RESEARCH REPORT SERIES
(Survey Methodology #1996-12)

Report on Results of Response Scale Research on the Diet and Health Knowledge Survey

Wendy Davis
Tracy R. Wellens
Theresa J. DeMaio

Center for Survey Measurement
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report issued: March 29, 2017

Disclaimer: This report is released to inform interested parties of research and to encourage discussion of work in progress. Any views expressed on the methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Abstract

The Center for Survey Methods Research conducted experimental research to evaluate the response scales used in the Diet and Health Knowledge Survey telephone questionnaire. Specifically, three characteristics of the response scale were examined. The length of the scale (i.e., the number of points on the scale), the extent of verbal labeling used, (i.e., partially labeled with only the end points labeled, or fully labeled with all points labeled) and the type of scale format used (i.e., branching with the direction and magnitude of an opinion collected in two questions; or standard with both the direction and magnitude collected in one question). The literature addressing these issues is often inconclusive. For example, several studies examining the type of scale format to use in a telephone interview have contradictory results. Some studies suggest that data quality is higher with a branching scales (Groves, 1979) and other studies suggest data quality is higher with a standard scale format (Miller, 1984). There are similar contradictions regarding the extent of labeling. There is more consensus in the literature regarding the number of points to use on a scale, but a range of points is offered as the optimal number of points (between 3 and 9 points) rather than a specific number of points. The study presented here examined each of these issues plus an additional survey characteristic currently used in the Diet and Health Knowledge Survey interview – providing respondents with a show card to refer to during the interview.

Keywords: response scales, diet, health

Suggested Citation: Wendy Davis, Tracy R. Wellens, Theresa J. DeMaio**. (1996). Report on Results of Response Scale Research on the Diet and Health Knowledge Survey**. *Research and Methodology Directorate, Center for Survey Measurement Study Series (Survey Methodology #1996-12).* U.S. Census Bureau. Available online at <http://www.census.gov/srd/papers/pdf/sm9612.pdf>.

UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

November 19, 1996

Dr. Linda Cleveland
USDA Agricultural Research Service
4700 River Road, Unit 83, RM6C36
Riverdale, MD 20737-1229

Dear Linda,

Enclosed is the final report of the response scale research conducted on the Diet and Health Knowledge Survey questionnaire under reimbursable agreement 47-91-01. Your comments and suggestions on the original draft version were quite useful, especially your editorial comments addressing the reference list. We have incorporated all of your suggestions but one. The only exception is that we didn't create a glossary of terms. Instead we defined within the text of the report all the terms you noted as unfamiliar. It seemed easier for the reader to have the definition as part of the text rather than having to flip through a separate glossary.

Also note that we have changed our recommendation concerning the length of scale. In the draft report you noted an inconsistency between the values in a table and our discussion of the table. In fact, the table values were the correct values and were much more in line with our expectations about the length of the scale. We have revised our recommendations to reflect this.

We have enjoyed working with you and appreciate your involvement and patience with this research project.

Sincerely,

Theresa J. DeMaio
Center for Survey Methods Research

Enclosure

cc:

C. Bowie	(DSD)
T. Wright	(SRD)
E. Martin	(SRD/CSMR)
W. Davis	"
T. Wellens	"

Report on Results of Response Scale Research on the Diet and Health Knowledge Survey

Prepared by

Wendy Davis, Tracy R. Wellens, and Theresa J. DeMaio
U.S. Bureau of the Census

November 18, 1996

EXECUTIVE SUMMARY

The Center for Survey Methods Research conducted experimental research to evaluate the response scales used in the Diet and Health Knowledge Survey telephone questionnaire. Specifically, three characteristics of the response scale were examined.

The length of the scale (i.e., the number of points on the scale), the extent of verbal labeling used, (i.e., partially labeled with only the end points labeled, or fully labeled with all points labeled) and the type of scale format used (i.e., branching with the direction and magnitude of an opinion collected in two questions; or standard with both the direction and magnitude collected in one question). The literature addressing these issues is often inconclusive. For example, several studies examining the type of scale format to use in a telephone interview have contradictory results. Some studies suggest that data quality is higher with a branching scales (Groves, 1979) and other studies suggest data quality is higher with a standard scale format (Miller, 1984). There are similar contradictions regarding the extent of labeling. There is more consensus in the literature regarding the number of points to use on a scale, but a range of points is offered as the optimal number of points (between 3 and 9 points) rather than a specific number of points.

The study presented here examined each of these issues plus an additional survey characteristic currently used in the Diet and Health Knowledge Survey interview -- providing respondents with a show card to refer to during the interview. Three hundred respondents recruited from the Washington DC metropolitan area were randomly assigned into one of the twelve groups shown below:

WITHOUT A SHOW CARD
- short, fully labeled, standard scale
- long, fully labeled, standard scale
- short, partially labeled, standard scale
- long, partially labeled, standard scale
- fully labeled, branching scale
- partially labeled, branching scale

WITH A SHOW CARD
- short, fully labeled, standard scale
- long, fully labeled, standard scale
- short, partially labeled, standard scale
- long, partially labeled, standard scale
- fully labeled, branching scale
- partially labeled, branching scale

Interviews were conducted by telephone between October 1995 and April 1996.

Our results suggest the following:

1) As compared to branching scales, standard scales were more reliable, less prone to interviewer and respondent error, and took less time to administer. Since they took less time to administer they may also be less costly to administer.

2) Having a show card to use during the interview had no profound or consistent effect on data quality, interviewer performance, or even administration time. This would suggest that it is probably not worthwhile to include a show card in this survey if budgets are tight.

3) Though not definitive, the results also suggest that a longer scale may produce better data. Longer scales had a slightly lower proportion of responses in the positive end points (i.e., strongly agree, very important, often) suggesting less extremity bias which is often asociated with social desirability. In addition, in two cases the shorter scale showed evidence of a

recency bias for which respondents answer with the last response choice offered, regardless of the value (Krosnick, 1992).

4) Lastly, the results pertaining to the extent of labeling were contradictory. In some cases fully labeled scales produced higher quality data (as defined for this study), but in other instances partially labeled scales were better. In addition, a flaw in the design made the results on this characteristics even harder to interpret. We suggest using fully labeled scales as a tentative recommendation based on the most recent findings in the literature which indicate that fully labeled scales may produce more reliable data.

Taking all of the above results together, we suggest a fully labeled, standard scale be used with either 5 or 6 points (the number of points depends on the construct being measured, i.e., disagree-agree, importance, frequency). However, we strongly suggest that this scale be field tested before being incorporated into the survey. We also believe that a show card is probably not necessary.

TABLE OF CONTENTS

INTRODUCTION 1

LITERATURE REVIEW 1
- Number of Scale Points 1
- Extent of Verbal Labeling 3
- Branching versus Standard scales 4
- Other Practical Issues 5

METHODS 6
- Experimental Design 6
- Scales 7
- Recruitment and Data Collection Procedures 11

RESULTS and RECOMMENDATIONS 12
- OVERVIEW OF ANALYTIC TECHNIQUES 12
  - Descriptions of Distributions 12
  - Analysis of Variance (ANOVA) 12
  - Reliability Analysis 13
  - Behavior Coding 14
- DISAGREE-AGREE SERIES (Q5 and Q9) 16
  - Description of Scale 17
  - Response Distribution 17
  - Analysis of Variance 18
  - Reliability Analysis 19
- IMPORTANCE SERIES (Q16 and Q19) 21
  - Description of Scale 22
  - Response Distribution 22
  - Analysis of Variance 23
  - Reliability Analysis 24
- FREQUENCY SERIES (Q17 and Q20) 25
  - Description of Scale 26
  - Response Distribution 26
  - Analysis of Variance 26
  - Reliability Analysis 27
- ADMINISTRATION TIME 28
- SUMMARY OF ANALYSIS and RECOMMENDATIONS 29
  - Branching 29
  - Extent of Labeling 30
  - Length of Scale 30
  - Presence of Show Card 31
  - Recommended Scale Type 31
  - General Recommendations 31

REFERENCES 32

TABLES 35

Report of Results of Response Scale Research
on the Diet and Health Knowledge Survey

Wendy Davis, Tracy R. Wellens and Theresa J. DeMaio
U.S. Census Bureau
November 18, 1996

INTRODUCTION

The Diet and Health Knowledge Survey (DHKS) questionnaire presents many questionnaire design challenges. Since 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.

This report presents the results of research conducted by the Census Bureau's Center for Survey Methods Research. In the sections that follow, we first present the results of a literature review that addresses three research issues related to the design of response scales: the number of scale points that should be included in a scale, whether each of the scale points should be labeled or only the end points, and, for bipolar scales, whether all the response categories are presented in a single question or obtained in sequential questions (called "branching"). Then we present the design and results of an experiment that we conducted to investigate these issues in the DHKS questionnaire. Finally, we present recommendations for changes to improve the DHKS questionnaire. Our recommendations fall into two areas--data quality and practical operational concerns.

LITERATURE REVIEW

This section is divided into three parts, dealing with the three response scale issues noted above. 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; Finn, 1972; Ramsey, 1973; Cox, 1980; Churchill and Peter, 1984; Alwin, 1992). Cox (1980) concludes that the optimal number of points on a scale is seven plus or minus two. He notes that two or three response alternatives are generally inadequate in terms of capturing the valid variation in responses. In addition, he suggests that only marginal returns are gained from using more than nine response alternatives. Alwin (1992) reports that reliability increases with the number of response categories when looking at scales from 3 to 9 points, but then plateaus beyond 9 categories. In comparison, Matell and Jacoby (1972) conclude that there is no difference in the proportion of scale used for scales between 4 and 19 points, though these authors were looking specifically at self-administered scales.

Thus, the literature addressing response scales recommends a range of scale points. A more precise recommendation cannot be extracted from the available literature primarily because the "optimal" number of points depends on several things such as the construct being measured and the mode of administration. For example, in current practice 8- or 9-point response scales are not frequently used in large scale national surveys. Rather than being a reflection of a preference among survey researchers for shorter scales, this may actually reflect an operational concern. Personal visit interviews are increasingly being replaced by telephone surveys as a more cost-effective way to conduct national surveys and get timely results. However, telephone surveys lack the visual aspect of presenting response scales that is possible with a self-administered questionnaire or a show card in a face-to-face interview. With only auditory input available, it may be difficult to administer a scale much longer than 7 points over the telephone. This belief may be based on information processing theory which holds that about 7 bits or chunks of information is the maximum that can be held in working memory without rehearsal (Miller, 1956; Newell and Simon, 1972).

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 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.

The current unipolar attitude scales in the DHKS telephone instrument are 4-point scales. (A unipolar scale uses only a single dimension, ranging, for example, from "very important" to "not at all important," whereas a bipolar scale uses two opposing dimensions ranging for example, from "very important" to "very unimportant." The former is believed to communicate an <u>absence</u> of importance, whereas the bipolar scale is believed to communicate the <u>presence</u> of non-importance.) For this study, we were asked to consider extending the number of points to 5, to include a middle alternative as well as to capture more of the variance in responses. The decision as to whether to include an explicit middle alternative centers on, for example, the effect of the middle alternative on the univariate distributions and the relationships between variables. In a classic article, Schuman and Presser (1981) conclude that including an explicit middle alternative does increase the proportion of respondents in that category but the people who choose the middle alternative are drawn in equal proportions from the other substantive categories. Bishop (1987) replicates these findings and adds that even referencing a middle alternative in the preface to a question, without offering it as an explicit response choice, increases the proportion of respondents in that (volunteered) category. Bishop (1987) also found an effect on the proportion of extreme responses for two of four items. Schuman and Presser (1981) report only weak support for this polarization effect. Therefore, these findings suggest that including a middle alternative on the response scale depends on the intended use of the data. If the substantive interest is in the marginals or simple relationships between respondent characteristics and the question, then offering a middle alternative

seems reasonable. Further, from the respondent's perspective, including a middle alternative may help to clarify the intended meaning of the scale.

