

**THE INFLUENCE OF WORDS, SYMBOLS, NUMBERS, AND GRAPHICS ON
ANSWERS TO SELF-ADMINISTERED QUESTIONNAIRES:
RESULTS FROM 18 EXPERIMENTAL COMPARISONS**

by

Don A. Dillman and Leah Christian*

ABSTRACT

This paper reports results from 18 experimental comparisons designed to test 10 hypotheses about the effects of verbal language (words) and nonverbal languages (numbers, symbols, and graphics) on responses to self-administered questionnaires. The experiments were included in a large-scale survey of 1,042 university students. Significant differences were observed for most comparisons, providing support for nine of the ten hypotheses. The results confirm that people's responses to questions in self-administered questionnaires are influenced by more than words. Thus, the visual composition of questions must be taken into consideration when designing such surveys and, especially, when comparing results across surveys in which the visual presentation of questions is varied.

It has been recognized for many years that answers to self-administered questionnaires are influenced by the way in which the questions and answers are displayed on questionnaire pages (e.g., Wright and Barnard, 1975, 1978; Rothwell, 1985; Smith, 1993). However, our scientific understanding of the natures of those effects is not well developed. Although, it has been argued on theoretical grounds that visual layout and design make a difference in how people answer questionnaires (Sless, 1994; Jenkins and Dillman, 1997), little experimental evidence exists that changing the visual presentation of individual survey questions influences people's answers.

In contrast to interviews, which rely only on verbal language (or words) for presenting questions to respondents, questions in self-administered questionnaires are presented in nonverbal languages as well. These other modes of communication include symbolic language (the use of symbols with shared cultural meaning), numeric language (the use of numbers), and graphical language (the use of multiple design features such as font size, brightness, color, and spacing) that may convey certain meanings, apart from those conveyed solely by words.

* Don A. Dillman is the Thomas S. Foley Distinguished Professor of Government and Public Policy in the Departments of Sociology and Rural Sociology and Deputy Director of the Social and Economic Sciences Research Center (SESRC), and Leah Christian is Research Assistant in the Department of Sociology and SESRC at Washington State University, Pullman, Washington 99164-4014. The authors wish to acknowledge the financial support provided by The Agricultural Research Center under Western Region Project W-183, The Social and Economic Sciences Research Center, The National Science Foundation, the USDA-National Agricultural Statistics Service, and The Gallup Organization. Appreciation is also expressed to Brian McQueen for his help with part of this analysis and Thom Allen who served as study director for collection of data analyzed here. Questions should be addressed to dillman@wsu.edu.

Our purpose in this paper is to report results from tests of 10 hypotheses based upon 18 experimental comparisons in which one or more of these four languages are manipulated to determine whether answers to survey questions were affected by these manipulations. All 18 experiments were included in a survey of university students designed to obtain an evaluation of their current university experience. We seek to answer the general question of to what extent survey designers must take into account the visual presentation of information in questionnaires.

THEORETICAL BACKGROUND AND HYPOTHESES

Telephone interviews require that questions be communicated to respondents through words. However, when the same questions are asked in self-administered questionnaires, meaning may be derived from nonverbal languages as well. These nonverbal languages (numerical, symbolic, and graphical) may influence whether the questions are read, the order in which they are read, and the meaning conveyed to respondents. Thus, the nonverbal languages seem likely to affect how respondents interpret the verbal language of a self-administered questionnaire.

A conceptual framework for explaining why and how visual languages may influence respondent behavior in self-administered questionnaires has been provided by Jenkins and Dillman (1997). Two limited tests of that theoretical perspective and the influence of the four visual languages on respondent behavior have been performed using questionnaire branching instructions. In both of these tests it was found that manipulating several aspects of the visual languages simultaneously, i.e., size of font, introduction of arrows, changes in spacing, and the addition of verbal language, in an effort to improve compliance with branching instructions, reduced respondent errors significantly. In a student classroom experiment in which 1,266 students completed one of three versions of a questionnaire, commission errors (not skipping ahead when directed to do so) were reduced from 20% to between 7-9% for the newly designed instructions (Redline and Dillman, 2002). In a follow-up test imbedded in the 2000 Decennial Census of the United States, commission errors were reduced from 21% to 13-15% for two similar methods (Redline, Dillman, Dajani, and Scaggs, 2002). A shortcoming of both of these experiments is that the independent effects of the verbal, symbolic, and graphical language changes were impossible to disentangle. These experiments have shown that the use of visual languages can make a significant difference in whether respondents follow branching instructions. However, they did not reveal which specific languages were responsible for the changes in respondent behaviors and only two discrete examples of combined manipulations were provided.

By conducting a series of experiments involving many different manipulations of individual language elements, we attempt here to provide insight into a range of potential response effects. In contrast to the work mentioned in the preceding paragraph, most of the tests involve individual questions that all respondents were expected to answer. Thus, this paper is an effort to move beyond branching instructions as a focus of language effects in self-administered surveys, as well as further investigate how manipulations of verbal and nonverbal languages affect response behavior to self-administered questionnaires.

A Symbolic Language Hypothesis

Symbolic language (such as arrows, answer spaces, and boxes) is often used without additional verbal instructions to convey meaning or give direction to respondents when answering questionnaires. For example, the cultural meaning assigned to an arrow should guide the respondent's attention in the direction the arrow is pointing. Thus, symbolic language can provide another means of communicating to respondents as well as to work in support of other visual languages of the questionnaire.

Hypothesis 1: Addition of an arrow (symbolic language) between an answer and a subordinate question increases the likelihood that respondents will answer the subordinate question.

Sometimes questionnaire designers wish to direct respondents who choose a particular answer to a subordinate question for which an answer is desired. An example, and the one tested here, was imbedded in a question that asked where the respondent wanted to live after completing college, "Eastern Washington" or "Somewhere Else" (as shown in Figure 1). After the answer choice "Somewhere Else", the word "Where" was listed on the same line as the answer choice, approximately 10 spaces beyond the end of the category description and 26 spaces beyond the answer box. In one version, an arrow was placed between the answer "Somewhere Else" and the subordinate question "Where."

It has been shown that when respondents read text they focus on a space of about 2 degrees, or 8-10 characters in width (Kahneman, 1973). This distance is known as the *foveal view*. Thus, the arrow was placed so the subordinate question would be visible in the foveal view. An arrow is a symbol that is culturally defined to focus one's attention in the direction the arrow is pointing. It was reasoned that respondents were more likely to see and, as a result, respond to the subordinate question when an arrow was placed between the category description and the subordinate question. This is a test of the independent effect of symbolic language on item nonresponse. An arrow was tested because it was one of the manipulations of a combined method that was most effective at reducing branching errors in the study by Redline et al. (2002).

Five Graphical Language Hypotheses

Graphical language is the conduit through which all other languages are expressed (Wallschlaeger and Busic-Snyder, 1992) and includes such features as variations in size, color or brightness, and location of words, symbols, and numbers. The influence of these graphical variations is realized through a number of higher level visual behaviors that have been shown to guide the processing and interpretation of information. For example, the *Law of Pragnanz* states simpler shapes are easier to perceive and remember, and the *Law of Similarity* says that similar shapes and figures will be seen as a group (Wallschlaeger and Busic-Snyder, 1992). The five following hypotheses about individual graphical language changes evaluate only a few of the many ways in which graphical language may influence respondent behavior.

Hypothesis 2: Providing a larger space for answers to open-ended questions will produce answers that are longer and contain more themes.

