name

Company Phone () -

Company FAX () -

Company E-mail

Figure 27: Design A included an embedded format for both the company telephone number and FAX number.

Item 29
Provide the contact information of the person submitting this report

Filer's Name

Company Phone

Company FAX

Company E-mail

Figure 28: Design B includes a free-form text-entry field for the telephone number and FAX number.

Results in Table 10 show that neither design was significantly faster or more accurate than the other. In addition, there was no significant difference in the satisfaction score between the two designs. The trend in these data suggests the embedded format might produce slightly more accurate data, and the free-form format tends to be a little faster to complete. We suspect that test users might have played with the embedded format a little more than they did with the free-form design, thus taking a little longer, but this hypothesis is based only on our own experience with the design, and not from review of the videotape recordings of the sessions.

Table 10: Results of an automated embedded format versus a free form for telephone number and FAX number fields
 Design A = Embedded format (n=32); Design B = Free-form (n=33)
 N.S.=not significant

Question	More accurate**	Faster	More satisfied
Phone number	Embedded format (N.S.)	Free-form (N.S.)	Free form (N.S.)
FAX number (Item 29)	Embedded Format (N.S.)		

**The analysis of mean score and the coefficient associated with the analysis of covariance model point in the same direction.

Almost all of the test users used the business card information to enter the information for this item since only 7 of the 65 test users visited the records when they were on the questionnaire page containing the telephone and FAX number request. Both the business card (Figure 1) and the records page (not shown) used the telephone format of 614-555-8945. In design B (the free-form), 88% of the test users responded in the same manner as both the business card and the records with “614-555-8945.” Two test users did not embed any characters within the telephone number and typed “6145558945” and one test user entered spaces between the numbers and typed “614 555 89 45.” Interestingly no one entered the telephone number using the format with the parenthesis for the area code as was used in the embedded design. Maybe they shied away from using “(” and “)” because they know from experience that “special characters” often cause trouble in computer response fields. To test which format people naturally prefer, we would have to set up the experiment for test users to enter their own telephone number, instead of relying on a prewritten number. However, if we had done that, we could not have compared accuracy.

Presenting amount response fields with a separate field to capture “none”

The last of the eight design issues on our original list was how to present amount response fields with a separate field to capture “none.” In many Census Bureau economic surveys and censuses, respondents are asked to report amounts for different line items. In order to account for the situation where the line item is not applicable, the paper form design includes a separate checkbox for the respondent to check if their business did not have an amount for that line item. The column heading for these checkboxes is usually titled, “Mark ‘X’ if None” as seen in Figure 29. There are really two tasks or questions associated with this design. The first question the respondent needs to answer is whether the business was in operation in 2002, and if it was in operation, the second question the respondent needs to answer is how many months the business was in operation. The design combines the two tasks into one quasi-question, which might confuse respondents. At the very least, it is not clear whether the respondent needs to fill in anything other than the checkbox, such as an N/A or a 0 in the amount field. Identical problems could occur if the same design were used in an electronic form.

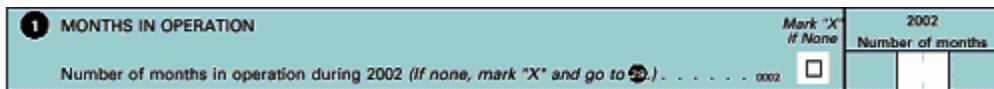


Figure 29: Example of the Mark “X” if None layout from the 2002 Economic Census Form #TW-48460

We decided to test the paper-based “Mark ‘X’ if None” design (Figure 30) against a revised design (Figure 31). In design B, we changed the paper-based instruction of “Mark ‘X’ if None” to “Check box if none.” We thought our change made the instruction more precise, yet was in keeping with the general principle of the paper form. In this design, all fields were enabled as the default. If the “Check box if none” field was checked, the amount field was cleared and disabled. In design A, we kept the matrix-like format, but in place of the paper-based column heading “Mark ‘X’ if None” we embedded a question, “Do you do this monthly?” and in place of the checkbox, two response options “Yes” and “No” were available. Test users had to answer “Yes” to the question in order to enable the amount field in the next question, “How much do you

typically spend monthly?" We hypothesized the revised design would aid navigation through the question, easing any respondent burden.

Item 14
 Enter the average dollar amount you spend monthly for these expenses.
 If no monthly expense, check the appropriate box.