From an interviewer's perspective, increasing the number of scale points may be more important than whether or not the concept of a middle alternative is offered. To interviewers, adding even one additional point means that more information has to be read to the respondent. This may increase administration time, the number of interruptions and break-offs, and other interviewer errors. This perspective is not typically reported in the literature, but it is a real concern in an applied setting. It is from the interviewer's perspective that we consider the 5-point scale to be different from the 4-point scale.

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; Madden, 1960; 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 and Hippler, 1987; Schwarz and Hippler, 1991). 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. Andrews (1984) found partially labeled scales to have higher data quality than fully labeled scales, though he does not comment on this finding. In a more recent study, Krosnick and Berent (1993) report 8 different experiments that looked at both the effects of branching and the extent of verbal labeling on the reliability of a scale. Two of the studies isolated the effects of the extent of verbal labeling on the reliability of bipolar scales. Interestingly, the results of these two studies contradict one another. In one laboratory study, face-to-face interviews were conducted with undergraduates about their political ideology. Two interviews were conducted, about a month apart. Contrary to expectations, Krosnick and Berent (1993) found that reliability was not significantly affected in any direction by the extent of labeling used on a scale (i.e., there was no difference in reliability between partially and fully labeled scales.) In a second study involving adults, face-to-face interviews on the topic of political ideology were followed a few months later by telephone interviews. In this study, fully labeled scales increased reliability, especially for less educated respondents. However, comparisons across these two studies are difficult because in one study there was a change in mode between the first and second administration of the questionnaire. It is possible that the change in mode may have influenced the reliability between the two administrations.

Although the literature in this area is inconsistent (Andrews, 1984; Krosnick and Berent 1993), the popular opinion among survey researchers seems to be that, in comparison to partially labeled scales, fully labeled scales communicate more information and thus yield more reliable, higher quality data. Data quality is defined in numerous ways across studies, ranging from the proportion

of responses falling in the extreme end points to various measures of reliability. 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 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.

From the respondent's perspective, there may be an overload of information when a fully labeled scale is longer than 4 or 5 points, particularly when the interview is administered over the telephone. Overloading respondents with too much information may have adverse effects on data quality because respondents simply cannot process all the information presented to them. Rather, they focus on some portion of the information presented to them (e.g., one end of a scale) and base their answers only on that portion. An example of one such negative effect on data quality would be an increase in response sets. Response sets is the term used when respondents choose the same answer, say strongly agree, for all items or all but one item in a series regardless of what their "true" opinion is. An increase in response sets represents an increase in measurement error.

In addition, there may be an interaction between the extent of verbal labeling and the number of scale points. For respondents and/or interviewers, a fully labeled scale of four points or fewer may not be problematic, whereas a longer scale may.

Branching versus Standard scales

The response tasks 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, or unfolding (Groves, 1979). This technique is most often used with bipolar scales (that is, items that were measured on a scale with two opposing dimensions--"agree vs. disagree" as compared to a scale using a single dimension--"agree vs. not agree.") Branching changes the respondent's task of choosing the direction and the strength of their attitude from a single step, as with a standard scale, to a two-step process. For example, in a standard scale respondents are asked whether they strongly disagree, somewhat disagree, somewhat agree or strongly agree. In comparison, in a branching scale respondents are asked first whether they disagree or agree. Then in a second separate question, respondents are asked whether they strongly

or somewhat disagree/agree. Krosnick and Berent , (1993) report eight studies that looked at the effects of branching and verbal labeling on the reliability of the scale. The previous section described the results of experiments that isolated the effects of verbal labeling and achieved contradictory results. He also isolated the effects of branching in two of the eight studies and found contradictory results. It may be that mode of administration is an important factor in the reliability of the branching and standard scales.

Inconsistent findings have been reported between other studies. Groves and Kahn (1979) compared a single-stage 7-point "satisfaction" question to branching "satisfaction" question in a telephone interview. They found fewer extreme responses and higher inter-item correlations with the two-step scale and concluded it was a better scale. Albaum and Murphy (1988), in comparison, found a higher percentage of extreme responses with the two-stage format. Miller (1984) did a study similar to Groves and Kahn, (1979) comparing a branching and standard version of a satisfaction scale but came to a different conclusion than Groves. He found fewer positive answers (a positive bias is often found with satisfaction scales), less missing data, and higher intercorrelations with the one-step, standard scale and concluded it was the better scale. None of these studies reported reliability of the scale.

However, in the Miller (1984) study, the one-step scale had only the end and middle points of the scale labeled whereas the two-step, branching version labeled all response options. Miller acknowledges that this is not a conceptually clean comparison, but notes that "presenting the seven verbal categories to respondents in a single step is an unlikely field solution, practically speaking." In fact, as noted above, the choice to use a branching scale often results from concern about administering a fully labeled scale that is longer than 4 points. So this technique evolved from practical concerns, although the assumption that administering a fully labeled scale greater than four or five points is problematic has not received a great deal of experimental evaluation. Unlike the other scale characteristics discussed previously, the literature addressing whether data quality (i.e., reliability) is greater with a branching scale as compared to a standard scale is not so clear.

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.

Including a visual aid or a show card in a personal visit interview requires the interviewer to show the visual aid to the respondent but doesn't dramatically affect survey procedures. However, the

use of show cards in a telephone interview may require substantial changes to the survey procedures. For example, incorporating a visual aid makes it necessary to collect address information prior to conducting the telephone interview. If a survey does not already collect address information and send out pre-notification letters, this is a major change in survey procedures and thus survey costs.

The justification for incurring the potential increased cost is that including show cards might improve data quality by decreasing the cognitive burden placed on respondents and simplifying the response task. Since the answer choices would be available to the respondent during the interview, response sets may be diminished and interviewer errors due to inadequate probing or failure to repeat the answer categories may also be diminished. In addition, administration time for the interview may decrease, which could offset some of the increased costs incurred by having to mail out a visual aid. The extent to which including a visual aid in a telephone interview impacts costs or data quality is, of course, also contingent upon whether respondents actually receive and retain the card for use during the interview.

METHODS

Experimental Design

To examine the issues that we discussed in the literature review, we designed an experiment to measure the effects of the number of points on a scale, the extent of verbal labeling, and the type of scale (branching or standard) on data quality as well as operational variables. The starting point for this research was a version of the DHKS questionnaire that incorporated revisions based on cognitive testing of the 1994-1996 DHKS questionnaire (Davis and Wellens, 1995). Selected questions from the revised version of the 1994-1996 questionnaire were incorporated into a test questionnaire designed specifically for this response scale research. (An example of one version of the test questionnaire is included as attachment A.) The data were collected over the telephone, since that is the typical mode of data collection for the DHKS.

We operationalized the variables as follows:

1. There were two treatments for the number of scale points. The short scale was always four points as used in the 1994-1996 DHKS; the long scale had five or six points, depending on the measure;

2. 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);

3. There were two treatments for type of scale that were manipulated only for bipolar items. The treatments were a standard scale and a branching scale that obtained the same information but in two separate questions;

4. There were two treatments for a show card. 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 was a between subjects design for each of the manipulated scale characteristics; that is, across experimental conditions, all respondents received identical questions, only characteristics of the response scales differed by condition. Figure 1 shows the experimental design.

Figure 1: Experimental Design

	Type of Scale			
Number of Points	Standard		Branching[1]	
Short (4 pts.)	Partially Labeled	Fully Labeled	Partially Labeled	Fully Labeled
show card				
no card				
Long (5-6 pts.)				
show card				
no card				

Scales

Three different types of subjective statements included in the DHKS questionnaire served as the targets of the experimental manipulation. These included the extent to which one disagrees/agrees with a statement, the importance of a statement, and the frequency of a behavior. Each of these types of subjective statements had what we refer to as a "target question series" in this report. For example, there was a series of ten disagree-agree statements that we used as the focus of our evaluation for that subjective concept. These ten statements were the target question series for the disagree-agree measure. All of the scale characteristics were manipulated for the disagree/agree

[1]The branching version is depicted as using a 4-point scale because in the second step when respondents are read a scale, they only hear the 4-points of the scale specific to the direction of their response. Moreover, only the 4-point versions of the frequency and importance measures (both unipolar scales) were included in the questionnaire with the branching format of the disagree-agree (bipolar) measure.

scale. Since it is the only bipolar scale statement in the DHKS questionnaire, it was the only one that could be administered using the branching technique. All the other manipulations were implemented with the behavior frequency and importance measures.

As noted above, the number of points in the long version of the scale was either 5 or 6 points, depending on the type of statement. The long versions of the frequency and importance items both contained 5 points which reflected the addition of a mid point. The longer version of the disagree-agree scale was 6 points. Two points were added to this scale because adding one middle point to the bipolar scale (e.g., neither disagree or agree) did not seem to be very different from a "no opinion" category which was already offered.

Figure 2 gives an example of each type of scale manipulation for the disagree/agree measure to help clarify the experimental scale conditions. The disagree/agree measure is given as an example since it is the only measure that includes the branching manipulation. The other two measures (e.g., frequency and importance) are identical in format to the disagree/agree measure for all other conditions.

Figure 2: An example of the experimental scales for the disagree/agree measure

TARGET STATEMENT: I should maintain a healthy weight.	
EXPERIMENTAL SCALES	RESPONSE CATEGORIES
Standard	
partially labeled, short	Choose an answer between 1 and 4 with 1 being strongly disagree, 4 being strongly agree, and 2 and 3 being somewhere in between or tell me "no opinion."
fully labeled, short	Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree or have no opinion about the statement?
partially labeled, long	Choose an answer between 1 and 6 with 1 being strongly disagree, 6 being strongly agree, and 2, 3, 4 and 5 being somewhere in between or tell me "no opinion."
fully labeled, long	Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree or have no opinion about the statement?
Branching	
partially labeled	1) Do you disagree, agree, or have no opinion about that statement? 2) How much do you dis/agree with that statement? Choose an answer between 1 and 4 with 1 being slightly dis/agree, 4 being strongly dis/agree and 2 and 3 being something in between.
fully labeled	1) Do you disagree, agree, or have no opinion about that statement? 2) Do you slightly, somewhat, mostly, or strongly dis/agree with that statement?

One of the objectives of the research was to measure the reliability of the different types of treatment/target statement combinations. 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 respondents were first asked a series of "filler items" for each subjective measure to get them familiar with the experimental scale (e.g., short, partially labeled, standard scale). The second time respondents heard the experimental scale

was when answering the target question series, that is, the questions that are the focus of this research. And the third time they heard the scale was when answering the identical target question series again using the same experimental scale. The first time respondents answered the target series is considered to be "time 1" (T1) and the second time they answered the target series is considered "time 2" (T2) for calculating the reliability of the experimental scale. Between T1 and T2, respondents were asked several items using different subjective measures than those used with the target question series. These items were included to increase the time between the first and second time the respondents answered the same target series. Figure 3 shows the general sequence of the interview for each scale type. For example, there were 7 disagree-agree items in the "building familiarity" series that introduced respondents to the scale format (e.g., short, partially labeled, standard scale) for the disagree-agree questions. There were 10 disagree-agree statements which served as the target questions using the same scale format. This target series was asked twice of respondents during the interview, but there were 12 unrelated items which were not disagree-agree statements in between the two administrations of the target question series.

