Designing Response Scales in an Applied Setting

Wendy Davis, Tracy R. Wellens and Theresa J. DeMaio, U.S. Census Bureau
Wendy Davis, U.S. Census Bureau, SRD, Rm. 3125-4, Washington DC, 20023

Key Words: response scales, show cards, telephone surveys

INTRODUCTION

The Diet and Health Knowledge Survey (DHKS) questionnaire presents many questionnaire design challenges. The questionnaire collects data about peoples' knowledge and attitudes about various health- and nutrition-related issues. The design of the response scales is an important ingredient in determining the quality of the data collected. In addition, more practical issues also come into play in developing a questionnaire that is efficient to administer in the field. Literature addressing response scales and data quality is often not consistent in its conclusions. Furthermore, practical concerns such as ease of administration, are often left completely unaddressed. The objective of this study was to further examine some of the data quality issues with response scales as well as incorporate measures of more practical concerns.

LITERATURE REVIEW

This section is divided into three parts. Each sub-section briefly reviews literature addressing one of three issues important to consider in the design of a response scale.: the number of scale points, the extent of verbal labeling, and the use of branching scales. Some other practical concerns are also considered. At the outset, we should note that much of the response scale literature deals with multiple indicators, combining them into indices to measure broad concepts. The questions in the DHKS, however, are intended as measures of specific pieces of information and are analyzed individually.

Number of Scale Points

For decades it has been widely accepted that scales between 3 and 9 points are optimal in terms of capturing the most variance without suffering losses in reliability for any single survey item (Bendig, 1953; Bendig and Hughes, 1953; Miller, 1956; Finn, 1972; Ramsey, 1973; Cox, 1980; Churchill and Peter, 1984; Alwin, 1992).

In an applied setting, the exact number of points used in a scale is a concern. As noted above, one or two points may make a difference in data quality in terms of both the extent of the true variance captured and the reliability of the measure. Other, more practical, concerns about the exact number of scale points involve increased administration time in a telephone interview, the perceived difficulty of administration from the interviewers' and the respondents' perspectives, the increased opportunity for interruptions and other break-offs by respondents, and the added complexity of formatting and printing. Whereas perceived difficulty of administration, interruptions, and break-offs may translate into interviewer error, administration time and the formatting and printing of an instrument translate directly into survey costs, a very big issue for government agencies.

Extent of Verbal Labeling

There are many reports in the literature of enhanced data quality, specifically enhanced reliability, when more of the response scale points are labeled (Bendig, 1953; Peters and McCormick, 1966; Zaller, 1988). The theory behind this is that verbal labels communicate information to respondents that is ambiguous or absent without the labels (Schwarz et al., 1991; Schwarz et al., 1987, 1995). Thus, the more labeling included on a scale, the greater the amount of information available to respondents to interpret and use for responding. The increased information provided by the labels makes it more likely that respondents will use the scale consistently, or with greater reliability. However, there have also been studies presenting contrary results (e.g., Andrews, 1984; Krosnick et al., 1993).

Although the literature in this area is inconsistent, the popular opinion among survey researchers seems to be that fully labeled scales communicate more information and thus yield more reliable, higher quality data than partially labeled scales. Some of the more practical issues related to interviewers using a fully labeled scale, however, have been left largely unaddressed. For example, how are administration time and interviewer error due to break-offs and interruptions affected by fully labeled scales? One would expect that it takes more time to read five or more labels than it does to read to two or three labels. But is the difference in administration time significant, and will it significantly increase costs for the survey? Similarly, if it does take more time to read verbally labeled scales, do respondents interrupt interviewers more

often and as a result hear only a portion of the response options? Do interviewers tend not to repeat the scales as often when they are fully labeled for the same reason? If there are more break-offs and interruptions, the effect on administration time may be insignificant, but data quality may decrease. For example, if respondents begin to interrupt the interviewer with their answer before the whole scale has been read to them, they only hear and may therefore only use the first few points on the scale.

Branching versus Standard scales

The response task in a telephone interview and in a face-to-face interview are slightly different. In a personal visit interview, show cards are often provided to help respondents use and remember a scale, especially when the scale is used for a series of items. A show card is not as convenient to use in telephone surveys, especially surveys conducted by random digit dialing.