One of the major shortcomings of self-administered questionnaires is that respondents typically provide shorter, less complete answers to open-ended questions than when surveyed in interviews (Dillman, 2000). Whereas interviewers can probe and cajole respondents to provide additional detail, this is not possible on self-administered questionnaires.

We expect that respondents will draw information about the questionnaire designer's expectations (Beatty and Herrmann, 2002) from the amount of answer space given for open-ended questions. Based upon work by Stember (1956), Smith (1993) has argued that allowing more space for recording open-ended answers in interview-administered surveys produces longer recorded responses that may be closer to actual verbatim. In this study three separate experimental questions were tested, with the space on one form being about twice that provided on the other (see Figure 1).

Hypothesis 3: Special instructions are more likely to be followed correctly if they are located just before, rather than just after, the place of their intended use.

It has been argued that an important goal of questionnaire construction is to use graphical language in a way that makes the elements (number, query, instructions, and answer choices) of a question appear as a distinct group, and that instructions in questionnaires should be provided exactly where they are needed (Dillman, 2000, p. 96-99). This is based on the *Law of Proximity* (Wallschlaeger and Busic-Snyder, 1992), which states that items graphically near each other will be seen as a group. In this experiment a Yes/No question was developed that was expected not to apply to a number of respondents (see Figure 1). In one version a special instruction to skip to the next question when applicable was placed ahead of the response categories, and in the other version placed immediately after the categories. It was expected that more people would skip the question when the instructions were located before the response options because they were more likely to see and thus read the instructions before answering. Thus, the location of the instructions would influence whether the respondents read them, subsequently influencing whether they chose to skip or answer the question.

Although the logic of placing instructions just prior to where they apply seems compelling, we have observed many self-administered questionnaires that place response categories ahead of instructions pertaining to their use. In addition to testing this effect of location, we were also interested in learning whether the latter placement might lead some respondents to connect the instructions with the question that immediately followed in accordance with the Law of Proximity and result in skipping over that question to go directly to the one that followed.

Hypothesis 4: Nonlinear scale layout achieved by double- or triple-banking of scale categories results in different answers than does linear scale layout.

Responding to ordinal scale questions appears to involve a different mental process than responding to nominal scale questions. For ordinal questions, the respondent must decide where her answer fits on an implied continuum, whereas responding to nominal scales requires respondents to compare all categories in order to select the best answer (Dillman, 2000, pp. 43-46). A linear layout of choices would seem to facilitate the process of identifying where a respondent best fits on the continuum. However, nonlinear layouts of ordinal scale questions, known as double-, triple-, or even quadruple- banking, depending upon the number of answer columns used, are often used in mail surveys to save vertical space and paper.

One reason that answer choices may be affected by the nonlinear format is that some respondents may be inclined to read horizontally, while other respondents read vertically. If respondents have become used to reading vertically, as the design of the test questionnaire encouraged through listing answer choices to other questions on the same page in a column, the introduction of a multiple-banking format seems likely to introduce potential confusion as the respondent processes the information. For this hypothesis, two experimental tests were conducted with double- and triple-banking formats each being tested against a vertical linear format (Figure 1).

Hypothesis 5: Increasing the distance between one response category and other choices results in greater selection of that answer.

No rules appear to exist for how far apart response categories should be placed from one another. Intuitively, it would seem that response categories should be placed equal distances from one another, particularly for ordinal scales. However, the ramifications of unequal distances between categories are still unclear. Smith (1993) reports that in a self-administered cross-national paper survey, respondents in one country were more likely, by a large margin, than were respondents in other countries, to choose lower answers to a social rank question when the boxes for the lower answers were larger.

It was found in a branching instruction experiment (Dillman and Carley-Baxter, 1999) that 3 out of 52 questions produced significantly different responses among treatment groups. Three questionnaire versions were used and the distances between answer category boxes were varied unintentionally on one of the three forms. For one of these questions, responses were given four choices for their most important life goal. Respondents were more likely to choose the answer "have a life partner with whom you have a satisfying relationship" (59% vs. 52%) when two additional vertical line spaces were placed between the answer box for this category and the one that preceded it. A significant difference among forms also occurred for respondent answers to a speculative question about the number of adults likely to have a cellular telephone in 15 years; four response categories were given. On one version of the questionnaire the option "about 3/4" was expressed in a way that placed one additional line of vertical space between this category and both the preceding and follow-up choices. On this version, 38% chose "about 3/4". However, when the distances were equal, the percent choosing this answer was significantly lower – 28% and 29% respectively. In the versions where these responses were chosen more

often, answer boxes were placed to the right (instead of the left) of response options. The possibility that this change in answer box location influenced the difference in some unknown way suggested the need for testing the spacing of the response options independently of the placement of the answer boxes.

These question formats have been retested in this study. The question on cellular phone usage has been changed to use of a recreational center, while the wording of the second question has been retained (Figure 1). We expect that more people will choose the answers to the substantive categories that are given greater visual prominence by being spatially set off from the others.

Hypothesis 6: Reversing the order of answer categories for Check-All-That-Apply questions will result in more people choosing the first-offered categories.

Considerable research has been done on response category order effects. It has been argued by Krosnick et al. (1996) that respondents to surveys often satisfice, that is, choose answers from earlier items in a list. The satisficing phenomenon appears to underlie findings from primary effects research which has shown that respondents who choose answers from written lists are more likely than respondents to telephone surveys to choose answers near the top rather than the bottom of a list of answers (Krosnick and Alwin, 1987). However, most of the research on these issues has focused on questions in which people are asked to select only one answer.

An exception is a study by Israel and Taylor (1990) which showed that the order in which categories are presented in such questions affects people's answers to Check-All-That-Apply questions on mail questionnaires, with the earlier categories receiving more selections than when listed later. Although the reasons are not clear, it has been suggested that in addition to satisficing, the lack of mutual exclusivity and social desirability may have contributed to these differences (Dillman, 2000, p. 66).

This experimental test evaluates the effects of a complete reversal of categories, rather than the partial change in ordering used by Israel and Taylor (1990), in a question that asked students to indicate which of 10 attributes described their university. The substance of this question was such that we expected most respondents to check several categories (see Figure 1).

Two Verbal Language Hypotheses

An enormous amount of research has been conducted on verbal language effects in answers to survey questions (e.g., Schwarz and Sudman, 1992; Sudman, Bradburn, and Schwarz, 1996). Our focus on verbal language manipulations in this research is limited to specific connections between words and "nonverbal" languages. One hypothesis (three experimental comparisons) tests the effects of removing verbal language from interior ordinal scale categories, and the other hypothesis (two experimental comparisons) tests how the verbal language of a scale may interact with the graphical layout.

Hypothesis 7: Removing word descriptions (verbal language) from interior scale points will change the response distribution.

The development of telephone interviewing has encouraged the use of polar-point-labeled scales rather than fully-labeled scales because it is easier for interviewers to say, e.g., “On a 1 to 5 scale, where 1 is very desirable and 5 is very undesirable...” than to read a full set of categories, e.g., “...very desirable, somewhat desirable, neither desirable nor undesirable, somewhat undesirable, or very undesirable...” Previously, in self-administered questionnaires, there has appeared to be no particular advantage to removing interior scale labels.

In an experiment conducted by the Gallup Organization (Dillman, Phelps, Tortora, Swift, Kohrell, and Berck, 2001), it was found that the use of polar-point scales to measure satisfaction with long distance service led to nearly twice as many telephone respondents choosing the positive-end labeled category as did mail survey respondents (38% vs. 21%). Primacy/recency effects previously identified by Krosnick and Alwin (1987) were ruled out by means of a sub-experiment as a cause of these differences. It was reasoned that the presence of graphical and symbolic language on the self-administered questionnaire, i.e., the presence of answer boxes located between the polar-points, may have made those categories more prominent and led to their greater use.