Figure 3: Sequence of Interview by Subjective Measure

Subjective Measure	**No. of Items in Series**
DISAGREE/AGREE MEASURE:	
'Building familiarity' series	7
Target question series	10
No. of items between T1 and T2	**12**
IMPORTANCE MEASURE:	
'Building familiarity' series	7
Target question series	6
No. of items between T1 and T2	**11**
FREQUENCY MEASURE:	
'Building familiarity' series	3
Target question series	5
No. of items between T1 and T2	**12**

Recruitment and Data Collection Procedures

Implementing this design required that we develop 12 different questionnaires. (Copies of question series 5, a disagree-agree series, are included as Attachment B to exemplify the differences across the 12 questionnaires).

Three hundred and eleven respondents were recruited to participate in this experiment through local organizations, advertisements in local papers, or word of mouth. Of the 311 respondents recruited, 300 completed interviews. Respondents were recruited to fit into one of two educational categories: high, having at least some college, or low, having a high school degree or less. There were 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. Roughly half of the respondents in each experimental condition were from the lower education group. All respondents were paid $10.00 for their participation. (To gain cooperation from organizations that could provide large numbers of respondents, we marketed our research as a fundraising activity for the organization. In some cases, respondents donated the $10.00 to the organization that recruited them.)

Respondents were allocated randomly to the 12 treatment groups. (See figure 4.)

Figure 4: Respondent Allocation to Experimental Conditions

	Type of Scale				
Number of Points	Standard		Branching		
Short (4 pts.)	Partially Labeled	Fully Labeled	Partially Labeled	Fully Labeled	TOTAL
show card	23	22	29	25	99
no card	23	30	24	25	102
Long (5-6 pts.)					
show card	25	26			51
no card	24	24			48
TOTAL	95	102	53	50	300

All respondents were initially contacted on the phone, a screener collecting demographic information was administered, and an appointment was made to conduct the telephone interview. Most appointments were scheduled for between 3 and 5 business days after the demographic

screener was administered. All respondents were mailed a postcard reminding them of their appointment time, as well as a "Census Surveys" magnet. Approximately one half of the respondents (53%) received a postcard which also had the answer categories for each of the three measures printed on the reverse side. Respondents were instructed to keep this postcard for use during the interview and told that the magnet could be used for storing the card on the refrigerator.

Interviews were conducted between mid-October, 1995 and mid-April, 1996. (The federal government furlough was responsible for this lengthy data collection period.) All telephone interviews were tape recorded. All interviews were conducted by research staff at the Center for Survey Methods Research.

RESULTS and RECOMMENDATIONS

Several different types of analysis were conducted to evaluate the different scale characteristics being tested. Three types of analysis were done to directly examine data quality. These were: general descriptions of the distributions; Analysis of Variance (ANOVA) for each item in a target-question series; and reliability analysis. Two other analytic techniques were used to evaluate field related issues: behavior coding analysis and regression models using administration time as a response variable. Although behavior coding analysis is often described as a tool for evaluating field procedures, it also provides an indirect measure of data quality. Each of these analytical techniques are briefly described in the next section.

OVERVIEW OF ANALYTIC TECHNIQUES

Descriptions of Distributions

The data are described in three ways for each question series. First, to indicate the shape of the distribution, the proportion of respondents who used the end points on the scale for each treatment group is given. This is important because a basic assumption of most statistical analysis techniques is that the data are normally distributed. Second, the amount of skew in the distribution is provided for each scale length separately. Lastly, the percent of respondents who demonstrated response set patterns for each question series are provided. A response set pattern is defined as having all, or all but one of the responses in a series being the same value. This is important because response sets make it impossible to assess if the responses given by a respondent represent their true value for the question, or if the respondent has simply chosen to always respond with a "4" or some other value on the scale, regardless of what question is being asked.

Analysis of Variance (ANOVA)

An ANOVA examines whether the variation in responses can be explained by the experimental conditions. For example, an ANOVA model including the extent of verbal labeling as an independent variable addresses whether the variation in responses between those respondents who

received partially labeled scales and those respondents who received fully labeled scales is statistically different. When interpreting the results of an ANOVA it is advantageous to have an expected distribution to determine which experimental condition yields the "best" response distribution. In this survey, as with other surveys, respondents may feel some social pressure during the interview to give the response that is most socially desirable. Specifically, people may feel pressure to answer in a way that is most favorable from a nutritional standpoint regardless of how they really feel about the topic. Given this, the "best" response distributions would be those that do not have a mean value which falls on the extreme positive end of the scale (i.e., strongly agree, very important, always).

For each question series, the following regression model was run for the long and short versions of the questions separately:

Question Response =Extent of Labeling + Presence of Card + Level of Education

The model for the disagree-agree series (Q5) also included a variable for the branching manipulation.

Each question within a series had to be analyzed separately since the items are considered independent of one another. Thus, we had to come up with a way to determine whether results of each of these independent runs were generalizable to the whole series. As our rule of thumb, at least half of the items in a series should have statistically significant results in order to generalize the results to the whole series. Using half of the items as a rule of thumb is actually a fairly loose criterion. However, many of the experimental conditions have less than 30 respondents, which does not provide much power for detecting significant differences. For this reason we set a somewhat loose criterion for generalizability.

Reliability Analysis

One way to evaluate a scale is to calculate reliability for more than one administration of the same question or series of questions, all of which use the same response scale. Reliability calculated as a test-retest correlation is a measure of how much variability there is between the first and second administrations of the same item. As long as the amount of time that passes between the two administrations is reasonably short relative to what is being measured, the observed variability in responses across the two administrations can be attributed to measurement error. Specifically for this research, the variance can be attributed to difficulty respondents had in using the response scale.

For the analysis presented here, reliability was operationalized using test-retest correlations (Pearson's r^2). Correlations were calculated for Time 1 (T1) and Time 2 (T2) responses to each question within experimental treatments (e.g., long vs. short; fully labeled vs. partially labeled). To test for significant differences between treatment groups, a Fischer's transformation was done on each r^2 value and a z-score was calculated. Consistent results within a question series are summarized. For our purposes, consistent is defined according to two criteria. First, for the results

to be consistent there must be significant differences between the comparison groups in the same direction for at least 50% of the items in the series. Second, there must also be at least a medium effect size as defined by Cohen (1988)[2]. If both of these criteria are met, then differences are considered consistent.

Behavior Coding Analysis

Behavior coding involves applying a predetermined, structured coding scheme to each question individually to provide information about how interviewers read questions and interact with the respondent. In this case, interviews were tape recorded and then coded to capture information such as whether questions are read as worded, whether the entire response scale is read, whether respondents give inadequate or uncodable answers, etcetera. (The behavior coding scheme is included as attachment C.) Data from behavior coding can provide information about the expected quality of the data for each question. If interviewers are not reading the question as worded, then all respondents may not be answering the same intended question. If respondents are consistently answering questions inadequately, then there may be a problem with the question wording. In addition to indirectly measuring data quality, we also use the behavior coding data to explain and interpret some of the outcomes of the ANOVA and reliability analysis.

Each question was assigned three different types of behavior codes. One type of code captured respondent behaviors, and the other two captured interviewer behaviors. One interviewer code captured how the interviewer read the question part of the item (e.g., "How important is it that the food be safe to eat?"). The second interviewer code captured how the interviewer read the scale part of the item (e.g., "Not at all important, not too important, somewhat important, very important"). The second type of interviewer codes were added to the behavior coding scheme to provide more information for evaluating the scales. These codes were only used for those questions which had the scale printed as part of the question (e.g., Q5a, Q5e, Q5h in attachment A). Specifically, the codes were assigned according to whether the whole scale was read to the respondent, whether only a portion of the scale was read, or whether none of the scale was read.

The question part of the two interviewer codes evaluated whether the question was read exactly as worded, whether there was a major change so that the meaning of the question might be altered, whether the question was omitted altogether, whether the stem of the question was omitted for those items in a matrix format or whether any other behavior occurred that did not match what was expected. The respondent codes captured whether the respondent gave a codable (e.g., adequate) response or not, whether the response was qualified in any way (e.g., "maybe a 2", "not too important," "if you think about the question in 'that' way"), whether the respondent requested clarification before answering, and any other behavior. In addition, codes were developed to

---

[2] An effect size is a measure used to appraise differences in degrees of correlation. Generally, a medium effect size translates into about 20% more shared variance in the group with the larger reliability value.

distinguish whether people in the branching condition provided their final answer at the first or second step of the question, and whether or not people tended to respond more with numbers or words when given a choice. Unfortunately, coders were not able to consistently use the codes that distinguish between numeric and verbal responses, so that can not be addressed in this analysis.

Cannell and Robison (1971) have suggested that a question is problematic when it is assigned a combination of codes other than "exact reading" and "adequate answer" for 15% or more of the sample. This is one of the criteria used in the work presented here. However, as noted previously some of our codes were specific to the response scale and separate from the question codes. We also include as part of the criteria whether or not the interviewer read all of the scale.

DISAGREE-AGREE SERIES (Q5 and Q9)

This section contains a discussion of the aggregated results of the questions in the disagree-agree question series. The specific questions that comprise the series are the following:

5. Now I am going to read you some statements about diet and nutrition. Please tell me if you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion.

Statement	Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	N/O
a. I should use salt or sodium only in moderation. -- Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	8
b. I should eat at least 5 servings of fruit and vegetables a day.	1	2	3	4	8
c. I should use sugars only in moderation.	1	2	3	4	8
d. I should eat a variety of foods.	1	2	3	4	8
e. I should maintain a healthy weight. -- Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have "no opinion" about that statement?	1	2	3	4	8
f. I should choose a diet low in fat.	1	2	3	4	8
g. I should choose a diet low in cholesterol.	1	2	3	4	8
h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	8
i. I should eat at least two servings of dairy products a day.	1	2	3	4	8
j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	8

Description of Scale

Since this is a bipolar scale (i.e., includes two opposing dimensions), there were two scale formats for this question series -- standard and branching. The standard scales presented all of the response options (that is all points on the scale) at once. The short version of the standard scale is a 4 point scale. The fully labeled version of it was used in the 1995 DHKS data collection. It reads: strongly disagree, somewhat disagree, somewhat agree, strongly agree. The partially labeled version of the standard, 4 point scale went from 1 to 4 with 1 being strongly disagree and 4 being strongly agree.

The long version of the disagree-agree standard scale is a six point scale. In the fully labeled version the scale read: strongly disagree, somewhat disagree, *slightly disagree*, *slightly agree*, somewhat agree, or strongly agree. The partially labeled, standard version went from 1 to 6 with 1 being strongly disagree and 6 being strongly agree.

In the branching version, the response options are presented in two separate steps. The direction of their response (i.e., whether they disagree or agree) is determined first, followed by the magnitude of their response. The original wording of the fully labeled, branching scale was: Do you disagree, agree, or have no opinion? Do you slightly, somewhat, mostly or strongly (dis)agree with that statement? The partially labeled, branching questions had the exact same first question, and the second question asked respondents to choose a number between 1 and 4 with 1 being slightly (dis)agree and 4 being strongly (dis)agree. For the analysis of variance reported later the branching version of the scale was collapsed to be equivalent to the long version of the standard scale. The two lower points ("slightly" and "somewhat") were combined into one point, resulting in a 6- point bipolar scale. We chose to combine these particular points because they contained so few responses combining them would not really alter the overall shape of the distribution.