One technique developed to make the telephone response task more comparable to a personal visit interview with show cards is called branching, (Groves, 1979). This technique is particular to bipolar scales (that is, items that were measured on a scale with opposite dimensions.) Branching changes the respondent's task of choosing the direction and the strength of their attitude from a single step to a two-step process. Inconsistent findings have been reported between studies in regards to the affect on data quality when using branching scales (e.g., Groves and Kahn, 1979; Albaum and Murphy, 1988; Miller, 1984).

Other Practical Issues

As we have noted, one method for decreasing the cognitive demands of the response task, especially for longer fully labeled scales, is to provide respondents with a visual aid or "show card." A visual aid allows respondents to refer back to the scale as needed, rather than forcing them to retain the scale in memory. As a result, the cognitive demands placed on respondents in an interview are decreased. In a telephone interview, the respondent's ability to store and maintain a response scale in memory for a series of questions may be hindered when respondents are faced with other distractions during the interview (e.g., television, radio, other people). As a result, the respondent either requests or the interviewer finds it necessary to repeat the scale more than once per question. Thus administration time may increase, interviewer errors may increase, etc.

The use of show cards in a telephone interview may require substantial changes to the survey procedures. The justification for incurring the potential increased cost is that including show cards might improve data quality. This study directly examines the effect of show cards on data quality in a telephone interview.

METHODS

To examine the issues discussed in the literature review, we designed an experiment to measure the effects that the number of points on a scale, the extent of verbal labeling, and the type of scale (branching or standard) might have on data quality as well as on other practical concerns such as ease of administration and administration time. Each of these manipulations were done using three different subjective measures: the extent to which one disagrees/agrees with a statement; the importance of a statement, and; the frequency of a behavior. We operationalized the scale characteristic variables as follows:

- There were two treatments for the number of scale points. The short scale was always four points; the longer scale had five or six points depending on the measure.
- There were two treatments for the extent of verbal labeling. The first was partially labeled (that is, only the endpoints had verbal labels); the second was fully labeled (that is, all the scale points were labeled).
- The type of scale was only manipulated for the bipolar items, and there were two levels. The first was a standard scale, and the second was a branching scale that obtained information in two separate questions.
- There were two levels of the show card condition. In one group, respondents received a postcard that had the response scales printed on the back for reference during the interview. The other group of respondents merely received a postcard that reminded them of the time of their interview.

This study used a between subjects design for each of the manipulated scale characteristics. Across experimental conditions, all respondents received identical questions, only characteristics of the response scale differed by condition.

Reliability was one of the measures of data quality. Therefore, we built into the questionnaire repeated administrations of the same items. Within each experimental manipulation, respondents were asked a series of questions using each type of subjective measure (e.g., disagree-agree, importance, and frequency) at three different times during the interview. The first administration was to get respondents familiar with the experimental scale (e.g., short, partially labeled, standard scale). The second administration was one of the three

target question series. And the third administration was the identical target question series, again using the same experimental scale. Eleven to twelve questions were asked in between the two administrations of each target question series.

Three hundred and eleven respondents were recruited. Of those, 300 completed the telephone interview. Respondents were recruited to fit into one of two education categories: high, having at least some college, or; low, having a high school degree or less. There was a total of 151 high education respondents and 149 low education respondents. Respondents within each of the education groups were randomly assigned to one of the experimental conditions. All interviews were conducted over the phone by Census Bureau research staff. Respondents were paid $10 for their participation.

RESULTS

Several different types of analysis were conducted to evaluate the different scale characteristics being tested. Data quality was examined with mean score analysis and reliability analysis. Two other analytic techniques were used to evaluate field related issues: behavior coding and a regression model using administration time as a response variable.

Each measure (i.e., disagree-agree, importance, frequency) was analyzed separately. As noted earlier, mean score and reliability analysis were done separately for each item within a measure since the items are intended to measure independent and distinct concepts. As a result it was necessary to define criteria for whether or not a result was generalizable to the whole series of items within a measure. Criteria for generalizability was set as follows:

- Mean score analysis. Regression models were run separately for the long and short versions of the scale for each item in a series. All of the experimental conditions except length of scale were included as predictors. The response variable was the value respondents gave as their answer to an item. If at least 50% of the items in a question series had an overall F value significant at $p<0.05$, then the result was considered generalizable to the whole series.
- Reliability analysis. Test-retest correlations (Pearson's r^2) were calculated and then transformed using Fisher's Z. A z-score was calculated for significance testing. If at least 50% of the items had z-scores significant at $p<0.05$ and an average effect size (e.s.) of at least 0.30 (Cohen, 1988), the result was considered generalizable.