This experiment represents a partial extension of that experiment. Inasmuch as the addition of symbolic and graphical languages appeared to influence answers, we reasoned that the presence or absence of verbal language that defines each scale category might further influence the use of particular response categories in self-administered questionnaires. The three experimental tests of this hypothesis compare the defining of interior categories with symbolic (answer boxes) and graphical language (equidistant spacing between categories) with also using verbal language to give meaning to each category (Figure 1). One of the three experimental tests also uses numbers to define the categories.

Hypothesis 8: Changing the relative proportion of positively vs. negatively worded categories in a five-point scale results in answers that take into account both words and graphical composition of the scale.

Schwartz et al. (1985) have shown that respondents take into account the word labels as well as the position of scalar categories in selecting answer categories. Furthermore, in a previous mail survey of university students by Rockwood, Sangster, and Dillman (1997), university students reported studying more hours per day when the five response categories began with “less than 2.5” and ended with “more than 4.5” than when the choices began with “less than .5” and ended with “more than 2.5”. In the larger range (<2.5 to >4.5), 69% of the students reported studying greater than 2.5 hours. However, in the low set, only 23% of the respondents reported studying more than 2.5 hours. It was concluded from these experiments that respondents were influenced by the number of categories, as well as the category labels to each response category.

Based on research of this nature, it has been argued that positive and negative categories should be balanced on satisfaction and other scales (Dillman, 2000, pp. 57-58). However, previous research has not shown the extent to which graphical versus verbal language affects such scales. For this hypothesis, two satisfaction questions were changed from having a positive middle category (Good) with Very Good and Excellent above it, to having a less positive middle category (Fair) that separated Excellent and Good from Poor and Very Poor (Figure 1). If only word labels affect answers, then we would expect the proportion of Good, Very Good, and Excellent on one scale to equal the proportion of Good and Excellent on the other scale. And, if the scalar or balance attributes of the scale also influence answers, we would expect the proportions of Good to Excellent answers to increase when they are allocated to three instead of two categories.

Two Multiple Language Change Hypotheses

Although combined manipulations of different languages result in losing one's ability to determine the specific cause of any differences, attempts to improve question formats often involve joint changes in two or more languages. This concern over question improvement leads to our final two hypotheses. The general conception guiding these two combined manipulations of the nonverbal languages is that they can be utilized to support the verbal language or that support can be withheld, resulting in differences in the way the meaning of a question is interpreted, or even confusion on the part of the respondent. The first hypothesis involves the simultaneous manipulation of two languages (graphical and verbal), and the second involves three languages: graphical, symbolic, and numerical.

Hypothesis 9: Changing a Check-All-That-Apply question to a Yes/No format will increase the proportion of respondents who select each answer.

The purpose of these combined graphical and verbal language manipulations is to attempt to reduce satisficing by bringing the total self-administered stimulus more in line with the way it is likely to be asked in an interview mode (Figure 1). In the interview mode, individual consideration of each sport would be required of the respondent. The verbal language change in the test question about being a fan of specific varsity sports is in two locations – the addition of a second (no) answer choice and a change in the stem of the question where the wording is aimed at encouraging an answer to each item – rather than only the ones that apply.

This change is motivated in part by Beatty and Hermann's argument (2002) that the respondent must decide how much effort he/she wants to put into the response process, how much effort is required to decide, and the precision the respondent thinks is needed. The answer then is partly determined by the clues sent by the researcher on the level of accuracy desired. Asking respondents to choose between two response options (Yes or No) suggests that a more accurate answer is desired than when asked to select from a list.

Past use of the Yes/No format in nonexperimental settings has shown that respondents sometimes check only the "Yes" answers, thus treating it as a Check-All-That-Apply question. Thus, the effects of "No" answers must also be evaluated.

Hypothesis 10: The elimination of graphical and symbolic language that supports the linear definition of a scale produces different answers than a number box for selecting an answer to a scale with labeled polar categories.

This experiment tests the effects of eliminating graphical (linear layout of choices), symbolic (answer boxes), and numerical support to the verbal question. It has been argued that use of a number box might make it possible to provide equivalent stimuli in interview and self-administered surveys where the additional graphical and symbolic language cannot be provided in both modes (Dillman, 2000, pages 235-236). Determining the effects of removing these languages within self-administered questionnaires, and thus relying solely on words, is a related concern that needs to be tested before further tests across modes are conducted. Three items (Figure 1) tested the effect of number boxes vs. linear scales for recording answers.

One of the potential difficulties of the number box format is that it requires respondents to remember the specifics of the scale when providing their answer, whereas the inclusion of graphical and numerical information in the alternative layout (Figure 1) provides a reminder of how the scale is constructed within the foveal view of the response options. Thus, this test, with its simultaneous manipulation of words, symbols, and numbers, provides an example of how the visual languages can work to support each other.

PROCEDURES

The 18 experiments were embedded on pages 2, 3, and 4 of a four-page questionnaire developed for assessing the student experience at Washington State University and conducted from March to April 2001. It was printed in a two-column format on 8-1/2 x 11 inch pages, with a colored background being used to contrast with white answer spaces provided for both open- and closed-ended questions.

Four versions of the questionnaire were mailed to equal subsamples (450) of a random sample of 1,800 undergraduate students living in the Pullman, Washington area (students on other campuses or enrolled in the distance degree program were excluded). A \$2 incentive was enclosed with the first mailing. A follow-up postcard and one replacement questionnaire were mailed, obtaining a response rate of 57.9% (1,042) of the 1,800 questionnaires mailed.

The experimental questions reported here were the same in two of the four versions of the questionnaires. Results from three other experiments having to do with the construction of pages one and three have been reported elsewhere (Sawyer and Dillman, 2002).

Inclusion of so many experiments in one questionnaire raises issues of whether some of the experiments may have affected results for others. That possibility cannot be ruled out. However, individual cognitive interviews of 22 students not in the survey sample who were asked to complete and comment on the questionnaire produced no responses that would suggest a differential effect across questionnaires. For the most part, the alternative treatments being tested here are not ones that seem likely to influence respondents in such a way that answers to

subsequent questions would somehow be affected. It is quite possible that percentage distributions to individual questions were influenced through order effects, but we were unable to find evidence that such effects might have impacted the kinds of hypotheses being tested here.

Statistical tests made to evaluate the hypotheses include chi-square tests for differences and t-tests for mean differences, where appropriate. The tests we made vary depending upon the questions and therefore will be reported for each hypothesis.

FINDINGS

A Symbolic Language Change

Addition of arrow to identify subordinate question. Adding the arrow to identify a subordinate question resulted in more respondents noticing and answering the subordinate question (see Table 1). First, nonresponse to the subordinate question was reduced significantly ($x^2 = 5.2$, $p = .02$) from 8% to 4.2% when the arrow was present. 93.9% of the respondents who chose "Somewhere Else" answered the "Where" question on the arrow version, whereas only 91.5% answered the subordinate question without the arrow, and this difference is significant at the .02 level. The arrow also increased the number of respondents from .4 to 1.8 percent who answered the "Where" question without checking an answer first. Thus, the presence of the arrow appears to draw the respondents' attention to the subordinate question making its verbal language more prominent.