Response Distribution

The results for the long version of the disagree-agree scale were disappointing. Because distributions for this item were highly skewed (see Table 1, at the back of the report) not many conclusions can be drawn from the data in reference to the scale characteristics. For example, table 1 shows that in the branching condition 9 of the 10 items had a negative skew beyond -1.25 indicating that the data points are all clustered at the extreme end of the scale (i.e., strongly agree). Similarly, table 2 shows the proportion of the distribution in the top two categories by branching, long and short scales. As can be seen, there was very little variation in responses. These questions ask about respondents' knowledge about various aspects of the dietary guidelines; on the surface it appears that our respondents were exceptionally knowledgeable about the guidelines as operationalized in these questions. While the widespread promotion of the generally broad concepts included in the guidelines make this a possibility, alternative explanations are possible. The top category on the scale is also the most socially desirable response and may not reflect a true attitude (social desirability response set). Since each of the 10 questions were written so that the "correct" as well as the most socially desirable answer fell in the top category (i.e., strongly agree), and since

the skew is extreme across all items in the series, this could reflect a tendency for respondents to choose the same answer over and over simply because it takes less cognitive effort to do so (acquiescence response set). In this question series approximately 20% of the respondents answered with the exact same answer to at least 9 of the 10 questions in the series. A distribution like this most likely reflects a problem in the format used for this series of questions.

It is impossible to tell whether the response set pattern reflects social desirability or lack of effort, or whether this is the true distribution in the population. However, because the percentages are so high, we feel that the question series should be re-written to make it easier to determine whether the distribution reflects a response set. The easiest way to do this is to change the direction of approximately half of the items in the series. For example, instead of the current wording "I should eat a variety of foods," the question could be written to say something along the lines of "It is not necessary for me to eat a variety of foods."

Analysis of Variance

Tables 3 and 4 give the significant results of the ANOVA by the long and short versions of the scale, respectively. As a result of the highly skewed distributions, and the high percentage of response sets, the responses were very similar across the experimental conditions (i.e., length of scale, amount of labeling, format, or presence of show card). Overall, none of the experimental manipulations had any consistent effect on responses for this question series. The presence of a card had no consistent effect on responses, though one item in the long version produced significantly higher responses for respondents with a show card as compared to those without a show card ($F=4.37$, $p<0.04$) and one item in the short version of the disagree-agree scale had higher responses with a show card ($F=3.87$, $p<0.05$). One item out of ten, though, is not stable enough to consider this a generalizable finding.

The extent of labeling produced no consistent effect in either the long or short version, though respondents in the fully labeled condition for one item of the short version tended to give higher responses than respondents in the partially labeled condition ($F=4.59$, $p<0.035$). Again, since this result only applies to a single item, we do not consider it representative of the whole disagree-agree question series.

There were no consistent differences in terms of length of scale either. As noted, neither labeling or the presence of a card was differentially affected by length of the scale. There was an effect of education in the longer version that was absent in the shorter version ($F=6.49$, $p<0.011$; $F=16.89$, $p<0.0001$), but again, only for two of the ten items. Thus, this difference between the short and long versions is not considered representative.

Reliability Analysis

Test-retest correlations were calculated across all respondents within each experimental condition, as well as for high and low education groups separately. Table 5 shows the range and average correlation for each experimental condition by education group.

As noted, Fischer's transformations were done on all r^2 values and z-scores were calculated to test significance. The only experimental condition that had an effect on reliability was branching scales. The extent of labeling, the length of the scale, and the presence of a show card did not have any consistent effect on reliability across the ten items. In fact, with the exception of the branching versus standard comparison, results in each condition were such that a few of the ten items were significant in one direction and others were significant in the opposite direction (see table 6). For example, the first row in the "overall" section of the first column shows that across education groups length of scale had a significant effect on reliability for 3 items. However, for two items the short scale was more reliable, and for one item the longer scale was more reliable. This was still the case for the more educated respondents when the analysis was done separately by education groups. Clearly, these results are too inconsistent to be generalized to the whole question series.

The exception was the branching versus standard scale comparison, as previously noted. The standard scales were statistically more reliable in 7 of the 10 items overall, with an average effect size of 0.40 ($z>=1.96$; $p<0.05$, across the 7 items). This was also true for both the high and low education groups separately. Five of the items in the higher educated group and 5 of the items in the less educated group were more reliable in the standard condition as compared to the branching condition ($p<0.05$), with average effect sizes of 0.43 and 0.37 respectively. These effect sizes indicate that roughly 20% more of the variance can be attributed to the format of the scale in the standard condition than in the branching condition. In other words, there is more random error in the branching condition.

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 and Berent, 1993) suggested that branching scales may be more reliable. Since our results for the disagree-agree scale contradicted this assertion, we looked to the behavior coding data for an explanation. The behavior coding data reveal that there was an unacceptably high percentage of interruptions by respondents at the second step of the branching version of a question regardless of whether it was partially or fully labeled, or whether there was a show card present. On average 86.2% of respondents who were supposed to be read the second step (which contained the scale) interrupted before they were read any of the scale. Whereas, when the question was read in one step, on average only 30.6% of respondents interrupted before being read any of the scale for the same items. This could explain why the standard form of the question appears more reliable than the branching version. Respondents simply were not provided the necessary information to respond thoughtfully in the branching version of the question. What is interesting about this finding is that the problem is not specific to any one interviewer, nor is it specific to any one question. No particular interviewer was consistently being interrupted in the branching version, suggesting that

this is more of a problem with the question format than it is a problem with particular interviewer behavior which lead to interruptions.

As a result, we repeated the reliability analysis, this time only examining the standard version of the scale (see table 7). When the branching items are dropped from the analysis, the extent of labeling on the scale does make a significant difference in reliability, but only for the higher educated group. The shaded row of table 7 shows that reliability was significantly greater in the fully labeled condition than the partially labeled condition for 7 of the 10 disagree-agree items for the more educated group. The effect size of 0.73 for this group indicates that roughly over 1/3 more variance is accounted for in the fully labeled condition relative to the partially labeled condition. Interestingly, there is nothing similarly conclusive with the less educated group. The extent of labeling did not have an impact on reliability for that group for this question series. Though it is not clear why there would be a difference by education groups, the findings do suggest using a fully labeled scale with the disagree-agree questions, since the higher educated would benefit from the verbal labels without having any negative impact on the less educated. Krosnick and Berent (1993) draw similar conclusions in their work. They found that the fully labeled scales improved reliability in one of two political opinion studies. They concluded that since full verbal labels did not have any negative effect in the other study, fully labeled scales should be used for assessing political opinions.

There is one additional finding from the behavior coding data that suggests that the branching form of the disagree-agree series may not yield the best data. After the first item in the series, on average, 43.6% of the respondents provided the information to answer the second step question at the first step (see table 8). Presumably respondents did this because they were trained about what to expect in the first item and they felt they did not need two stages to answer, or that they did not want to take the time to answer two questions rather than one. The result is that respondents in the branching condition heard all of the scale options less often than the respondents in the standard condition. Thus, they have to remember all the response options from the first (and perhaps only) time they heard them when asked the first item in the series. More likely, however, is that they remember some subset of the response options and use that subset as their repertoire of possible answers, potentially leading to biased results. This, in addition to the problems noted above with the branching format, leads us to suggest that a standard format be used with the disagree-agree items.

As a last comment with regard to the disagree-agree series (as well as the other two series), there is a potential bias with the overall format we used for testing. Krosnick (1992) has shown that attitude questions tend to have a "recency bias" in telephone interviews. A "recency bias" refers to the effect of respondents having a tendency to answer with the last response choice they hear. In the fully labeled format we used for testing, the last response choice is also the most socially desirable response choice. Since we feel there is potential for respondents to be influenced by social desirability, having this also as the last response choice is problematic. We suggest that the scale be reversed to go from strongly agree to strongly disagree so that the potential for a recency bias is not added to the potential for a social desirability bias. Moving from "strongly agree" to "strongly disagree" is the direction used in the 1994-1996 questionnaire.

In summary, the results from the z-score tests of significance and the behavior coding imply that a standard one-stage scale should be used. Although the extent of labeling on the scale did not significantly impact the responses given by the less educated respondents, there was an effect for the higher educated respondents. Higher educated respondents did better with fully labeled scales, suggesting that fully labeled scales should be used with disagree-agree items.

There was no statistical difference between the 4-point and the six-point scale in terms of reliability or mean scores. There was a greater number of items with a notable negative skew in the longer scale compared to the shorter scale (7 versus 5 respectively). However, the average percent of responses across the 10 items falling in the top category of the scale (i.e., strongly agree) was greater in the shorter version of the scale. Thus, there is no clear difference in length of scale for these data quality measures.

Finally, the presence of a show card was not related to any of the measures of data quality for this question series.

IMPORTANCE SERIES (Q16 and Q19)

This section contains a discussion of the aggregated results of the questions in the importance question series. The specific questions that comprise the series are the following:

16a. **Now think about buying food. When you buy food, how important is it that the food be safe to eat -- not at all important, not too important, somewhat important, or very important?**

[1] Not at all important
[2] Not too important
[3] Somewhat important
[4] Very important

16b. **When you buy food, how important is (FACTOR)?**

IF NEEDED:	How important is (FACTOR) - not at all important, not too important, somewhat important, or very important when you buy food?

Factor	Not at all Important	Not too Important	Somewhat Important	Very Important	N/O
1. nutrition? -- not at all important, not too important, somewhat important, or very important.	1	2	3	4	8
2. price?	1	2	3	4	8
3. how well the food keeps?	1	2	3	4	8
4. how easy the food is to prepare?	1	2	3	4	8
5. taste?	1	2	3	4	8

Description of Scale

Since the importance scale is a unipolar dimension, (i.e., contains only a single dimension) all versions of the scale were in the standard format (i.e., both direction and magnitude are assessed in one question). The short version of the importance scale is a 4-point scale. The fully labeled version reads: not at all important, not too important, somewhat important, very important. (This is the same number of points and the same verbal labels as used in the 1995 DHKS instrument, but the order was reversed for the purpose of this test.) The partially labeled version of the standard, 4-point scale went from 1 to 4 with 1 being not at all important and 4 being very important.

The long version of the scale is a five-point scale. The fully labeled version of the scale read: not at all important, not too important, somewhat important, *important*, very important. The partially labeled, standard version went from 1 to 5 with 1 being not at all important and 5 being very important.

Response Distribution

As with the disagree-agree scale, the responses for the importance question series were not normally distributed. Half of the items in the series (i.e., 3 of 6) were negatively skewed for the short version and two-thirds of the items in the series (i.e., 4 of 6) for the long version had a non-trivial negative skew (see table 9). Because of this, there was little variability in responses, making it difficult to detect differences between experimental conditions.

As can be seen in table 10, over half of the responses, on average, fell in the highest response category, "very important." Over 80% fell in the top two categories for both the long and the short scale. Moreover, about 38% of the respondents gave the same answer to all or all but one of the items in the series, strongly indicative of a problem with response sets. As indicated by the negative skew, the most common response set pattern is to choose the points on the end of the scale indicating great importance. However, in comparison to the disagree-agree question series, it is not immediately clear that this pattern is driven by social desirability. While it may be socially desirable to say that "nutrition" or "taste" is very important when buying food, it does not seem that respondents would perceive it more socially acceptable to say "price" or "how well the food keeps" is very important to them when buying food. On the other hand, this could be a difference between the two education groups. With education being highly correlated to income, it could be that price is "very important" to the less educated. Thus, what seems like a response set may actually be reflecting true values for some respondents. To test this pattern of responses, we did a chi-square test of independence between education level and the presence of response sets. Significantly more of the less educated respondents answered all of the items or all but one of the items as "very important" as compared to the higher educated respondents ($X^2=17.94$, $p<0.001$). Since this difference between the education groups did not occur in the disagree-agree series (nor in the

frequency series), we believe that the "very important" responses reflect the true distribution for this question series.

Analysis of Variance

As with the disagree-agree scale, there was not enough variability in the data to detect many significant differences across conditions. The large negative skew, as noted above, is an indication of low variability. As a result, only one of the experimental conditions yielded significant differences, but with marginal consistency. Tables 11 and 12, respectively, contain the significant results of the ANOVA models for the long and short versions of the importance scale.