Monthly Expense	Check box if none	Amount
Advertising	<input type="checkbox"/>	\$ <input type="text"/>
Store Maintenance	<input type="checkbox"/>	\$ <input type="text"/>
Truck Maintenance	<input type="checkbox"/>	\$ <input type="text"/>
Equipment Maintenance	<input type="checkbox"/>	\$ <input type="text"/>
Other Expenses	<input type="checkbox"/>	\$ <input type="text"/>

Figure 30: Design B is a slight variation of the paper form design with a separate checkbox for no amount.

Item 14
 Mark "Yes" or "No" to indicate whether your business had these expenses. If "Yes", enter the average dollar amount spent monthly.

Monthly Expenses	Do you do this monthly?	How much do you typically spend monthly?
Advertising	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Don't Know	\$ <input type="text"/>
Store Maintenance	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Don't Know	\$ <input type="text"/>
Truck Maintenance	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Don't Know	\$ <input type="text"/>
Equipment Maintenance	<input type="radio"/> Yes <input type="radio"/> No <input type="radio"/> Don't Know	\$ <input type="text"/>

Figure 31: Design A replaces the checkbox with a question. The question must be answered to enable the amount field

We found no difference in accuracy or satisfaction scores between the two designs. Our time comparison for this issue was confounded since we also included on the same page two other items (12 and 13), of which Item 13 varied the banking (single or double) of the response categories between the A and B design. More study of this issue is needed in order to draw any conclusions about either of the designs compared.

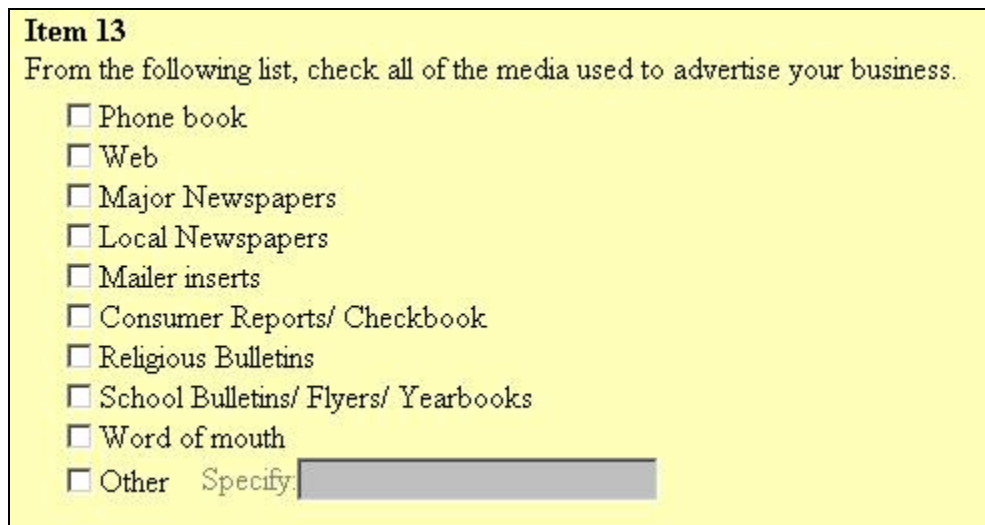
Other design issues studied in this experiment

In addition to the eight original design issues, we also captured information on a few other design issues within this experiment. Specifically, we added a question to ascertain whether or not double banking a list of items is an aid or a detriment to data accuracy. We also collected data on what test users remembered about the design to learn about what test users implicitly remembered to help them complete the task of answering the survey. They were not told that they would be quizzed at the end of the session. We gathered data about what test users remembered about links embedded within the instrument, about the name and sponsor of the questionnaire, about the item numbering used within the questionnaire, and about the format used for the amount figures. We review each of these issues in this section and mention what future experiments could modify in order to test different hypothesis.

“Mark all that apply”: Single versus Double Bank

We included in the questionnaire a mark-all-that-apply question where the layout of the response categories differed. We wanted to see whether accuracy improved with either a single list or a double-banked list of response categories. In design A (Figure 32), the 10 response categories of Item 13 were laid out as a single column checkbox list. In design B (Figure 33), this same list was double banked with 5 response categories in each column. The response categories in the double-banked list were in the same order as the single-banked list if the test user read the first column and then the second.

In mark-all-that-apply questions, respondents are allowed to check off multiple response categories. In the records, the answer to the Item 13 question was in text format and included, in order, the phone book, internet, local newspapers, radio spots and word-of-mouth. To answer correctly, the test user needed to check “Phone book,” “Web,” “Local Newspapers,” “Word-of-Mouth,” and “Other” with a write-in of radio or radio spots.