Graphical Language Changes

Use of larger answer spaces on open-ended answers. Varying the answer spaces shows that a larger answer space influences both the number of words and the number of themes provided in respondent open-ended answers (Table 2). The number of words was hand-counted for each open-ended response and the themes were counted as number of topics mentioned. The coding of themes was completed by one researcher and 10% were verified by another researcher, with 90% agreement. The results indicate that a larger space produced a higher number of words and themes. For the first question, there was an average of 13.3 words and 2 themes when given a larger space, whereas when given a smaller space responses averaged 9.7 words and 1.8 themes. Responses to the second question averaged 12.9 words and 2.1 themes with the larger space, and 6.6 words and 1.7 themes with the smaller space. Finally, the third question responses averaged 12 words and 1.5 themes when given a larger answer space, and 10.2 words and 1.4 themes on when given a smaller space. The difference in the number of words was significant at the .01 level or higher on all three questions, whereas the difference in the number of themes was significant at the same level or higher for only the first two questions. These results indicate that in general the manipulation of the graphical language by changing the size of the response space significantly affected both the amount and content of information the respondents wrote in the space provided.

Location of special instruction before vs. after answer categories. This graphical language change dramatically affected whether respondents used the special instructions. When the special instruction “If you haven’t had many one-on-one meetings, just skip to Question 9” was placed after the response options, 54.6% of the respondents said “Yes”, 42.4% said “No”, and 4.4% provided no answer (Table 2). However, when these instructions were placed before the responses, 54.7% said “Yes”, 19.1 said “No”, and 26.2% provided no answer ($x^2 = 25.9$, $p = .000$). Thus, when special instructions were graphically placed where respondents were more likely to see them, prior to answering the question, the location influenced their decision to answer or not answer the question. Furthermore, placing the instructions after the responses introduced confusion as some respondents used the instructions for the following question. Nonresponse to the following question increased from 2.6% to 11.0% when the instructions were placed after the responses. It appears that respondents may have mentally grouped the special instruction with the following question and interpreted the instructions as directing them to skip over it to Question 10.

Linear vs. nonlinear scale layout. Changing the scale layout from a linear to nonlinear layout, affected respondent behavior in one test but not the other one. In the nonlinear, triple-banked version, 40.4% of respondents chose “Good” and 42.4% chose “Very Good”. Whereas in the linear format, 31.3% of respondents chose “Good” and 48.8% chose “Very Good” ($x^2 = 10.8$, $p = .029$) (Table 2). More respondents chose “Good” when the scale was triple-banked and less respondents chose “Very Good”, suggesting that some respondents were reading horizontally whereas others were reading vertically. A slight trend in the same direction exists for the second question, which used double-banking, but it is not significant ($x^2 = 2.3$, $p = .509$). On this scale, the two horizontal categories “Very Satisfied” and “Somewhat Dissatisfied” do not display the sense of a complete scale, as might be inferred from the three horizontal categories – “Excellent, Good, Poor” – on the question for which differences are statistically significant. Thus, differences in the verbal language of the two tests of this hypothesis could have produced the difference in results.

Equal vs. unequal spacing between response boxes. The results from testing the effect of equal versus unequal spacing between response categories were not significant ($x^2 = 6.8$, $p = .079$ and $x^2 = 1.4$, $p = .844$) in either test (Table 2). However, both responses where the graphical spacing increased their prominence resulted in slight increases in the frequency of their selection by respondents. Thus, the results were in the predicted direction but not significant.

Reversal of response category order for Check-All-That-Apply question. The results from reversing the order of answer choices in a Check-All-That-Apply question were striking. Respondents who were given the reverse order checked a significantly greater number of responses for 8 of the 10 categories. Overall, the mean number of answer categories chosen increased from 3.24 to 4.13 ($t = 7.386$, $p = .000$) (Table 2). The largest differences were observed for “Electronic or Wired University” (increasing from 53% in the first order to 73.2% in the reverse order) and “Outdoors oriented” (increasing from 26% to 33.8%). Greater differences were seen at one end of the scale than the other. Clearly, order has a dramatic effect on the frequency with which answer choices are chosen.

Verbal Language Changes

Removal of world labels for mid-scale categories. In the three tests of removing scale labels, all differences were significant for at least one of the tests, with the polar-point scales tending to result in greater use of the middle categories. On the first question the number of respondents choosing the middle category increased from 11% on the fully labeled version to 23.4% on the polar-point version ($\chi^2 = 62.4$, $p = .000$) (see Table 3). On the second question this increase was from 23.6% to 28.4% ($t = 1.7$, $p = .045$, but the χ^2 test was not significant), and on the third question the increase usage of the middle category went from 14.3% to 21.4% ($\chi^2 = 14.7$, $p = .005$). Due to the increased usage of the middle category, respondents were less likely to choose the descriptors on the positive end of the scale, especially the “Somewhat Desirable” and “Somewhat Confident” responses.

Balanced vs. unbalanced word labels for scales. The second test of verbal language, the effecting of changing scale labels to unbalance the scale, resulted in more respondents choosing positive responses when given a positively skewed scale (see Table 3). Respondent choices appeared to be affected by both the verbal language descriptors, as well as the graphical layout of the scale, because respondents were more likely to choose positive answers when more of the options were positive. When the Very Good category was added, the percent of Good or Better responses increased from 22.7% to 37.2% on one of the scales, and 67.1% to 79.9% on the other one. This change is also evident by comparing the means between the unbalanced and balanced versions. For Question 1, the mean for the balanced version was 2.7 and 2.2 ($t = 7.9$, $p = .000$) for the unbalanced version. For Question 2, the means also decreased from 3.8 to 3.3 ($t = 7.6$, $p = .000$) in the unbalanced version. Thus, the means are greater (i.e., the evaluations are more negative) for the balanced version of the questions.

A Graphical and Verbal Language Change

The effect of changing Check-All-That-Apply question to Yes/No format. Changing a Check-All-That-Apply question to a Yes/No format also influenced respondent behavior. In the Yes/No format, respondents were more likely to select 14 of the 15 choices by a margin of 4 to 13.8 percentage points, all of which were statistically significant (Table 4). The mean number of choices chosen was also greater for the Yes/No format ($t = -6.954$, $p = .000$). Eleven percent (11%) of respondents treated the question as a Check-All-That-Apply question and only selected “Yes” responses, leaving many others blank (i.e., neither Yes nor No was checked).

A Graphical, Symbolic, and Numerical Language Change

The effect of a removal of numbers, separate answer boxes, and linear layout of scale. The final experiments tested the combined effects of manipulating the graphical, symbolic, and numerical languages by changing a polar-point scaled question to a number box. This combined manipulation produced dramatic differences. First, the use of the number box significantly increased the mean responses for each of the questions tested (see Table 5). The means changed from 2.4 to 2.8 ($p = .000$ for both) in the first and second questions, and from 2.7 to 2.9 ($p = .000$) in the third question.

Additional coding was performed to test the possibility that respondents became confused and did not remember the direction (from positive to negative; consistent with other items on the page) of the scale. Responses to each of the three number box questions were coded based on whether any scratched out or changed answers were evident. If an answer had been changed, it was then determined whether the previous answer was “legible”. Only obviously legible answers were coded, noting both the changed and actual answer. On the number box version, 10% (versus 1% for the polar-point scales) of respondents scratched out answers to at least one of these questions and provided a different answer. A total of 74 respondents made 86 changes in their answers. Most of these errors occurred by respondents reversing the scale when selecting their response: from 4 to 2 (44 respondents) and from 5 to 1 (10), while a few also changed from 2 to 4 (4) and 1 to 5 (1). These data suggest that removing the supporting graphical and numerical language by displaying the scale in the polar-point format introduced confusion in respondent’s understanding of the scale. This finding suggests that other respondents may have made this error without catching or changing them resulting in the larger means for the number box questions.