The show card had no effect in either the long or short versions of the scale. Extent of labeling did not have a consistent effect in the long version of the scale but it did have a marginally consistent effect in the short version. For one of the six items in the long version of the scale, respondents in the partially labeled condition had a higher mean response than those in the fully labeled condition ($F=5.91$, $p<0.017$). But since this is only for one item, we do not consider it a stable finding which can be generalized to the whole question series. For two items in the shorter version of the scale, respondents in the fully labeled condition had a higher mean response ($F=16.54$, $p<0.0001$; $F=16.49$, $p<0.0001$). It is difficult to definitively say whether this is consistent or not. Significance only occurs in two of six items, but both are highly significant and the direction of significance is the same in both cases. At first, this suggested to us that fully labeled scales may be more prone to social desirability effects than partially labeled scales, leading to the conclusion that partially labeled scales may have better data quality. However, this result may actually represent a problem in our design rather than a true difference between the scales.

In the partially labeled condition, the scale was read to respondents as: "choose a number between 1 and 4, with 1 being 'not at all important,' 4 being 'very important' and 2 and 3 being something in between." 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 " very important." 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. In addition, as noted previously, the literature does seem to suggest (though not definitively) that fully labeled scales yield higher quality data. Based on Krosnick's work and the other response scale literature, we believe this finding may be an artifact of the confound in the design rather than a real difference in the extent of labeling.

A difference by length of scale seems plausible given that the recency bias only appeared to be a problem in the shorter version of the scale. The shorter version may have more potential for response bias problems, suggesting that the longer scale may yield higher data quality. No other findings suggested a difference between the long and short scales. Education level produced significant differences in two of six items, with the less educated responding higher than the higher educated but it was consistent for both the long and short version (see tables 11 and 12). In addition,

the two items were the same two items noted above ("price" and "how well food keeps"). We interpret this as a real difference between education groups for these particular questions in the series, and do not think it reflects anything about the scale, especially since it occurred in both the long and short versions.

Reliability Analysis

There were no differences in reliability for any of the scale conditions looking at both education groups together. However, when the analyses are done separately for the two education groups, a different picture emerges.

While length of the scale had no consistent effect, the presence of a show card and the extent of labeling did. However, this was only for the less educated respondents. The presence of a show card improves reliability in four of the six items for the less educated group with an average effect size of 0.45 (see table 13). The presence of a show card has no effect on reliability for the higher educated group. It seems that those with less education benefitted from having two cues, one visual and one aural, when answering the questions. In addition, there were significantly fewer response sets in the less educated group when they had a show card to use during the interview (X^2=5.46, $p<0.019$). The presence of a show card did not diminish response sets for the higher educated respondents. Since this survey oversamples low income and income is highly positively correlated with education, a large portion of the sample is likely to have a high school education or less. Under this assumption, including a show card may improve data quality when measuring importance.

In regard to the extent of labeling, the less educated respondents in the partially labeled condition had more reliable responses for 3 of the 6 items with an average effect size of 0.40 (see table 13). This is surprising since it is in the opposite direction of what we found with the disagree-agree question series and counter to what we expected based on the literature. To explain it we looked at the response set and behavior coding data. There is no difference in the percent of response sets by extent of labeling for either education group. Thus, response set patterns cannot explain this finding.

The behavior coding data does not explain this finding either. It seemed plausible that the verbal response categories were longer and more tiresome for interviewers to read. As a consequence, we thought that perhaps interviewers in the fully labeled condition were not reading all the response categories as often. Whereas, this might not affect respondents with more education, it could be that respondents with less education may need to hear the whole scale to respond. If that is the case, then partially labeled scales would appear to be more reliable for the less educated. However, the behavior coding data did not support this hypothesis. As table 14 shows, interviewers actually read all the response categories in the fully labeled condition more frequently than the partially labeled condition, for both education groups.

An alternate hypothesis is that the verbal labels were difficult for less educated respondents to remember and respond with consistently. Though interviewers weren't affected by the verbal labels, respondents may have been. Thus, respondents in the less educated group may have found the

partially labeled scale easier to use. Unfortunately, we can only speculate that this is the reason, since no data are available to investigate it any further.

Overall, none of the scale characteristics affected reliability across education groups. However, when looking at the two education groups separately two consistent and significant patterns emerge. First, lower educated respondents have more reliable answers and a lower frequency of response sets when they have a show card available to use during the interview. Thus, in a survey oversampling low income which is highly correlated with education, a show card may improve data quality when measuring importance.

Second, less educated respondents had more reliable responses in the partially labeled condition than the fully labeled condition. This is a surprising finding in that it contradicts what was found with the disagree-agree scale, and what the literature suggests. Since it was not replicated with the other measures, and since it is not well supported in the literature, at this point we can only recommend further investigation rather than a change in scale format for this series of items.

FREQUENCY SERIES (Q17 and Q20)

This section contains a discussion of the aggregated results of the questions in the frequency question series. The specific questions that comprise the series are following:

17. Now think about food labels. How often, if at all, do you use (SECTION) to help you decide whether or not to purchase a product? -- Never, rarely, sometimes or often?

IF NEEDED:	When deciding whether to purchase a product do you use (SECTION) never, rarely, sometimes, or often?

SECTION	Never	Rarely	Sometimes	Often (Always)	NEVER SEEN	N/O
a. The nutrition panel that tells the amount of calories, protein, fat and such in a serving of the food.	1	2	3	4	5	8
b. The short phrases on the label like 'low-fat' or 'light.'	1	2	3	4	5	8
c. The list of ingredients.	1	2	3	4	5	8
d. The information about the number of servings in a package.	1	2	3	4	5	8
e. The information about the size of a serving.	1	2	3	4	5	8

Description of Scale

The frequency scale is also a unipolar scale, (i.e., contains only a single dimension) so like the importance scale, all versions were in the standard format (i.e., both direction and magnitude were assessed in a single question). The short version of the frequency scale is a 4-point scale. The fully labeled version uses the same labels as used in the 1995 DHKS instrument, but in reverse order. The partially labeled version of the standard, 4-point scale used in this study went from 1 to 4 with 1 being never and 4 being often.

The long version of the scale is a five-point scale. In the fully labeled version the scale read: never, rarely, sometimes, often, *always*. The partially labeled version went from 1 to 5 with 1 being never and 5 being always. Notice, different than the longer versions of the disagree-agree and the importance scales, the additional point was added at the end of this scale, whereas in the other two scales the additional points were added somewhere in the middle. Adding the additional point on the end for this scale required a difference in the label on the end points between the two length of scale conditions, allowing for a possible confound in results . However "always" was a silent option (i.e., not read aloud to respondents, but taken as an acceptable response if volunteered) which presumably would minimize this potential confound..

Response Distribution

The distributions for this question series closer approximate a normal distribution compared to the other question series, since none of the items are severely skewed (see table 15). Similarly, there is not as high a proportion of responses falling in the top and top two categories (see table 16), as compared to the other two measures. Approximately 30% of respondents answered in a response set pattern. However, there were no significant differences in the percentage of response sets by any experimental condition. This suggests that the scale options are not what drives the response set. It more likely reflects that the questions are perceived by respondents to be measuring the same or very similar concepts.

Analysis of Variance

For this measure, the presence of a show card had no effect on the mean scores. The extent of verbal labeling did, but only for the short version of the scale suggesting an effect of scale length. In addition, there seemed to be a difference in the way the education groups used the short and long scales.

The responses in the fully and partially labeled conditions for four of the five items in the short version of the scale were significantly different (see table 17). Respondents in the partially labeled condition tended to give lower responses on the 4-point scale than respondents in the fully labeled condition. However, this also could be reflective of the recency bias previously noted in the results of the importance scale.

What is interesting about this finding, however, is that the confound in the design only seems to have an impact in the short version of the scale. The longer version of the scale is unaffected by the "recency bias" which was observed in the shorter version. This would suggest that there is an effect of scale length, and in favor of the longer scale. However, there was also an effect of scale length by education group, but in this case, in favor of the shorter scale.

In the long version, the average score was significantly higher for the less educated respondents relative to the higher educated respondents for four of the five items in the series[3] (see table 18). The tendency for the less educated respondents to give more responses that indicate a greater frequency of behavior is likely a reflection of social desirability. It does seem more socially desirable to say that you "always" pay attention to the short phrases on the label such as "lowfat," or that you "always" use the nutrition panel in your purchasing decisions. Narayan and Krosnick (1995) have noted that less educated respondents have a greater tendency towards response effects such as social desirability than do respondents with a more education. Thus, there is support in the literature for this finding.

In terms of length of scale there are two contradictory findings for this measure. A "recency bias" is observed in the shorter version of the scale, but not in the longer. On the other hand, a response effect due to education was observed in the longer version of the scale, but not the shorter. This inconsistency makes it necessary to consider the results of the other two measures in order to make a suggestion for scale length.

As with both the disagree-agree and the importance scales, the presence of a show card did not have an effect on mean scores. Though the extent of labeling did make a difference for the short version of the scale, we believe it reflects a confound in the design rather than a true difference. Thus, the length of the scale is the only true effect.

Reliability Analysis

The reliability of responses was not affected by any of the experimental conditions. The length of the scale had no effect, the presence of a show card had no effect, nor did the extent of labeling. Table 19 shows these results. This marked stability of responses suggests that the type of scale selected for this survey should be based on what works best for the other measures (e.g., disagree-agree, importance) since seemingly none of the scale characteristics should affect the reliability of responses for this measure.

---

[3] In the short version, there was also a significant difference between the education groups. However, the difference was only for two items and was in the opposite direction of that observed with the longer scale. Since the difference in the short version is only for 2 items, it is not clearly generalizable to the series.

ADMINISTRATION TIME

Administration time is an issue in that it is one of the biggest determinants of cost in a telephone survey. The longer the administration time, the higher the cost. Thus, we have included this variable in our analysis, since data quality can never be fully separated from costs.

One would expect that a fully labeled scale, if administered correctly, might take longer to administer than a partially labeled scale. On the other hand, it might also be reasonable to expect a shorter administration time for interviews in which respondents have a show card. Though our previous analysis demonstrated that the presence of a show card only affected mean scores for one measure and did not affect reliability for any of the measures, a show card may decrease administration time since interviewers do not need to repeat the scale as often. As a result, there might be a reduction in the average cost per interview. Note that administration time refers to the amount of time it took to administer the whole interview instrument. We do not have data about administration time for any particular question series so no conclusions can be drawn about any one series (e.g., importance) independent of the others.

An analysis of variance (ANOVA) was done using administration time as the dependent variable. The independent variables were the extent of verbal labeling (e.g., partial or full), the presence of a show card (e.g., with or without), and type of scale (e.g., branching, long or short).

The overall ANOVA was significant ($F=6.46$, $p<0.0001$). Table 20 shows the means, the F value and the significance level for each independent variable in the model. Education did not produce a significant effect, so it was excluded from the final model. The extent of verbal labeling and the type of scale were both significant factors. Results show that partially labeled scales took longer to administer than did an interview consisting of fully labeled scales. This could reflect the fact that interviewers tended to read less of the fully labeled scale, or that they read it less often than directed. We looked at the behavior coding data, and this hypothesis was not supported. As can be seen from table 21, in fact, the partially labeled scales were read less often than fully labeled scales -- the opposite of what we had expected. As it turns out, even though there are fewer labels to read in the partially labeled condition, more words are used to describe the scale to respondents in comparison to the fully labeled condition. Given this, it seems logical that it took longer to administer the instrument when it contained partially labeled scales. Krosnick and Berent (1991) report this same finding in regards to the administration time of partially versus fully labeled scales.

Not surprisingly, interviews that included the branching or unfolding technique for bipolar items took longer to administer. There was a significant difference between branching and the shorter scale, but not between branching and the longer scale. It is difficult to interpret this comparison given that only one of the scales in the interview was a branching scale (disagree/agree scale). However, we can presume that using a branching scale for a larger proportion of the interview would increase the total administration time even more. This, coupled with the finding that reliability was lower with branching scales than standard scales, clearly suggests that standard scales may produce better data at a lower cost.