Item 13
From the following list, check all of the media used to advertise your business.

- Phone book
- Web
- Major Newspapers
- Local Newspapers
- Mailer inserts
- Consumer Reports/ Checkbook
- Religious Bulletins
- School Bulletins/ Flyers/ Yearbooks
- Word of mouth
- Other Specify:

Figure 32: Design A, Mark-all-that-apply question with response categories in a single-banked layout.

Item 13

From the following list, check all of the media used to advertise your business.

<input type="checkbox"/> Phone book	<input type="checkbox"/> Consumer Reports/ Checkbook
<input type="checkbox"/> Web	<input type="checkbox"/> Religious Bulletins
<input type="checkbox"/> Major Newspapers	<input type="checkbox"/> School Bulletins/ Flyers/ Yearbooks
<input type="checkbox"/> Local Newspapers	<input type="checkbox"/> Word of mouth
<input type="checkbox"/> Mailer inserts	<input type="checkbox"/> Other Specify: <input type="text"/>

Figure 33: Design B, Mark-all-that-apply question with response categories in a double-banked layout.

There was no significant difference in accuracy between the single-banked and the double-banked layout for this item. We could not measure speed since this item was on the same page as Item 14, which also varied between the A and B design. Because this was not one of the original issues, we did not include an ease-of-use question in the post-questionnaire pertaining to this layout. Like other issues studied in this experiment, this issue also should be studied more before coming to any conclusion regarding the effect of banking on respondent performance and satisfaction. In addition to using a larger sample, we recommend varying the number of response categories to determine if the length of the list also affects the accuracy.

Design of the initial questionnaire welcome page

The first page of the questionnaire did not differ drastically between designs. We included the name of the survey, the sponsors, the OMB number, the Census logo, and a few instructions on both the A and B designs. Design A (Figure 5) included additional instructions pertaining to the functionality of the automated summing feature built into the instrument. This was the third paragraph on the screen. Design B was identical to design A except it excluded the third paragraph. The layout of the initial screen mimicked the layout of a typical Economic Census paper form.

We were interested in determining whether any of the information from the first page of the questionnaire was stored in long-term memory. Although the information from the first page is not needed to answer any of the questions within the questionnaire, there are advantages to a design which supports “incidental learning,” i.e., things that people “pick up” without conscious effort. For instance, Dillman (2000) recommends identifying the sponsor on questionnaires to convey the legitimacy and usefulness of the survey. Thus, it would be advantageous for the design to convey with ease the survey sponsor’s name. A similar argument can be made for conveying the name of the survey. Sometimes the survey does not apply to a particular company. If the respondent pays some attention to the name of the survey, then they can decide whether their company fits into the universe for which this survey is appropriate.

To determine whether test users stored the survey sponsor and the name of the survey in long-term memory, we added two questions to the post-test questionnaire. Those questions were:

What was the name of the survey?

List all the sponsors of the survey that were mentioned on the first page.

The correct answer to the first of these two questions was, “2002 Economic Census Study, Trucking and Warehousing.” The name was located in the top, center of the initial page of the questionnaire and the words, “2002 Economic Census Study” were bolded. As shown in Table 11, no one answered this question correctly. Twenty-five test users thought the name had the word Census in it, which it did. Very often however they thought the name was “Business Census” or “US Census.” No test user mentioned the words “Economic” or “Study” in their answer, and only one test user mentioned the year “2002.” That test user responded, however, the name of the survey was “packing and Manufacturing (sic) Census 2002.” No one mentioned “Trucking and Warehousing.” Census Bureau was the name of the sponsor and this was on the second line in the top left of the page.

Table 11: Answers to the question, “What was the name of the survey?”

<u>Answer</u>	<u>Frequency</u>
Contained the word ‘Census’ (e.g., Census, US Census, Business Census, Census Survey)	36
Missing or Don’t Know	25
Other	4

Test users did even worse when trying to remember who the sponsors of the survey were, as shown in Table 12. The welcome page listed three sponsors of the questionnaire: the U.S. Department of Commerce, Census Bureau, and Statistical Research Division. The names of the sponsors were listed in the upper left corner of the welcome page. No one mentioned either the US Department of Commerce or the Statistical Research Division in their answer. The large majority of test users did not answer this question or wrote that they didn’t know the information, didn’t remember the information, didn’t pay attention to the information, or didn’t read the information. Seven test users mentioned the U.S. Census Bureau or something similar. One test user thought there were no sponsors.