Description of Scales

Each measure (e.g., disagree-agree, importance and frequency) had a short and long version of the scale. Since the disagree-agree measure used a bipolar scale, it also had a branching version. For all three measures the short version of the scale contained 4-points. The low end of the scale was the most negative end. For example, "strongly disagree" was the low end of the disagree-agree scale "not at all important" was the low end of the importance scale, and; "never" was the low end of the frequency scale. The high end had the positive responses (e.g., "strongly agree," "very important," "often/always"). In the partially labeled condition respondents were given the end points and their labels, as well as being reminded that the middle points represented something in between. For example, the 4-point, partially labeled disagree-agree scale was read as: "Choose a number between 1 and 4 with 1 being `strongly disagree,' 4 being `strongly agree' and 2 and 3 being something in between."

The long version of the importance and frequency scales were each 5-point scales. For the importance scale, the extra point was added in the middle of the scale, so the end points were not affected. The extra point was added to the end of the frequency scale, however, so the labels in the partially labeled condition were different between the short and long versions of the scale.

The long version of the disagree-agree scale was 6 points rather than 5. Adding one point to the scale, such as "neither agree or disagree," didn't seem any different from a "no opinion" category which was already part of the response choices. So two points were added to the middle of the scale instead of adding one-point.

The branching version of the disagree-agree asked the direction of the respondents' opinion first, and then the magnitude in a second question. The magnitude question had 4 points. It read "slightly, somewhat, mostly, strongly" in the fully labeled condition, and went from 1 to 4 in the partially labeled condition. The first question, which asks the direction of their opinion, was identical in both the fully and partially labeled conditions.

Response Distributions

Overall, each item for all 3 measures had negatively skewed response distribution. The disagree-agree items and the importance items were the most severely skewed. Of the ten disagree-agree items, the negative skew was less than -1.25 for 9, 7 and 5 items in the branching, long and short versions of the scale respectively. Of the 6 importance items, the negative skew was less than -1.25 for 4 of the items in the long version of the scale, and 3 of the items in the short

version. However, none of the 5 frequency items in either the long or short versions of the scale had a negative skew of -1.25 or less.

Similarly, over 63% of responses on average across the ten disagree-agree items fell into the top category on the scale (e.g., strongly agree) for the branching, long and short versions of the scale. Close to 60% or more of responses across the 6 importance items, on average, fell into the top category for the long and short versions of that scale (e.g., very important). But on average across the 5 frequency items, only 30% of responses or less fell into the top category (e.g., always/often) for both the long and short versions of the scale.

There was not much variance in the data as a result of these distributions, especially for the disagree-agree and importance measures. Thus, differences between experimental conditions were hard to detect.

Branching

Only the disagree-agree measure included a branching manipulation, so only that measure is discussed here. Though the mean scores were not affected by the type of scale, reliability was. Seven of ten items had higher reliability in the standard condition with an average effect size of 0.40 ($z>=1.96$; $p<0.05$).

Previously we noted that the findings in the response scale literature were inconclusive in regards to the data quality of branching versus standard scales. However, the most recent work in this area (Krosnick et al., 1993) suggested that branching scales may be more reliable. Since our results for the disagree-agree scale were in contradiction to this assertion we looked to the behavior coding data for an explanation. Respondents interrupted interviewers before they read the scale on average over 50% more often in the two step (branching) version than they did in the one step version. In addition, approximately 44% of respondents answered the branching version at the first step of the question rather than waiting to hear the scale at the second step. This suggests that for this series of questions, standard scales produce higher data quality.

Not surprisingly, the questionnaire containing the branching scales took significantly long to administer than the shorter scale (Dunnett's $T=2.225$, $p<0.05$). There was no difference in administration time between the branching and the longer scale.

Extent of Labeling

Mean scores were not affected by the extent of labeling for either the short or long versions of the disagree-agree measure. Nor were mean scores affected in the long versions of the importance and frequency measures. However, labeling did affect mean scores for the short versions of the importance and frequency measures, but in opposite directions. Responses were significantly higher in the fully labeled condition for 4 of 5 frequency items ($F>=7.64$, $p<.006$). On the other hand, though only a marginally consistent finding, responses were significantly lower in the fully labeled condition for 2 of 6 importance items ($F=16.5$, $p<.0001$). Since social desirability is a concern in this survey, lower mean responses are assumed to be better. However, in the case of the frequency measure, the higher responses in the fully labeled condition may be a result of a problem with the design rather than the scale manipulation.