Contrary to what we expected, having a show card available during the interview did not affect the speed of the interview. As noted above, having a show card did not affect the mean scores of any of the measures and it only affected the reliability of the importance measure for the less educated respondents. This would suggest that there is not a noticeable benefit of using show cards for surveys using frequency scales of four or five points. However, when the interaction of card and type of scale is considered, this conclusion is altered. The show card did not seem to have any effect on administration time for the version of the questionnaire using only 4-point scales. On the other hand, the version of the questionnaire using longer scales, and the version of the questionnaire using branching scales for the bipolar items were shorter if a show card was available. The show card seems to mitigate the lengthening effects of some procedures. However, the interaction of show card and length of scale (i.e., short, long, branching) had no impact on data quality. Given this, cost issues should drive the decision whether or not to include a show card.

SUMMARY OF ANALYSIS and RECOMMENDATIONS

Evaluation of the different scale characteristics was done separately for each measure (i.e., disagree-agree, importance, and frequency). These are very different concepts and could be affected differently by the experimental conditions. In fact, this is what occurred. Analytic results across the three measures were different. Moreover, the distributions for the three measures were different, also suggesting they be analyzed separately. However, it is important that a single scale format be used consistently throughout the survey instrument, regardless of what combination of the experimental conditions is selected. A change in format (e.g., from fully to partially labeled), between the disagree-agree and the importance scales would greatly complicate the response task. Respondents would have to learn and understand the new format before responding. And it is likely that the first scale format respondents encounter in the survey would bias the way the second scale is interpreted and used. Thus, there would be an order effect, of sorts, within the instrument. For these reasons we developed a recommendation for a single scale type to be used across the three measures.

This section summarizes our findings according to the various experimental conditions. Recommendations are made for each condition separately and then a single scale format is suggested for the instrument as a whole. However, we strongly recommend a field test of this format before incorporating it into the final instrument.

Branching

The results for this scale characteristic were quite clear. 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. In addition, respondents interrupted interviewers before they read the scale on average over 40% more often in the two step (branching) version than they did in the one step version of the question. Approximately 44% of respondents answered the two-step version at the first step of the question anyway. Lastly, the version of the instrument containing the branching format took significantly longer to administer.

Recommendation:

Given these results we suggest maintaining a standard, one-step format in the instrument.

Extent of Labeling

The extent of labeling had no true effect on the mean scores for any of the 3 measures (i.e., disagree-agree, importance or frequency), though as noted previously there was a confound in the design that produced what we believe to be an artificial effect of labeling in the mean scores for both the importance and frequency measures.

In terms of reliability, respondents with a higher education had higher reliability in the fully labeled condition of the disagree-agree measure. Since reliability for the less educated respondents was not negatively affected by the verbal labels, this would suggest a fully labeled scale. However, in contradiction to this, reliability for the importance measure was better in the partially labeled condition for less educated respondents. This is surprising, especially since interviewers were less likely to read the whole scale to respondents in the partially labeled condition for this measure.

Recommendation:

In general, our findings were not conclusive in regards to the extent of labeling. The literature, though also not conclusive, does more often find higher data quality (typically in terms of reliability) for fully labeled scales than partially labeled scales. Moreover, the most recent work in this area by Krosnick and Berent (1993) suggests fully labeled scales are more reliable. Based on this, and the fact that partially labeled scales took significantly longer to administer than fully labeled scales, we suggest using the fully labeled scales. However, we make this suggestion with the caveat that field testing be done.

Length of Scale

For the most part there were not large effects on data quality by length of scale. In fact, reliability was never affected by length of scale for any measure. This is not too surprising given the little difference between the shorter and longer versions of the scale. However, there were a couple of interesting results in the ANOVA as well as with the response distributions. For both the importance and the frequency scales, we observed a recency bias in the ANOVA for the shorter version of the scale only. In addition, the proportion of respondents in the top categories for each scale (i.e., strongly agree, very important, always) was greater in the shorter version of the scale than the longer version. This suggests that people's opinions may be further differentiated by offering at least one additional category beyond the four-point scale. In other words, a greater amount of the variance will be captured with the longer version of the scale.

On the other hand, there was an effect of education in the longer version of the frequency scale that was not apparent in the shorter version. Less educated respondents in the longer version tended

to reply that they engaged in the health behaviors referred to in that question series with greater frequency than did the higher educated respondents. We interpreted this as a social desirability bias.

Recommendation:

The results are not clearly decisive. However, we feel that the two examples of a recency bias in the short versions of the scale coupled with the larger proportion of responses at the high end of the short scale suggest that the longer scale may provide better data quality. The longer scale will allow respondents to differentiate their positive responses (e.g., "often" versus "always"; "important" versus "very important") to a somewhat greater extent.

Presence of Show Card

With one exception the presence of a show card had no affect on data quality either in terms of reliability or mean scores. There was no real difference in interviewer behaviors with a show card. Administration time was only affected by a show card in the longer and branching versions of the scale. However, we do not have any cost data available to evaluate the extent of the mitigating effect of the card on administration time when using longer scales. (Since we are not recommending using a branching format this finding relative to the branching scale is of no consequence.)

Recommendation:

The show card does not seem to improve data quality, or ease interviewers' tasks. It does seem to shorten administration time when using a longer scale but whether or not the decrease in administration time will save enough money to offset the costs of printing and monitoring the show card in the field is difficult to determine. In the absence of this information and the absence of any obvious data quality benefits we feel that the money allocated to producing and monitoring the card may be better spent on improving other survey procedures, such as interviewer monitoring, increasing the number of contact attempts, etcetera.

Recommended Scale Type

Based on the above findings, a longer (i.e. five- or six-points depending on the measure), fully labeled, standard scale should provide the highest data quality for the items in this survey. We do suggest a larger field test under normal survey procedures be conducted however, since we had some contradictory results.

General Recommendations

For all 3 measures (i.e., disagree-agree, importance, and frequency), response sets were an issue. In the case of the disagree-agree and the frequency scales, we concluded that the response sets were most likely driven by social desirability. On the other hand, it seemed that the response set may reflect a true response pattern for at least some of the items in the importance scale. To make it

easier to detect true answers from response sets, we recommend reversing the direction of some of the items in each series, so that the most socially desirable response is not always last. We realize that this is not easily done for some of the items, but making this change would allow better measurement of the presence of response sets versus true responses. Of course, new wording to reverse the direction of the questions should be tested before going into the field.

In addition, we feel that the order of the response options should be the same as used in the 1994-1995 survey for each measure. For testing, we had reversed the order of the responses to move from negative (i.e., strongly disagree, not at all important, never) to positive (i.e., strongly agree, very important, always). We reversed the order as a result of discussions with ARS staff which convinced us that it would be more logical for the low numbers in the partially labeled condition (i.e. "1") to correspond with the negative responses (i.e., strongly disagree, not all important, never). However, this order makes it so that the last response option heard is also the most socially desirable. As noted previously, Krosnick (1992) has reported that respondents in telephone interviews have a tendency to respond with the last response option heard -- what he refers to as a "recency bias." Given this, we recommend that the order of the response categories go from positive to negative, as they were in the 1994-1996 instrument, regardless of the scale type selected for use in future instruments.

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).

Bishop, G. F. (1987) "Experiments with the Middle Response Alternative in Survey Questions" Public Opinion Quarterly, 51, 220-232.

Cannell, C.F. and Robison, S. (1971) "Analysis of Individual Questions" in J.B. Lansing et al. (Eds.) Working Papers on Survey Research in Poverty Areas, Chapter 11. Ann Arbor, MI, Survey Research Center, The University of Michigan.

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.

Davis, W. and Wellens, T. (1995) "Evaluation of the Diet and Health Knowledge Survey: Phase 1, Cognitive Interviews" Report prepared for U.S.D.A.-A.R.S.

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.), New York: Springer-Verlag.

Madden, J. M. (1960) "A Comparison of Three Methods of Rating-Scale Construction," WADD Technical Note.

Matell, M. S., and Jacoby, J. (1972) "Is There An Optimal Number of Alternatives For Likert-Scale Items? Effects of Testing Time and Scale Properties" Journal of Applied Psychology, 56(6), 506-509.

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.

Narayan, S. and Krosnick, J. (1995) "Education Moderates Some Response Effects in Attitude Measurement" Unpublished manuscript.

Newell, A. and Simon, H. A. (1972) Human Problem Solving, Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Peters, D.L. and McCormick, E.J. (1966) "Comparative Reliability of Numerically Anchored versus Job Task Anchored Rating Scales." Journal of Allplied 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).

Schuman, H. and Presser, S. (1981) Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording and Context, Academic Press.

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.), New York: 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 Meeting of the American Political Science Association, Washington, D.C.

TABLES

Table 1: Skewness for Branching, Short and Long Versions of the Ten Disagree-Agree Items by Type of Scale

Type of Scale	Min. Skew	Max. Skew	No. < -1.25
Branching	-0.174	-4.844	9
Short	-0.033	-3.276	5
Long	-0.161	-2.521	7

Table 2: Average Percent of Distribution Falling in the Top and Top Two Categories for all Disagree-Agree Questions by Type of Scale

Type of Scale	Avg. % "Strongly Agree"	Avg. % "Strongly or Somewhat Agree"
Branching	66.15	81.13
Short	67.76	85.51
Long	63.94	78.38

Table 3: Statistically Significant Results of the Analysis of Variance (ANOVA) for Long Version of the Disagree-Agree Items

(Scale: 1=Strongly Disagree, 6=Strongly Agree)

Question & Factor	**Mean Score (n)**	**F value**	**Pr >F**
salt and sodium in moderation...		6.49	.011
low education	4.90(87)		
high education	5.47(102)		
maintain a healthy weight...		4.37	.04
with card	5.88(104)		
without card	5.69(96)		
6 servings of breads...		16.89	.0001
low education	3.25(87)		
high education	4.40(99)		

Table 4: Statistically Significant Results of the Analysis of Variance (ANOVA) for Short Version of the Disagree-Agree Items

(Scale: 1=Strongly Disagree, 4=Strongly Agree)

Question & Factor	Mean Score (n)	F value	Pr >F
salt and sodium in moderation...		3.87	.05
with card	3.79(42)		
without card	3.47(51)		
5 servings of fruit...		4.59	.035
fully labeled	3.79(52)		
partially labeled	3.54(46)		

Table 5: Minimum, Maximum & Average r^2 values for Disagree-Agree Question Series by Experimental Condition per Education Group

Condition	Minimum		Maximum		Average	
	High Ed	Low Ed	High Ed	Low Ed	High Ed	Low Ed
Length of Scale						
Long (6 pt.)	.29	.44	.88	.74	.61	.63
Short (4 pt.)	.29	.43	.82	.87	.66	.67
Extent of Labeling						
Partially	.70	.69	.93	.91	.80	.77
Fully	.78	.60	.96	.91	.88	.75
Presence of Card						
With Card	.70	.62	.96	.94	.83	.78
No Card	.74	.58	.97	.85	.84	.75
Type of Scale						
Branching	.38	.40	.90	.83	.67*	.64*
Standard	.74	.63	.96	.88	.84*	.78*

* Denotes statistical significance (after Fischer's transformation): $z>=1.96$, $p<=.05$

Table 6:Test of Significant Difference in Reliability for the Disagree-Agree Question Series by Education (10 Items Total)

Education Level	# Signif. Items p< .05	Condition	Average Effect Size[4]
Low	3 0	Short Long	.25
Low	2 1	Card No card	.26
Low	3 2	Partial Fully	.37
Low	5 0	Standard* Branch	.37
High	2 1	Short Long	.26
High	2 0	No card Card	.29
High	3 0	Fully Partial	.38
High	5 0	Standard* Branch	.43
Overall	2 1	Short Long	.19
Overall	2 2	Card No Card	.24
Overall	5 2	Fully* Partial	.34
Overall	7 0	Standard* Branch	.40

* Denotes statistical significance (z>=1.96, p<=.05 and an average effect size >=.30) across items in the question series.