Table 12: Answers to the question, “List all the sponsors of the survey that were mentioned on the first page?”

<u>Answer</u>	<u>Frequency</u>
Missing or Don’t Know	50
‘US Census Bureau’ or something similar	7
No sponsors	1
Other	7

These questions, while not necessarily pertinent to any of the design issues studied, are important in evaluating how the cover page or welcome page supported incidental learning. Based on these results, we conclude test users did not store the general information of name and sponsors in short-term memory and transfer it to long-term memory, given this welcome page. We cannot conclude whether or not test users read the initial screen since they easily could have perceived the information and placed it into short-term sensory store, but not processed it sufficiently to move it into short-term memory. However, we speculate that a better design could help improve the incidental learning of this type of information. Future studies could modify the design of the initial screen. Administering the same post-test questionnaire questions could help evaluate whether the redesign helps respondents remember the information on the initial screen, even when they do not pay conscious attention to it.

Links within the electronic questionnaire

In the post questionnaire, we asked test users to check which hyperlinks were available within the questionnaire. We wanted to know whether test users assumed there were links, and if so, which ones were the typical links. The post-questionnaire statement read:

The following links appeared in the questionnaire.

- Help
- Contact Us
- Check for reporting errors

Test users could select one or more of the links as appearing in the questionnaire. This was a trick question since none of these hyperlinks were available within the questionnaire. Thirty-seven of the 65 or 57% of the test users answered correctly by not selecting any of the response choices. Most of the incorrect answers (28%) were with test users who assumed the questionnaire had a “Contact Us” link. Seven percent of users selected a Help link, and the remaining 8% chose the edit (reporting errors) link, or a combination of the links available. Respondents guess based on their mental model (knowledge structures) for the typical survey. These results suggest many respondents have a mental model of an electronic survey including a Contact Us link.

Numbering the questions in the questionnaire

Our questionnaire contained 29 questions, which we referred to in the questionnaire as “items.” Some of the items had multiple parts, but each item was labeled with the word “Item” and the number, e.g., **Item 1**. The label was above the question and left-justified. There were no references to the item numbers except for the few skip instructions embedded within the questionnaire. If the test users completed the questionnaire correctly, they should not have skipped any items; thus even the skip instructions should not have affected the majority of test users.

Numbering questions is often used and is a typical recommendation for questionnaires. It is hypothesized that numbering helps respondents navigate the questionnaire, quickly determine the length of a questionnaire, and refer to different parts of the questionnaire, whether that be for editing purposes, or skip instructions. We wanted to determine if test users were aware of our numbering; so during the post-test questionnaire we posed the question, “Were the items/questions numbered?” Response choices included Yes, No, and I don’t know. If they responded affirmatively, we asked, “How many items/questions were there in the survey?”

As shown in Table 13, most of the test users answered that the items/questions were numbered, with 14% of the test users incorrectly answering that they were not numbered. A large proportion of test users 26% also were not sure whether they were numbered. Of the 38 test users who claimed the items were numbered, only 3 remembered or guessed correctly that there were 29 items, but most of the test users (24/38) thought there were between 20 and 30 items. Three test users responded that there were between 35 and 50 items; two test users responded that there were 16 or 17 items, and nine test users did not venture a guess.

Table 13: Answers to the question, “Were the items/questions numbered?”

<u>Answer</u>	<u>Frequency</u>	<u>Percent</u>
Yes	38	58%
No	9	14%
Don’t know	17	26%
Missing	1	2%

Interpretation of these data is speculative. Ideally, it would be interesting to see whether different styles of numbering increased the percentage of test users who answer the question correctly. For example, what if the item numbering was in reverse print, which is a format frequently used on the paper forms. It would also be interesting to determine if the distribution changes if other factors, such as skipping instructions, or editing, were built into the form. It does seem that many test users (about 37%, not shown directly in the table) were aware that there was a numbering scheme and seemed to have a general feeling that there were between 20 and 30 items in the questionnaire, even though they didn’t necessarily need that information to perform the task. Perhaps the numbering helped test users navigate or gauge how long the questionnaire was, and they made a mental note of it even though they didn’t need the information to perform the task, but needed it to estimate workload (which can be considered a meta-cognitive level issue). Future experimentation could explore changing the numbering layout to see if respondents are more or less aware of it. One could also experiment with removing the numbering altogether to see what effect that has on respondents.