In the partially labeled condition, the last points on the scale the respondents heard were the middle points. In comparison, in the fully labeled condition, the last point on the scale that respondents heard was the end point " often." Thus, the differences between the partially and fully labeled condition may actually reflect what Krosnick (1992) refers to as a "recency effect" -- respondents in telephone surveys tend to respond more often with the last category they hear.

Reliability was also significantly affected by labeling, but in different directions across measures. Reliability was higher in the fully labeled condition for 7 of the 10 disagree-agree items (average e.s.=0.73), but only for respondents with at least some college education. In comparison, reliability was lower in the fully labeled condition for 3 of the 6 importance items (average e.s.=0.40) but only for respondents in the less educated group. The former result is what we had expected. The latter result for the importance measure is somewhat surprising. There is nothing in the behavior coding data to explain the differences. It may be that the particular verbal labels for this measure ("not at all important, not too important, somewhat important, *important*, very important") were perceived as wordy and, as a result, difficult for the less educated respondent to understand and use.

There was no effect of labeling on reliability for any of the frequency items.

The extent of labeling did make a significant difference in terms of administration time. However, in the opposite direction than we expected. Partially labeled scales took longer to administer than fully labeled scales ($F=10.12$, $p<0.0016$). Behavior coding data revealed, again counter to what we expected, that partially labeled scales were being read less often, and were being interrupted more frequently than fully labeled scales. This is surprising given that they took longer to administer. However, the actual wording of the partially and fully labeled scale descriptions differed in that the

partially labeled description contained more words than the fully labeled description.

Length of Scale

There were only a few, small effects on data quality by length of scale. In fact, reliability was never affected by length of scale for any measure. This is not that surprising given the slight difference between the shorter and longer versions of the scales. However, there were a few contradictory differences in mean responses between short and long scales across measures. In the shorter version of the importance scale, mean scores seemed to be increased due to social desirability for 2 of 6 items ($F=16.5$, $p<0.0001$), but there was no effect of social desirability in the longer version. Similarly, in the short version of the frequency scale, mean scores were affected by a recency bias for 4 of 5 items ($F>=7.64$, $p<0.006$), but not in the long version.

On the other hand, a response effect due to education occurred in the longer version of the frequency scale but not in the shorter. Less educated respondents gave higher answers than more educated respondents ($F>=4.54$, $p<0.036$) in the longer version of the scale.

There was no effect of administration time by length of scale.

Presence of Show Card

With one exception having a show card had no affect on data quality either in terms of reliability or mean scores. The exception was that reliability was improved for the less educated group in 4 of the 6 importance items when they had a show card (average effect size of 0.45).

The presence of a show card also had no direct effect on administration time. However, when included as an interaction with scale type and length (e.g., branching, long or short), the scale card had a mitigating effect on administration time. The branching scale which took significantly longer to administer without a show card than the long and short versions of the scale, was not significantly longer to administer with a show card ($F=5.41$, $p<0.005$).

CONCLUSIONS

There were really only two clear results in this study. First, as compared to branching scales, standard scales were more reliable, less prone to interviewer/respondent error, and took less time to administer. Since they took less time to administer they may also be less costly to administer.

Second, having a show card to use during the interview, had no profound or consistent affect on data quality, interviewer performance, or even administration time. This would suggest that it is probably not worth while to include a show card in a telephone survey if doing so would significantly alter survey procedures and increase costs.

In regards to the other scale manipulations, the results are not as clearly defined. This ambiguity is in part due to some problems in the study design. First, we are assuming that social desirability is affecting responses, and as a result, distributions with a lower mean score are "best." However, we do not know the true distribution, so this may be an incorrect assumption.

Second, there is a potential confound in the design of the fully and partially labeled scale manipulation. Krosnick (1992) has reported finding a recency bias in telephone surveys in which the last points heard are more often given as a response. In the fully labeled condition the last points heard are the end points, but in the partially labeled condition, the last points heard are the middle points. Thus, it is difficult to determine whether lower mean scores in the partially labeled condition are due to the labeling or to a recency bias. In addition, the absence of a difference between the two labeling conditions could be a result of the flaw in the design masking such a difference (e.g., fully labeled scales having lower mean scores).