To test for the possibility of respondent confusion, individual correlations were calculated between both the linear scale and number box answers, and 13 other questionnaire items about satisfaction with classroom experience that were not varied across experimental treatments (one was varied). Each of these items was expected to correlate positively with answers to these three test questions. If confusion exists among respondents who did not change their answers, we would expect the correlations between the number box answers and the 13 items to be lower than for the polar-point scalar format. As expected, all 13 correlations for each of the items using both scalar formats were positive, however, the mean for the polar-point format was .24 compared to only .14 for the box format. In only 4 of 36 instances was the correlation between the 13 items and the test items in the direction of the polar box correlation being higher. It was also observed that the mean difference for the first test item (.16) was higher than the mean (.10) for the second item, which was higher than the mean (.03) for the third item in the questionnaire sequence. This was consistent with the expectation that respondents were less likely to make errors on the later items.

DISCUSSION AND CONCLUSIONS

When researchers develop survey questionnaires, their attention is often focused on question wording as the sole conduit of question meaning. In this paper we have proposed that the use of symbols, numbers, and graphics communicate additional meaning to the respondent that may independently and jointly influence respondent behavior.

These 18 experimental tests of 10 hypotheses combine to form an initial analysis of how visual languages affect respondent behavior. Significant differences were found for all but one of the hypotheses (though not on every experimental item), and the remaining tests were all in the expected directions. Our general conclusion is that the visual design of questions on self-administered questionnaires has a significant impact on respondent behavior. In addition, the consistency of findings suggests that it is important for users of survey data to explicitly consider

such effects when designing questionnaires and when comparing results across studies; similarly worded questions may be presented to respondents in visually dissimilar ways, resulting in answers not being directly comparable.

Some of the specific findings reported here have straightforward implications. The results from adding an arrow between an answer choice and a subordinate question placed just outside the foveal (8-10 character) view, show that such symbols can independently guide respondents to appropriate follow-up questions. Also, the results of locating *after the answer choices* a special instruction that should be considered before answering (Hypothesis 3), produced negative consequences for both answering that question and correctly advancing to the next question. This suggests that placing instructions both in the navigational path exactly where they are needed and spatially grouped with the proper question are important. These results indicate that graphical and symbolic languages can influence respondent navigation through the survey and answers to specific questions.

Space limitations in survey questionnaires are often a concern to research designers; however, results from this research suggest that spacing conveys meaning to respondents so care should be taken when designing survey layout. Reducing the amount of space for open-ended answers resulted in respondents giving answers with fewer words in which fewer themes were identified in two of the three test items. However, the item that did not reach significance was an item in which brief answers were expected, and both the large and small answer spaces were much larger than for the other questions.

A second finding relevant to spacing and questionnaire design concerns the practice of double- or triple-banking scalar response choices. Dividing response choices so that Excellent and Very Good appeared in the first column, Good and Fair in the second column, and Poor in a third column resulted in significant changes in the relative proportions of respondents choosing Very Good and Good. The likely reason is that this kind of visual display made it unclear to respondents whether they should read horizontally (in which case they saw the choices Excellent, Good, and Poor), or vertically, seeing the entire scale as it was intended. A similar trend was noticed when only four categories were used for a satisfaction question, but it did not reach significance. Inasmuch as the task in responding to scalar questions is to determine where one fits on the scale, it would seem that an entirely vertical or horizontal display of answer choices is preferable to double- or triple-banking of response choices, otherwise, confusion is introduced.

The findings with regard to long list questions (Check-All-That-Apply and Yes/No) are also important in understanding how visual languages affect respondent behavior. Presenting Check-All-That-Apply questions in a reverse order confirmed findings by Israel and Taylor (1990) in which changing the order had substantial effects on the proportion of responses for each category. Reversing the order of 10 possible descriptors of Washington State University resulted in an increase of from 1.3 to 40.2 (mean = 12.1) percentage points for the 10 items. This pattern does not reflect a pure primacy effect in which the first items listed get more mentions than the items listed last, as observed by Krosnic and Alwin (1987) for lists in which only one item was selected. Instead, it appears possible that a primacy effect occurred. When the list began with positive descriptors (Diversity and World Class University), respondents were more likely to

choose those and all succeeding items than when the list began with more negative descriptors (Farm/Agriculture School and Party School). This issue clearly needs more research using different kinds of lists and ordering of items. However, it does raise serious questions about the often-used practice of converting questions that are asked in a Yes/No format in interview questionnaires to a Check-All-That-Apply format for self-administered questionnaires.

The related findings that changing Check-All-That-Apply questions to a Yes/No format increased the proportion of Yes answers for all items also has significant implications. From a cognitive standpoint it would seem that requesting an answer to each item would increase the likelihood that respondents would give deliberate consideration to each item separately, thus changing the response pattern from searching the list for sports to deliberately considering each sport and providing an answer. It appears that some respondents are continuing to treat the question as a Check-All-That-Apply and are only checking Yes answers. However, the increased cognitive consideration (as indicated by increases in the number of respondents choosing each sport) may still outweigh the number of respondents incorrectly treating the question as a Check-All-That-Apply.

The finding that the percent of respondents answering Good or Better goes up substantially when the choices of Good and Excellent are expanded to include Good, Very Good, and Excellent on a five-point scale, reconfirms previous research (Schwarz et al., 1985; Rockwood et al., 1996) that respondents draw information from both the graphical and verbal attributes of scales when choosing answers. Results from our current study provide further evidence that absolute interpretations of results based only upon scale labels, e.g., “a majority of the respondents feel the service they received was Good or Better” are inappropriate. The interpretation of results needs to take into account graphical composition of the scale as well as the words themselves. Although most of the experiments reported here focused on the manipulation of one aspect of an individual language (whether a symbolic, graphical, or word) change, one experiment manipulated simultaneously three different languages. This experiment, on use of a number box vs. a linear polar-point scale, suggests the need to consider in question design the supportive role that symbols and numbers play in helping respondents reply accurately to survey questions. Removal of answer boxes and numbers in a linear layout resulted not only in different answers being recorded, but produced evidence of respondent confusion, i.e., more erasures, which were likely to consist of number reversals such as changing 4’s to 2’s and 5’s to 1’s, and lower correlations between the number box answers and other satisfaction measures. Without the presence of symbolic, numerical, and graphical layout information, respondents became confused with regard to the direction of the scale. Although this combined manipulation of three languages does not allow us to determine the role that each played in reducing confusion, the results do suggest the supportive role that such visual information plays in helping people to respond to the wording of questions is important in questionnaire design. One implication is that when opinion scale questions in self-administered questionnaires are designed with answer formats that require applying information from special instructions in other locations, or even the stem of the question, answers are likely to be less accurate.

The research reported here was partly motivated by two studies (Tarnai and Dillman, 1992; Dillman et al., 2001) that have found significant differences between answers to telephone and mail questionnaires. In those studies respondents to the aural survey mode (telephone) were more likely than visual respondents to choose terminal scale points, particularly on the end towards which answers were skewed, i.e., the positive end of the scale. In neither of those studies was clear experimental evidence provided as to the specific cause of the differences, but both suggested differences in aural vs. visual communication. The results of the current study add to the challenge of finding an explanation for such telephone and mail differences. Instead of looking for one simple explanation based upon a general difference between visual and aural communication, it now seems important that the manner in which visual languages are portrayed to respondents must also be considered in attempting to explain such differences. The recent trend towards greater use of mixed-mode surveys in which researchers attempt to survey some members of a population via one mode, and some by another, suggest that a priority for future research is to understand the fundamental causes of such modal differences.