Discussion and Conclusions

The results of this research are admittedly preliminary. However, they are encouraging from the cognitive perspective because they are consistent with something we know well about people: In general, people are very adaptable to their environment and will use the tools available as long as they function in reasonably consistent and predictable ways. Thus, it is probably not necessary to compare the performance of every possible widget for every possible task. A reasonable widget set, used consistently will do the job in most cases. For example, our test users were able to master both drop-down lists and radio buttons in selecting response options. The data suggest radio buttons might be preferable to drop-down lists for data requiring record-lookup, but if there is a difference, it is only slight. This means that if there are other constraints, such as limited screen real estate, using drop-down lists is a viable option. A similar conclusion can be drawn with the use of matrices not needing a horizontal scroll. Test users appeared to master this design as well as the stacked design we tested.

Some of the research findings confirmed previous hypotheses about good Web design. Specifically, the use of a horizontal scrolling with a matrix should be discouraged (Weinschenk et al. 1997), and the use of customized fills should be encouraged. These guidelines were either mentioned in the literature, or used frequently in other electronic modes and

assumed to be helpful. Designing the questionnaire to minimize mental arithmetic, whether by automating the summation of data fields or in conversion of reporting units appears to aid in efficiency and accuracy.

In addition to these perhaps more obvious findings, we learned a bit about respondent expectations within an electronic instrument. For example, the additional switch to turn the automated summing off and on did not help test users complete the form more accurately or efficiently. The switch is not a function commonly seen in electronic surveys, and for those test users who did not carefully review the instructions, the functionality of the switch was not obvious. On the other hand, test users expected a “Contact Us” link to be present in the electronic survey, so much so, that many tests users assumed it was present, when in fact, it wasn’t. (Expectations like this tell us about users’ mental models of Web surveys and Web sites in general. Since “Contact Us” links are so often present, users have come to expect and assume their presence. “Contact Us” links are, thus, part of the user’s mental model of typical or archetypal Web-site components.)

A colleague has commented to us that it is too bad we didn’t find more statistically significant outcomes. Our response to this comment is that the lack of statistically significant differences can be seen as a positive outcome in that it is consistent with the view of people as adaptable tool users. We compared reasonable designs and found that reasonable people could use them about equally well. This is an antidote to the school of thought that says there must be one “best” way for the user to interact with an interactive survey. Our findings suggest that there are probably several design solutions that will work for any given situation.

A caveat to this suggestion is that the design was not always the primary reason for response accuracy in our study. The fact that test users’ self-reported SAT score could significantly predict overall accuracy suggests that the survey-completion task requiring record-lookup is not a simple task and that this research might differ from the results of other survey-completion tasks not requiring record-lookup.

A caveat to designers of user interfaces to electronic surveys is that users build up expectations based on the controls and displays they encounter at the beginning of an automated questionnaire. It is not a good idea to use one control (widget) for a certain purpose and then use a different one later. A design goal should be to help the user (respondent) develop a rhythm or flow as they go through the instrument. The best way to do this is to use control and visual display techniques as consistently as possible throughout (e.g., Nielsen, 1989; Tullis, 1997; Woods & Watts, 1997).

In future research, the design of the welcome page and the numbering format used within the questionnaire could be revised and retested. It seems as if the number of questions might be retained by respondents even when it is not necessarily needed for completing the questionnaire. This should be studied more, but it may be that numbering is somehow useful and expected by respondents, i.e., that numbering is part of their mental model of a questionnaire and that the numbers help respondents with the metacognitive task of managing their workload.

Continuing in the metacognitive vein, our results seem to show that at least some respondents are non-readers and will skim potentially important information, whether that be instructions, the title of the survey, or the sponsors of the survey in an effort to reduce the demand on their working memories. The key to the design of these pages is to get the skimmers to notice issues of importance. For example, at Statistics New Zealand, they number their instructions right into the flow of questions as a technique to get people to read instructions.

From the trends in the data we now hypothesize that formatted fields for telephone and FAX numbers could improve data, but take respondents a little longer to complete. We are not sure of the preferred format for these numbers, but speculate it could be easily determined if we asked an open-ended question about the respondent’s own numbers and then tallied the different formats used. We do not feel we gathered enough information, or tested even the ideal designs with a number of issues. The long choose-one list is an issue needing further study. For example, the navigational links within the page are frequently used, but it is not clear if they lead to improved accuracy. Likewise, the placement of the response field in relationship to the question needs more study, as does the impact of single verses double banking.