[4] An effect size is a measure used to appraise differences in degrees of correlation. Cohen (1988) sets forth the following guidelines for interpretation of effect sizes: Generally, a small effect size (E.S.=.10) indicates that the group with the larger r^2 value has between 5%-8% greater shared variance between the first and second administrations of the question than does the other group. Above, the larger r^2 value corresponds to the group with the significant z-score test of differences. Similarly, a medium effect size (E.S. =.30) indicates that there is about 20% more shared variance for the group with the larger r^2 value. A large effect size (E.S.=.50) indicates that there is about 1/3 more shared variance for the group with the larger r^2 value relative to the other group.

Table 7: Test of Significant Difference in Reliability for the Disagree-Agree Question Series by Education, Controlling for Branch Form

(10 Items Total)

Education Level	#Signif Items p<0.05	Condition	Average Effect Size[5]
Low	2 1	Short Long	.29
Low	3 1	Card No Card	.39
Low	2 1	Partial Fully	.27
High	1 2	Short Long	.35
High	3 0	No Card Card	.38
High	7 0	Fully* Partial	.73
Overall	1 2	Short Long	.24
Overall	3 2	Card No Card	.23
Overall	4 0	Fully Partial	.24

* Denotes statistical significance (z>=1.96, p<=0.05 and an average effect size >=0.30) across items in the question series.

[5] An effect size is a measure used to appraise differences in degrees of correlation. Cohen (1988) sets forth the following guidelines for interpretation of effect sizes: Generally, a small effect size (E.S.=.10) indicates that the group with the larger r^2 value has between 5%-8% greater shared variance between the first and second administrations of the question than does the other group. Above, the larger r^2 value corresponds to the group with the significant z-score test of differences. Similarly, a medium effect size (E.S. =.30) indicates that there is about 20% more shared variance for the group with the larger r^2 value. A large effect size (E.S.=.50) indicates that there is about 1/3 more shared variance for the group with the larger r^2 value relative to the other group.

Table 8 : Percent of Answers Given In Response to Prior Step
Within A Branching Question, and
Percent of Times Interviewers Omitted the Second Step Question[6]

Question No.	Intvwr. Omitted Second Ques.	Response given in Prior Step
5a1+2	9.4%	8.6%
5b1+2	36.0	36.0
5c1+2	**28.7**	**30.0**
5d1+2	51.2	51.2
5e1+2	42.2	43.4
5f1+2	58.5	59.8
5g1+2	**65.5**	**65.5**
5h1+2	34.9	31.0
5i1+2	36.5	33.8
5j1+2	41.3	41.9
Average excluding 5a[7]	43.9	43.6

[6]After first question in series (i.e., Q5a): between 30% and 65% of respondents answered 2nd step at first step across all remaining items; between 29% to 65% of interviewers omitted 2nd step question across all remaining items.

[7] Since Q5a was the first question in the series, interviewers were more likely to read the second step question, and respondents were more likely to listen to the second step question relative to the later items in the series. After the first item, however, the question format was established for respondents so they felt more able to answer with the second step information at the first step. This difference between the first question and later questions is the motivation for excluding Q5a from the average.

Table 9: Skewness for Short and Long Versions of the Six Importance Items, by Length of Scale

Length of Scale	Min. Skew	Max. Skew	No. < -1.25
Short	-0.758	-4.928	3
Long	-0.424	-2.890	4

Table 10: Average Percent of Response distribution Falling in the Top and Top Two Categories across all Importance Questions by Length of Scale[8]

Length of Scale	Avg. % "Very Important"	Avg. % "Very Important" or "Important"
Short	66.33	90.7
Long	59.05	82.3

[8] In addition, three of the six items in the long version and two of the six items in the short version had no responses in the first response category "not at all important."

Table 11: Statistically Significant Results of the Analysis of Variance (ANOVA) for Long Version of the Importance Items

(Scale: 1=Not at all important, 5=Very important)

Question & Factor	Mean Score (n)	F value	Pr >F
food is safe to eat...		5.91	.017
fully labeled	4.84(50)		
partially labeled	4.98(49)		
price...		17.23	.0001
high education	3.92(51)		
low education	4.66(47)		
how well the food keeps...		3.83	.053
high education	3.80(51)		
low education	4.26(47)		

Table 12: Statistically Significant Results of the Analysis of Variance (ANOVA) for Short Version of the Importance Items

(Scale: 1=Not at all important, 4=Very important)

Question & Factor	Mean Score (n)	F value	Pr >F
price...		16.54	.0001
fully labeled	3.73(102)		
partially labeled	3.58(99)		
high education	3.12(104)		
low education	3.65(97)		
how well the food keeps...		16.49	.0001
fully labeled	3.49(102)		
partially labeled	3.28(99)		
high education	3.12(104)		
low education	3.68(97)		

Table 13: Test of Significant Difference in Reliability for the Importance Question Series, by Education

(6 Items Total)

Education Level	#Signif Items p<0.05	Condition	Average Effect Size[9]
Low	1 0	Short Long	.18
Low	4 1	Card* No Card	.45
Low	3 0	Partial* Fully	.40
High	1 0	Short Long	.21
High	1 1	Card No Card	.26
High	2 0	Partial Fully	.35
Overall	1 0	Short Long	.16
Overall	1 0	No Card Card	.22
Overall	1 0	Partial Fully	.22

* Denotes statistical significance (z>=1.96, p<=0.05 and an average effect size >=0.30) across items in the question series.

[9]An effect size is a measure used to appraise differences in degrees of correlation. Cohen (1988) sets forth the following guidelines for interpretation of effect sizes: Generally, a small effect size (E.S.=.10) indicates that the group with the larger r^2 value has between 5%-8% greater shared variance between the first and second administrations of the question than does the other group. Above, the larger r^2 value corresponds to the group with the significant z-score test of differences. Similarly, a medium effect size (E.S. =.30) indicates that there is about 20% more shared variance for the group with the larger r^2 value. A large effect size (E.S.=.50) indicates that there is about 1/3 more shared variance for the group with the larger r^2 value relative to the other group.

Table 14: Average % of Respondents Who Were "Read All" of the Scale, by Education Group and Extent of Labeling For Importance Question Series

Extent of Labeling	Low Education		High Education	
	T1[10]	T2	T1	T2
Fully Labeled	82.2	72.0	75.4	60.8
Partially Labeled	66.7	63.9	57.5	44.3

[10] T1 refers to the first time the target question series was administered. T2 refers to the second time it was administered.

Table 15: Skewness for Short and Long Versions of the Five Frequency Items by Length of Scale

Length of Scale	Min. Skew	Max. Skew	No. < -1.25
Short	0.638	-0.890	0
Long	-0.035	-0.759	0

Table 16: Average Percent of Response distribution Falling in the Top and Top Two Categories across all Frequency Questions by Length of Scale

Length of Scale	Avg. % "Often/Always	Avg. % "Often" or "Always"
Short	30.6	65.3
Long	21.0	47.1

Table 17: Statistically Significant Results of the Analysis of Variance (ANOVA) for Short Version of the Frequency Items

(Scale: 1=Never, 4=Often)

Question & Factor	Mean Score (n)	F value	Pr >F
the nutrition panel...		11.64	.0001
low education	2.87 (97)		
high education	3.41 (104)		
fully labeled	3.31 (102)		
partially labeled	2.98 (99)		
short phrases on the label...		7.64	.006
fully labeled	3.08 (102)		
partially labeled	2.68 (99)		
list of ingredients		6.95	.001
low education	2.73 (97)		
high education	3.12 (104)		
fully labeled	3.08 (102)		
partially labeled	2.78 (99)		
the size of a serving...		11.02	.001
fully labeled	2.79 (102)		
partially labeled	2.33 (99)		

Table 18: Statistically Significant Results of the Analysis of Variance (ANOVA) for Long Version of the Frequency Items

(Scale: 1=Never, 5=Always)

Question & Factor	Mean Score (n)	F value	Pr >F
short phrases on label...		4.54	.036
low education	3.64 (47)		
high education	3.17 (52)		
list of ingredients...		4.91	.029
low education	3.77 (47)		
high education	3.29 (52)		
no. of servings in package...		9.75	.002
low education	3.51 (47)		
high education	2.77 (52)		
the size of a serving...		14.25	.0003
low education	3.61 (47)		
high education	2.71 (52)		

Table 19: Test of Significant Difference in Reliability for the Frequency Question Series by Education

(5 Items Total)

Education Level	#Signif Items $p<0.05$	Condition	Average Effect Size[11]
Low	0 0	Short Long	.25
Low	0 0	Card No Card	.09
Low	1 0	Partial Fully	.20
High	0 0	Short Long	.12
High	0 0	Card No Card	.21
High	0 0	Partial Fully	.11
Overall	0 0	Short Long	.14
Overall	0 0	No Card Card	.08
Overall	2 0	Partial Fully	.17

[11]An effect size is a measure used to appraise differences in degrees of correlation. Cohen (1988) sets forth the following guidelines for interpretation of effect sizes: Generally, a small effect size (E.S.=.10) indicates that the group with the larger r^2 value has between 5%-8% greater shared variance between the first and second administrations of the question than does the other group. Above, the larger r^2 value corresponds to the group with the significant z-score test of differences. Similarly, a medium effect size (E.S. =.30) indicates that there is about 20% more shared variance for the group with the larger r^2 value. A large effect size (E.S.=.50) indicates that there is about 1/3 more shared variance for the group with the larger r^2 value relative to the other group.

TABLE 20: ANOVA model of administration time by extent of labeling, type of scale, presence of show card, and the interaction of scale type and show card.

FACTOR	Mean administration time	F value	Pr > F
Extent of labeling		10.12	0.0016
partially labeled (n=148)	19.59		
fully labeled (n=152)	18.17		
Type of scale		7.14	0.0009
branching (n=103)	19.81		
short (n=98)	17.81		
long (n=99)	18.91		
Presence of card		---	n.s.
with card (n=150)	18.61		
without card (n=150)	19.14		
Card x Type		5.41	0.0049
with card*branch (n=54)	19.11		
with card*short (n=45)	18.57		
with card*long (n=51)	18.08		
no card*branch (n=49)	20.61		
no card*short (n=53)	17.14		
no card*long (n=48)	19.80		

Table 21: Average percent of "Read Part" of "Read None" codes
for Partial vs. Fully labeled item,
based on all items in the series with a scale, by question series.

	Fully		Partial	
Question Series	R Part	R None	R Part	R None
Disagree-Agree T1	7.6%	27.0%	9.4%	45.6%
Disagree-Agree T2	8.3	32.7	12.6	53.4
Importance T1	10.7	10.8	21.9	15.9
Importance T2	19.5	14.4	20.4	25.0
Frequency T1	5.6	11.9	18.6	16.9
Frequency T2	6.6	26.0	15.5	26.8

(11/27/95)

What We Eat in America
Diet and Health Knowledge Survey
(response scale test, version E)

START TIME: _____(am / pm)

INTRODUCTION

Today we are going to talk about your opinions on your diet, health, food shopping and related topics.

For some of the questions I am going to ask, you may not know the answer or you may not have an opinion. Actually, we expect that there may be some questions that you don't know about. In those instances, please tell me.