The findings reported here may also have significant implications in how Internet surveys are conducted. For example, the location of instructions after response categories is a phenomenon we have observed on many Internet surveys and is probably an undesirable construction practice. In addition, we have frequently observed scalar formats that use double- and triple-banking, a practice encouraged by the landscape orientation of web pages. Further, the use of HTML programming, in which boxes are reserved for questions in which people can check multiple answers, has encouraged the use of Check-All-That-Apply formats on the Internet which the data reported here suggests is an undesirable practice. However, great care should be taken in extrapolating results from these tests on paper questionnaires to electronic surveys, and it is important that these issues be researched in Internet surveys themselves.

The hypotheses evaluated here are only a few of many that might be tested. Only one symbolic change (addition of an arrow) was evaluated and no independent numerical changes were evaluated for this paper. Although seven independent graphical changes – ranging from size of the open-ended answer space to unequal distance between answer choices – additional manipulations of graphical languages need to be tested in other experiments to fully understand the effects of all aspects of graphical language on respondent behavior.

Based on the experimental evidence presented here and in previous writings (Redline and Dillman, 2002; Redline et al., 2002), it is apparent that survey questions consist of much more than words. Future research needs not only to test the current hypotheses on other populations using different substantive items, but also needs to evaluate other manipulations, e.g., the role of numbers in getting people to answer questions in a prescribed order, the role of consistency in the use and display of symbols throughout questionnaires, effects of figure/ground variations on whether information gets processed by respondents, and how spacing affects question comprehension. Without such research it will be difficult to develop more general principles of how questions should appear on the pages of questionnaires. Such work is essential for taking us beyond our current stage of understanding survey responses by verbal language alone, to beginning an understanding of how verbal and nonverbal languages work together to form a stimulus to respondents.

REFERENCES

- Beatty, Paul and Douglas Herrmann. (2002) "To Answer or Not to Answer: Decision Processes Related to Survey Item Nonresponse." In Groves, Robert, Don A. Dillman, John Eltinge and Roderick J. A. Little (eds.), *Survey Nonresponse*. Wiley: New York, pp. 71-86.
- Dillman, Don A. (2000) *Mail and Internet Surveys: The Tailored Design Method*. 2nd Edition. New York: John Wiley and Sons, Inc.
- Dillman, Don A. and Lisa Carley-Baxter. (1999) Unpublished data. Social and Economic Sciences Research Center, Washington State University, Pullman, WA.
- Dillman, Don A., Glenn Phelps, Robert Tortora, Karen Swift, Julie Kohrell, and Jodi Berck. (2001) "Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response, and the Internet." Unpublished paper presented at Annual Conference of American Association for Public Opinion Research, Montreal, Canada.
- Israel, G. D. and C. L. Taylor. (1990) "Can response order bias evaluations?" *Evaluation and Program Planning*, 123:1-7.
- Jenkins, Cleo R. and Dillman, Don A. (1997) "Towards a Theory of Self-Administered Questionnaire Design." In Lyberg, et. al. (eds.), *Survey Measurement and Process Quality*. New York: John Wiley and Sons, Inc.
- Kahneman, D. (1973) *Attention and Effort*. Englewood Cliffs, NJ: Prentice Hall.
- Krosnick, Jon A., Sowmya Narayan, and Wendy R. Smith. (1996) "Satisficing in Surveys: Initial Evidence." *New Directions for Evaluation*, 70:29-44.
- Krosnick, J., and D. F. Alwin. (1987) "An evaluation of a cognitive theory of response-order effects in survey measurement." *Public Opinion Quarterly*, 51:201-219.
- Redline, Cleo and Don A. Dillman. (2002) "The Influence of Alternative Visual Designs on Respondents' Performances with Branching Instructions in Self-Administered Questionnaires." In, R. Groves, D. Dillman, J Eltinge, and R. Little (eds), *Survey Nonresponse*. New York: John Wiley and Sons, Inc.
- Redline, Cleo, Don A. Dillman, Aref Dajani, and Mary Ann Scaggs. (2002) "The Effects of Altering the Design of Branching Instructions on Navigational Performance in Census 2000." In Proceedings of the Section on Survey Methods, The American Statistical Association.
- Rockwood, Todd, Robert L. Sangster, and Don A. Dillman. (1997) "The Effect of Response Categories on Survey Questionnaires: Context and Mode Effects." *Sociological Methods and Research*, 26:118-140.

Rothwell, Naomi D. (1985) "Laboratory and Field Response Research Studies for the 1980 Census of Population in the United States." *Journal of Official Statistics*, 1 (2): 137-157

Sawyer, Scott and Don A. Dillman. (2002) "How Graphical, Numerical, and Verbal Languages Affect the Completion of the Gallup Q-12 on Self-Administered Questionnaires: Results from 22 Cognitive Interviews and a Field Experiment." Social and Economic Sciences Research Center Technical Report 02-26, Pullman, WA.

Schwarz, Norbert and Seymour Sudman. (1992) *Context Effects in Social and Psychological Research*. New York: Springer-Verlag.

Schwarz, N., H.J. Hippler, B. Deutsch, and F. Strack. (1985) "Response scales: Effects of category range on reported behavior and subsequent judgments." *Public Opinion Quarterly*, 49:388-395.

Sless, David. (1994) "Public Forums: Designing and Evaluating Forms in Larger Organizations." *Proceedings of Public Graphics*, pp. 9.2-9.25. Lunteren, The Netherlands.

Smith, Tom W. (1993) "Little Things Matter: A Sampler of How Differences in Questionnaire Format Can Affect Survey Responses." GSS Methodological Report No. 78. National Opinion Research Center, Chicago, IL.

Stember, Charles Herbert. (1956) "The Effect of Field Procedures on Public Opinion Data." Unpublished Ph.D. dissertation, Columbia University.

Sudman, Seymour, Norman Bradburn, and Norbert Schwarz. (1996) *Thinking About Answers*. San Francisco: Jossey-Bass.

Wallschlaeger, C. and C. Busic-Snyder. (1992) *Basic Visual Concepts and Principles for Artists, Architects, and Designers*. Dubuque, IA: William C. Brown Publishers.

Wright, P. and P. Barnard. (1975) "Just fill in this form – A review for designers." *Applied Ergonomics*, 6:213-220.

Wright, P. and P. Barnard. (1978) "Asking Multiple Questions about Several Items: The Design of Matrix Structures on Application Forms." *Applied Ergonomics*, 9:7-14.

Figure 1: Question formats used to test hypotheses on symbolic, graphical, verbal, and numerical changes in questionnaire formats.

Symbolic Language Change

Hypothesis 1 – Addition of arrow to identify subordinate question.

A. After finishing school at Washington State University, where do you hope to live?

- Eastern Washington
 Somewhere else → Where? _____

B. After finishing school at Washington State University, where do you hope to live?

- Eastern Washington
 Somewhere else Where? _____
-

Graphical Language Changes

Hypothesis 2 – Use of large answer spaces on open-ended answers. First of 3 experimental items.¹

A. Why did you choose to attend Washington State University?



B. Why did you choose to attend Washington State University?



Hypothesis 3 – Location of special instruction before vs. after answer categories.