Lastly, the difference between the short and longer versions of the scale was quite small for all three measures. It isn't surprising that there were so few significant differences. In fact, it is harder to explain why there were any differences, especially since there wasn't any effect on interviewer behavior or administration time. A larger difference between the two scale lengths would make it more likely that true differences would be detected.

REFERENCES

Albaum, G. and Murphy, B. D. (1988) "Extreme Response on a Likert Scale" Psychological Reports, 63, 501-502

Alwin, Duane F. (1992) "Information Transmission in the Survey Interview: Number of Response Categories and the Reliability of Attitude Measurement" in Sociological Methodology, 1992 by Peter V. Marsden (ed); The American Sociological Association, Blackwell Publishers

Andrews, F. M. (1984) "Construct Validity and Error Components of Survey Measures: A Structural Modeling Approach" Public Opinion Quarterly, 48, 409-442

Bendig, A. W. (1953) "The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and the Number of Categories on the Scale" The Journal of Applied Psychology, 37(1)

Bendig, A. W., and Hughes, J. B. (1953) "Effect of Amount of Verbal Anchoring and Number of Rating-Scale Categories Upon Transmitted Information" Journal of Experimental Psychology, 46(2)

Churchill, G. A., and Peter, J. P. (1984) "Research Design Effects on the Reliability of Rating Scales: A Meta-Analysis" Journal of Marketing Research, vol. XXI, 360-375

Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences (2nd ed.), Lawrence Erlbaum Associates: Hillsdale, NJ

Cox, Eli P. (1980) "The Optimal Number of Response Alternatives for a Scale: A Review." Journal of Marketing Research, vol. XVII, 407-422.

Finn, R. H. (1972) "Effects of Some Variations in Rating Scale Characteristics on the Means and Reliabilities of Ratings" Educational and Psychological Measurement, 32, 255-265

Groves, R. (1979) "Actors and Questions in Telephone and Personal Interview Surveys" Public Opinion Quarterly, 43(2)

Groves, R. and Kahn, R. (1979) Surveys by Telephone: A National Comparison with Personal Interviews, New York, Academic Press

Krosnick, J. and Berent, M. K. (1993) "Comparisons of Party Identification and Policy Preferences: The Impact of Survey Question Format" American Journal of Political Science, vol. 37(3), 941-964

Krosnick, J. (1992) "The Impact of Cognitive Sophistication and Attitude Importance on Response-Order and Question-Order Effects" in Context Effects in Social and Psychological Research by Norbert Schwarz and Seymour Sudman (eds..), Springer-Verlag

Miller, G. (1956) "The magical number seven, plus or minus two" Psychological Review, 63(2), 81-97

Miller, P. (1984) "Alternative Question Forms for Attitude Scale Questions in Telephone Interviews" Public Opinion Quarterly, vol. 48, 766-778

Peters, D. L. and McCormick, E. J. (1966) "Comparative Reliability of Numerically Anchored versus Job-Task Anchored Rating Scales." Journal of Applied Psychology, 50: 92-96

Ramsey, J. O. (1973) "The Effect of Number of Categories in Rating Scales on Precision of Estimation of Scale Values" Psychometrika, 38(4)

Schwarz, N., Hippler, Hans-J. (1995) "The Numeric Values of Rating Scales: A comparison of their impact in mail surveys and telephone interviews." International Journal of Public Opinion Research, 7(1)

Schwarz, N., and Hippler, Hans-J. (1987) "What Response Scales May Tell Your Respondents: Informative Functions of Response Alternatives" in Social Information Processing and Survey Methodology by Hans J. Hippler, Norbert Schwarz, and Seymour Sudman (eds..), Springer-Verlag

Schwarz, N., Knauper, B., Hippler, H. J., Noelle-Neumann, E. and Clark, L. (1991) "Rating Scales: Numeric Values May Change the Meaning of Scale Labels" Public Opinion Quarterly, 55(4)

Zaller, J. (1988) "Vague Minds versus Vague Questions: An Experimental Attempt to Reduce Measurement Error" Paper presented at the Annual Meetings of the American Political Science Association; Washington DC