In the longer-term future, instead of focusing on widget-to-widget comparisons, it may be more productive to focus on the general approach to design and to ask whether it is grounded in a user-centered perspective (e.g., Norman & Draper, 1986). Presentations by several authors at recent FedCASIC meetings have pointed out the need to develop survey instruments from a perspective focused on the user, not on the technology (e.g., Couper and Nicholls 1998). This means understanding who the users are and understanding their roles in the context of the work environment: Are there different classes of users?

What are the user's tasks? What are the organizational pressures on the user? With whom does the user need to communicate while responding to a Census survey?

Taking a user-centered perspective means involving the user in the development of paper-based or electronic instruments. It means bringing in users (or going out to them) to have them try out prototypes long before any final coding is begun. It means testing prototypes for comprehension, for usability, and for accessibility by people with disabilities. We have found in other research that simply complying with Federal accessibility regulations is not enough to guarantee usability. Each person with disabilities has a different cluster of issues, and the user interface to an electronic survey needs to be responsive to those individual special needs. Taking a user-centered perspective to design is the best way to meet the needs of all users.

Acknowledgements

The authors extend their thanks to Drs. Manuel de la Puente, Kent Marquis, and Jeffrey Moore (U. S. Census Bureau) for their helpful reviews of earlier drafts. We appreciate the guidance of the U. S. Census Bureau's 2000 Economic Electronic Style Guide Group in helping us select the issues to investigate. The collaborative experience with the University of Maryland was rewarding, both professionally and personally. We are especially grateful to those who participated as test users at the university.

References

- Couper, M. P., & Nicholls, W. L. II. (1998). The history and development of CASIC. In M. P. Couper, R. P. Baker, J. Bethlehem, C. Z. F. Clark, J. Martin, W. L. Nicholls II, & J. M. O'Reilly (Eds.). *Computer Assisted Survey Information Collection* (Chapter 1). New York: Wiley.
- Couper, M., Tourangeau, R., and Conrad, F. (2004). "What They See is What We Get: Response Options for Web Surveys" *Social Science Computer Review*. 22, no. 1: 111-127.
- Couper, M., Traugott, M., and Lamias, M. (2001). "Web Survey Design and Administration." *Public Opinion Quarterly*, 65 (2), 230-253.
- Dillman, D. (2000). *Mail and Internet Surveys, The Tailored Design Method*. New York: Wiley.
- Galitz, W. (1993). *User-interface screen design*. New York: Wiley.
- Koyani, S., Bailey, R., and Nall J. (2003). *Research-Based Web Design & Usability Guidelines* (NIH Publication No. 03-5424) Washington, D.C. Department of Health and Human Services.
- Jenkins, C. and Dillman, D. (1993). "Combining Cognitive and Motivational Research Perspectives for the Design of Respondent-Friendly Self-Administered Questionnaires," revision of a paper presented at the Annual Meetings of the American Association for Public Opinion Research.
- Nielsen, J. (Ed.). (1989). *Coordinating User Interfaces for Consistency*. New York: Academic Press.
- Norman, D. A., & Draper, S. W. (Eds.). (1986). *User Centered System Design*. Hillsdale, NJ: Erlbaum.
- Norman, K. L., Slaughter, L., Friedman, Z., Norman, K. D., and Stevenson, R. (2000). Dual navigation of computerized self-administered questionnaires and organizational records. (HCIL-TR-2000-22, LAP-TR-02, CS-TR-4192, UMIACS-TR-2000-71). College Park, MD: Human-Computer Interaction Laboratory and the Laboratory for Automation Psychology, University of Maryland.
- Rayner, K. (1998) "Eye Movements in Reading and Information Processing: 20 Years of Research." *Psychological Bulletin*, 124(3), 372-422.
- Redline, C. and Dillman, D. (2002). "The Influence of Alternative Visual Designs on Respondents' Performance with Branching Instructions in Self-Administered Questionnaires." In R.M. Groves, D.A. Dillman, J.A. Eltinge, and R.J.A. Little (Eds.), *Survey Nonresponse* (pp. 179-193). New York: Wiley.
- Tullis, T. S. (1997). Screen design. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., 503-572). New York: Elsevier Science B.V.
- Weinschenk, S., Jamar, P., and Yeo, S. C. (1997). *GUI Design Essentials*. New York: John Wiley and Sons, Inc.
- Woods, D. D., & Watts, J. C. (1997). How not to have to navigate through too many displays. In M. Helander, T. K. Landauer, & P. Prabhu (Eds.), *Handbook of Human-Computer Interaction* (2nd ed., 617-650). New York: Elsevier Science B.V.