A. Have one-on-one meetings with professors contributed significantly to your WSU education?

If you haven't had many one-on-one meetings, just skip to Question 9.

- Yes
 No

B. Have one-on-one meetings with professors contributed significantly to your WSU education?

- Yes
 No

If you haven't had many one-on-one meetings, just skip to Question 9.

Hypothesis 4 – Linear vs. nonlinear scale layout (1 of 2 items).

A. Overall, how would you rate the quality of education that you are getting at WSU?

- Excellent Good Poor
 Very good Fair

B. Overall, how would you rate the quality of education that you are getting at WSU?

- Excellent
 Very good
 Good
 Fair
 Poor

Figure 1: Question formats used to test hypotheses on symbolic, graphical, verbal, and numerical changes in questionnaire formats. (Continued)

Hypothesis 5 – Equal vs. unequal spacing between response boxes. First of 2 experimental items.¹

A. What percentage of WSU students do you think use the Student Recreation Center?

- Less than one-fourth of WSU students
- About half of WSU students
- About three-fourths of WSU students
- More than three-fourths of WSU students
- No opinion

B. What percentage of WSU students do you think² use the Student Recreation Center?

- Less than one-fourth of WSU students
- About half of WSU students
- About three-fourths of WSU students
- More than three-fourths of WSU students
- No opinion

Hypothesis 6 – Reversal of response category order for check-all-that-apply question.

A. Which of the following descriptions do you feel describe Washington State University? Check all that apply.

- Farm/Agriculture school
- Party School
- Electronic or “Wired” university
- Competitive in Pac 10 Sports
- Conservative university
- Politically charged/socially conscious
- Religious
- Outdoors oriented
- World class university
- Diverse

B. Which of the following descriptions do you feel describe Washington State University? Check all that apply.

- Diverse
- World class university
- Outdoors oriented
- Religious
- Politically charged/socially conscious
- Conservative university
- Competitive in Pac 10 Sports
- Electronic or “Wired” university
- Party school
- Farm/Agriculture school

Verbal Language Changes

Hypothesis 7 – Removal of word labels for mid-scale categories. First of 3 experimental items.¹

A. How do you consider Washington State University as a place to go to school?

- 1 Very Desirable
- 2 Somewhat Desirable
- 3 Neither Desirable nor Undesirable
- 4 Somewhat Undesirable
- 5 Very Undesirable

B. How do you consider Washington State University as a place to go to school?

- 1 Very Desirable
- 2
- 3
- 4
- 5 Very Undesirable

Figure 1: Question formats used to test hypotheses on symbolic, graphical, verbal, and numerical changes in questionnaire formats. (Continued)

Hypothesis 8 – Balanced vs. unbalanced word labels for scales.

A. How would you rate the quality of retail shops available to you in the Pullman/Moscow area?

- Excellent
- Very Good
- Good
- Fair
- Poor

B. How would you rate the quality of retail shops available to you in the Pullman/Moscow area?

- Excellent
- Good
- Fair
- Poor
- Very Poor

Graphical and Verbal Language Change

Hypothesis 9 – The effect of changing check-all-that-apply question to yes/no format.

A. Which of the following Cougar varsity sports would you consider yourself to be a fan of? Please check all that apply.

- Men's baseball
- Women's basketball
- Men's basketball
- Women's cross-country
- Men's cross-country
- Men's football
- Women's golf
- Men's golf
- Women's rowing
- Women's soccer
- Women's swimming
- Women's tennis
- Women's track and field
- Men's track and field
- Women's volleyball

B. Do you consider yourself to be a fan of these Cougar varsity sports?

	Yes	No
Men's baseball	<input type="checkbox"/>	<input type="checkbox"/>
Women's basketball	<input type="checkbox"/>	<input type="checkbox"/>
Men's basketball	<input type="checkbox"/>	<input type="checkbox"/>
Women's cross-country.....	<input type="checkbox"/>	<input type="checkbox"/>
Men's cross-country	<input type="checkbox"/>	<input type="checkbox"/>
Men's football.....	<input type="checkbox"/>	<input type="checkbox"/>
Women's golf	<input type="checkbox"/>	<input type="checkbox"/>
Men's golf.....	<input type="checkbox"/>	<input type="checkbox"/>
Women's rowing.....	<input type="checkbox"/>	<input type="checkbox"/>
Women's soccer.....	<input type="checkbox"/>	<input type="checkbox"/>
Women's swimming	<input type="checkbox"/>	<input type="checkbox"/>
Women's tennis.....	<input type="checkbox"/>	<input type="checkbox"/>
Women's track and field.....	<input type="checkbox"/>	<input type="checkbox"/>
Men's track and field	<input type="checkbox"/>	<input type="checkbox"/>
Women's volleyball	<input type="checkbox"/>	<input type="checkbox"/>

Graphical, Symbolic, and Numerical language Change

Hypothesis 10 – The effect of a removal of numbers, separate answer boxes, and linear layout of scale. Replacement of number box.

A. On a scale of 1 to 5, with one being very satisfied and 5 being very dissatisfied, how satisfied are you with the classes you are taking this semester?

- 1 Very Satisfied
- 2
- 3
- 4
- 5 Very Dissatisfied

B. On a scale of 1 to 5, with one being very satisfied and 5 being very dissatisfied, how satisfied are you with the classes you are taking this semester?

Number of your rating

¹Wording and visual layout of other items is available from the authors.

²The query was displayed on one continuous line.

Table 1. Results from Symbolic Language Tests: Hypothesis 1.

Hypothesis 1. Percentage of respondents answering subordinate questions without arrow.

Responses	Arrow	No Arrow	Where question?	Arrow	No Arrow
(n)	509	518	Total mentions (n)	428	435
Eastern Washington	11.8	15.3	Mentions from eligible	93.9	91.5
Somewhere else	82.5	83.4	Mentions from noneligible	2.1	0.5
No answer checked	3.9	1.0	No response	4.2	8.0
No answer checked but answered where	1.8	0.4			
Total	100%	100%		100%	100%
Chi-square	$x^2 = 6.6$	$p = .036$		$x^2 = 5.2$	$p = .02$

Table 2. Results from Graphical Language: Hypotheses 2-6.

Hypothesis 2. Mean number of words and themes reported by respondents in larger vs. smaller answer spaces.

Responses	Average # of (n)	Larger Space	Smaller Space	Difference of Means	
Why attend WSU	Words	13.3	9.7	t=-6.5	p=.000
	Themes	2.0	1.8	t=2.7	p=.01
Description of advisor	Words	12.9	6.6	t=12.9	p=.000
	Themes	2.1	1.7	t=8.0	p=.01
Additional recreational activities	Words	12.0	10.2	t=1.8	p=.040
	Themes	1.5	1.4	t=0.7	n.s.

Hypothesis 3. Percent of respondents choosing each answer when instruction to skip located before vs. after answer boxes.

Responses	Instructions Before	Instructions After
(n)	519	522
Yes	54.7	54.9
No	19.1	40.3
Missing	26.2	4.4
Total	100%	100%
Chi-square	$x^2 = 25.9$	$p < .0001$

Hypothesis 4. Percentage of respondents choosing response categories for linear vs. nonlinear layout of scalar responses.

	Question 1		Question 2	
	Linear	Nonlinear	Linear	Nonlinear
(n)	518	517	519	518
(1) Excellent	11.4	11.0	(1) Very Satisfied 42.0	38.8
(2) Very Good	48.8	42.4	(2) Somewhat Satisfied 49.5	51.2
(3) Good	31.3	40.4	(3) Somewhat Dissatisfied 6.9	8.9
(4) Fair	7.0	5.4	(4) Very Dissatisfied 1.5	1.2
(5) Poor	1.5	0.8		
Total	100%	100%	100%	100%
Mean	2.4	2.4	1.7	1.7
Difference of Means	$t = 0.8$	$p = .206$	$t = 1.1$	$p = .146$
Chi-square	$x^2 = 10.8$	$p = .029$	$x^2 = 2.3$	$p = .509$

Table 2. Results from Graphical Language: Hypotheses 2-6. (Continued)

Hypothesis 5. Percentage of respondents choosing response categories with unequal vs. equal spacing between categories.

	<u>Question 1</u>			<u>Question 2</u>	
	Equal	Unequal		Equal	Unequal
(n)	474	471		510	514
To have a life partner with whom you have a satisfying relationship	30.8	37.6	(1) Less than one-fourth of WSU students	17.3	19.1
Enjoy your work	43.9	42.9	(2) About half of WSU students	47.5	46.5
Earn a high income	10.6	7.9	(3) About three-fourths of WSU students	24.7	22.6
Raise a family	14.8	11.7	(4) More than three-fourths of WSU students	4.3	4.7
			No opinion	6.3	7.2
Total	100%	100%		100%	100%
			Mean	2.3	2.3
			Difference of Means	t=.07	p = .472
Chi-square	x² = 6.8	p = .079		x² = 1.4	p = .844

Hypothesis 6. Percentage of respondents choosing answer choices for Check-All-That-Apply questions when order of presentation is reversed.

<u>Responses</u>	<u>Order on Left</u>	<u>Reverse Order</u>	<u>Chi²</u>	<u>P</u>
(n)	519	523		
Farm/Agriculture school	53.0	54.3	.18	.670
Party school	49.7	52.8	.98	.323
Electronic or "Wired" university	53.0	73.2	45.0	.000
Competitive in Pac-10 sports	39.3	51.4	15.4	.000
Conservative university	14.1	19.1	4.8	.029
Politically charged/socially conscious	17.9	30.2	21.1	.000
Religious	5.6	9.2	4.8	.028
Outdoors oriented	24.7	40.2	28.1	.000
World class university	26.0	33.8	7.6	.006
Diverse	40.9	49.0	6.9	.009
Overall test: x² = 100.6, p = .000				
Mean # of responses/respondent	3.24	4.13		
t-test = -7.386, p = .000				

Table 3. Results of Verbal Language Tests: Hypotheses 7-8.

Hypothesis 7. Percentage of respondents choosing response categories on polar-point vs fully labeled scales.

Desirability Scale (n)	Question 1		Question 2		Question 3 ¹	
	Fully Labeled	Polar-Point	Fully Labeled	Polar-Point	Fully Labeled	Polar-Point
(1) Very Desirable	39.0	25.0	15.9	15.2	31.2	34.4
(2) Somewhat Desirable	44.8	38.4	35.9	29.5	46.1	37.7
(3) Neither Desirable nor Undesirable	11.0	23.4	23.6	28.4	14.3	21.4
(4) Somewhat Undesirable	4.6	9.6	16.9	17.3	7.3	5.1
(5) Very Undesirable	0.6	3.7	7.7	9.7	1.2	1.4
Total	100%	100%	100%	100%	100%	100%
Mean	1.8	2.3	2.6	2.8	2.0	2.0
Difference of Means	t = 7.7	p = .000	t = 1.7	p = .045	t = 0.0	p = .499
Chi-square	x² = 62.4	p = .000	x² = 6.8	p = .146	x² = 14.7	p = .005

¹Question 3 scale read from Very Confident to Very Unconfident

Hypothesis 8. Percentage of respondents choosing labeled categories on balanced vs. unbalanced scales.

Category Labels (n)	Question 1		Question 2	
	Unbalanced	Balanced	Unbalanced	Balanced
Excellent	(5) 1.8	(5) 2.7	(5) 13.7	(5) 20.0
Very Good	(4) 9.5	N/A	(4) 28.3	N/A
Good	(3) 25.9	(4) 20.0	(3) 37.9	(4) 47.1
Fair	(2) 36.2	(3) 40.1	(2) 16.0	(3) 25.4
Poor	(1) 26.7	(2) 22.5	(1) 4.1	(2) 5.2
Very Poor	N/A	(1) 14.8	N/A	(1) 2.3
Total	100%	100%	100%	100%
Mean	2.2	2.7	3.3	3.8
Difference of Means	t = 7.9	p = .000	t = 7.6	p = .000
Chi-square	x² = 70.2	p = .000	x² = 74.2	p = .000

Table 4. Test of Graphical Plus Verbal Language Change: Hypothesis 9.

Hypothesis 9. Percent of respondents checking or choosing “Yes” answers for Yes/No vs. Check-All-That-Apply formats.

Responses	Yes/No Format (n=523)			Check-All-That-Apply Format (n=519)	Chi-Square (Check vs. Yes)	P
	Yes	No	No Answer	Checked		
Men’s baseball	42.3	48.6	9.2	34.1	7.31	.007
Women’s basketball	18.6	68.8	12.6	8.9	19.82	.000
Men’s basketball	43.4	48.4	8.2	33.5	10.69	.001
Women’s cross-country	9.0	77.2	13.8	3.9	10.78	.001
Men’s cross-country	9.6	76.7	13.6	3.7	13.64	.000
Men’s football	79.7	17.6	2.7	80.5	0.11	.744
Women’s golf	6.5	80.1	13.4	2.5	8.98	.003
Men’s golf	10.1	76.9	13.0	5.4	7.91	.005
Women’s rowing	17.4	69.4	13.2	5.4	33.44	.000
Women’s soccer	33.5	54.7	11.9	19.7	24.98	.000
Women’s swimming	14.7	72.3	13.0	6.2	19.23	.000
Women’s tennis	12.2	74.8	13.0	5.8	12.67	.000
Women’s track and field	18.4	69.0	12.6	11.9	8.21	.000
Men’s track and field	22.4	65.6	12.0	12.5	17.11	.000
Women’s volleyball	37.3	53.0	9.7	30.4	5.43	.020

Overall test: $\chi^2 = 4.6$ p = .032

**Mean responses
per respondent**

4.3

3.0

t-test = -6.954, p = .000

Table 5. Results for Test of Simultaneous Graphical, Symbolic, and Numerical Change: Hypothesis 10.

Hypothesis 10. The effect of a number box vs. linear display of answer choices.

Responses	Question 1		Question 2		Question 3 ¹	
	Polar-Point	Number Box	Polar-Point	Number Box	Polar-Point	Number Box
(n)	517	513	517	512	506	466
(1) Very Satisfied	15.9	9.8	10.8	6.3	15.2	14.2
(2)	43.5	29.2	47.6	35.4	31.8	23.8
(3)	31.0	34.5	31.3	33.8	31.6	29.6
(4)	7.7	21.3	9.5	20.9	15.4	21.7
(5) Very Dissatisfied	1.9	5.3	0.8	3.7	5.9	10.7
Total	100%	100%	100%	100%	100%	100%
Mean	2.4	2.8	2.4	2.8	2.7	2.9
Difference of Means	t = 7.7	p = .000	t = 6.9	p = .000	t = 3.5	p = .000
Chi-square	x² = 63.4	p = .000	x² = 48.1	p = .000	x² = 18.0	p = .001

¹Response categories to this item were (1) Outstanding and (5) Terrible.