1a. A Food Guide Pyramid has been used to help explain the food groups and dietary guidelines. How familiar are you with the Food Guide Pyramid? Would you say you are not at all familiar, familiar in name only, somewhat familiar, or very familiar?

[1] not at all familiar –> *skip to Q2*
[2] familiar in name only
[3] somewhat familiar
[4] very familiar

1b. Let's talk about the recommended number of servings from different food groups that a person should eat each day. According to the Food Guide Pyramid, how many servings from the (FOOD GROUP) would you say you should eat each day for good health, or don't you know?

What about the (FOOD GROUP)?

	Food Group	No. Of Servings	DK
a.	**Fruit group?**		8
b.	**Vegetable group?**		8
c.	**Milk, yogurt and cheese group?**		8
d.	**Bread, cereal, rice and pasta group?**		8
e.	**Meat, poultry, fish, dry beans and eggs group?**		8

2. Now I am going to read some statements about what people eat. Please tell me if you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about the statement.

IF NEEDED: Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree or have no opinion about the statement?

STATEMENT	Strongly Disagree	Somewhat Disagree	Slightly Disagree	Slightly Agree	Somewhat agree	Strongly agree	N/O
a. Choosing a healthy diet is just a matter of knowing what foods are good and what foods are bad. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree or have no opinion about the statement?	1	2	3	4	5	6	8
b. Eating a variety of foods each day probably gives you all the vitamins and minerals you need. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree or have no opinion about the statement?	1	2	3	4	5	6	8
c. Some people are born to be fat and some thin; there is not much you can do to change this.	1	2	3	4	5	6	8
d. Starchy foods like bread, potatoes, and rice make people fat.	1	2	3	4	5	6	8

3a. The next few statements are about your own diet. Please tell me whether you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree or have no opinion about each statement.

There are so many recommendations about healthy ways to eat, I find it hard to know what to believe. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree or have no opinion about that statement?

[1] Strongly DISagree
[2] Somewhat DISagree
[3] Slightly DISagree
[4] Slightly Agree
[5] Somewhat Agree
[6] Strongly Agree

[8] No opinion

3b. **What I eat can make a big difference in my chance of getting a disease like heart disease.**

[1] Strongly DISagree
[2] Somewhat DISagree
[3] Slightly DISagree
[4] Slightly Agree
[5] Somewhat Agree
[6] Strongly Agree
[8] No opinion

3c. **My diet is healthy now, so there is no reason for me to make changes.**

[1] Strongly DISagree
[2] Somewhat DISagree
[3] Slightly DISagree
[4] Slightly Agree
[5] Somewhat Agree
[6] Strongly Agree
[8] No opinion

4. **Still talking about your own diet. The next few questions ask about the nutrients you obtain from foods, not from vitamin pills.**

To be your healthiest, do you think your diet has too much, too little or about the right amount of (NUTRIENT)?

How about (NUTRIENT)?

NUTRIENT	Too Much	Too Little	About Right	DK
a. Calcium	1	2	3	8
b. Iron	1	2	3	8
c. Fat	1	2	3	8
d. Cholesterol	1	2	3	8
e. Fiber	1	2	3	8

5. Now I am going to read you some statements about diet and nutrition. Please tell me if you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion.

Statement	Strongly Disagree	Somewhat Disagree	Slightly Disagree	Slightly Agree	Somewhat Agree	Strongly Agree	N/O
a. I should use salt or sodium only in moderation. – Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
b. I should eat at least 5 servings of fruit and vegetables a day.	1	2	3	4	5	6	8
c. I should use sugars only in moderation.	1	2	3	4	5	6	8
d. I should eat a variety of foods.	1	2	3	4	5	6	8
e. I should maintain a healthy weight. - - Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
f. I should choose a diet low in fat.	1	2	3	4	5	6	8
g. I should choose a diet low in cholesterol.	1	2	3	4	5	6	8
h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
i. I should eat at least two servings of dairy products a day.	1	2	3	4	5	6	8
j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8

6. **The next few questions ask you to make comparisons between foods. Each question is referring only to "regular" products; that is, foods that have NOT been specially formulated to be lower in fat. You may not know the answers to all of these questions. If so, just tell me that you do not know.**

Based on your knowledge, which has more fat: (READ PAIR. DO NOT PROBE "DON'T KNOW" ANSWERS.)

AS NEEDED: **Which has more fat** (NEXT PAIR)?

	Pairs	
a.	**Liver, or**	1
	T-bone steak	2
	THE SAME	3
	DON'T KNOW	8
b.	**Egg white, or**.....	1
	Egg yolk	2
	THE SAME	3
	DON'T KNOW	8
c.	**Skim milk, or**	1
	Whole milk	2
	THE SAME	3
	DON'T KNOW	8
d.	**Regular hamburger, or**	1
	Ground round	2
	THE SAME	3
	DON'T KNOW	8
e.	**Loin pork chops, or**	1
	Pork spare ribs	2
	THE SAME	3
	DON'T KNOW	8
f.	**Hot dogs, or**	1
	Ham	2
	THE SAME	3
	DON'T KNOW	8
g.	**Peanuts, or**	1
	Popcorn	2
	THE SAME	3
	DON'T KNOW	8
h.	**Yogurt, or**	1
	Sour cream	2
	THE SAME	3
	DON'T KNOW	8

7. If a food product is labeled "light," does that mean that compared to a similar product not labeled "light" it is: lower in calories, lower in fat, lower in calories and fat, or does it mean something else?

[1] LOWER IN CALORIES
[2] LOWER IN FAT
[3] LOWER IN CALORIES AND FAT
[4] SOMETHING ELSE
[8] DON'T KNOW

8. If a food label says a food is (DESCRIPTION), would you say you are not too confident, somewhat confident or very confident that the description is a reliable basis for choosing foods?

What about (NEXT DESCRIPTION)?

IF NEEDED:	"How confident are you that the description is reliable? Would you say not too confident, somewhat confident, or very confident?"

Description	Not too Confident	Somewhat Confident	Very Confident	N/O
a. light?	1	2	3	8
b. healthy?	1	2	3	8
c. extra lean?	1	2	3	8

9. Now I am going to read you some statements about diet and nutrition. Please tell me if you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion.

Statement	Strongly Disagree	Somewhat Disagree	Slightly Disagree	Slightly Agree	Somewhat Agree	Strongly Agree	N/O
a. I should use salt or sodium only in moderation. – Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
b. I should eat at least 5 servings of fruit and vegetables a day.	1	2	3	4	5	6	8
c. I should use sugars only in moderation.	1	2	3	4	5	6	8
d. I should eat a variety of foods.	1	2	3	4	5	6	8
e. I should maintain a healthy weight. - - Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
f. I should choose a diet low in fat.	1	2	3	4	5	6	8
g. I should choose a diet low in cholesterol.	1	2	3	4	5	6	8
h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
i. I should eat at least two servings of dairy products a day.	1	2	3	4	5	6	8
j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8

10. Nutrition may be a more important factor when you decide to buy some foods as compared to others. How important is nutrition when you decide to buy (KIND OF FOOD) – not at all important, not too important, somewhat important, important, or very important?

How important is nutrition when you buy (KIND OF FOOD)?

Kind of Food	Not at all Important	Not too Important	Somewhat Important	Important	Very Important	DON'T BUY THIS	N/O
a. fruits?	1	2	3	4	5	9	8
b. breakfast cereals?	1	2	3	4	5	9	8
c. lunch meats?	1	2	3	4	5	9	8
d. beef?	1	2	3	4	5	9	8
e. salad dressings?	1	2	3	4	5	9	8

11a. How important is your diet to your health? Not at all important, not too important, somewhat important, important, or very important?

[1] Not at all important
[2] Not too important
[3] Somewhat important
[4] Important
[5] Very Important

11b. How important is exercise to your health?

[1] Not at all important
[2] Not too important
[3] Somewhat important
[4] Important
[5] Very Important

12. How often do you or someone else wash fresh fruits and vegetables before you eat them – never, rarely, sometimes, often or always?

[1] Never
[2] Rarely
[3] Sometimes
[4] Often
[5] Always

13. When you eat fresh fruits with peels that can be eaten, how often do you eat the peel? Never, rarely, sometimes, often or always?

[1] Never
[2] Rarely
[3] Sometimes
[4] Often
[5] Always

14. When you eat fresh vegetables with peels that can be eaten, how often do you eat the peel?

[1] Never
[2] Rarely
[3] Sometimes
[4] Often
[5] Always

15. Do you eat the outer leaves of leafy vegetables like lettuce and cabbage? Would you say yes or no?

[1] Yes
[2] No
[8] DON'T KNOW
[9] DON'T EAT LEAFY VEGETABLES

16a. **Now think about buying food. When you buy food, how important is it that the food be safe to eat – not at all important, not too important, somewhat important, important, or very important?**

[1] Not at all important
[2] Not too important
[3] Somewhat important
[4] Important
[5] Very important

16b. **When you buy food, how important is (FACTOR)?**

IF NEEDED:	How important is (FACTOR) - not at all important, not too important, somewhat important, important, or very important when you buy food?

Factor	Not at all Important	Not too Important	Somewhat Important	Important	Very Important	N/O
1. nutrition? – not at all important, not too important, somewhat important, important, or very important	1	2	3	4	5	8
2. price?	1	2	3	4	5	8
3. how well the food keeps?	1	2	3	4	5	8
4. how easy the food is to prepare?	1	2	3	4	5	8
5. taste?	1	2	3	4	5	8

17. Now think about food labels. How often, if at all, do you use (SECTION) to help you decide whether or not to purchase a product – Never, rarely, sometimes, often or always?

IF NEEDED:	When deciding whether to purchase a product do you use (SECTION) never, rarely, sometimes, often or always?

How about (SECTION)?

SECTION	Never	Rarely	Sometimes	Often	Always	NEVER SEEN	N/O
a. The nutrition panel that tells the amount of calories, protein, fat and such in a serving of the food?	1	2	3	4	5	6	8
b. The short phrases on the label like 'low-fat' or 'light?' Never, rarely, sometimes, often or always?	1	2	3	4	5	6	8
c. The list of ingredients?	1	2	3	4	5	6	8
d. The information about the number of servings in a package? Never, rarely, sometimes, often or always?	1	2	3	4	5	6	8
e. The information about the size of a serving?	1	2	3	4	5	6	8

18. Now I'm going to read some statements. Please tell me if you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about each statement. (READ STATEMENT)

STATEMENT	Strongly DISagree	Somewhat DISagree	Slightly DISagree	Slightly Agree	Somewhat Agree	Strongly Agree	N/O
a. The nutrition information on food labels is useful to me. – Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
b. The nutrition information on food labels is hard to interpret. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
c. Reading food labels takes more time than I want to spend.	1	2	3	4	5	6	8
d. I would like to learn more about how to use food labels to choose a nutritious diet.	1	2	3	4	5	6	8
e. Sometimes I try new foods because of information on the food label. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
f. When I use food labels, I make better food choices.	1	2	3	4	5	6	8

19a. **Now think about buying food. When you buy food, how important is it that the food be safe to eat – not at all important, not too important, somewhat important, important, or very important?**

[1] Not at all important
[2] Not too important
[3] Somewhat important
[4] Important
[5] Very important
[8] No opinion

19b. **When you buy food, how important is (FACTOR)?**

IF NEEDED:	How important is (FACTOR) - not at all important, not too important, somewhat important, important, or very important when you buy food?

Factor	Not at all Important	Not too Important	Somewhat Important	Important	Very Important	N/O
1. nutrition? – not at all important, not too important, somewhat important, important, or very important	1	2	3	4	5	8
2. price?	1	2	3	4	5	8
3. how well the food keeps?	1	2	3	4	5	8
4. how easy the food is to prepare?	1	2	3	4	5	8
5. taste?	1	2	3	4	5	8

20. Now think about food labels. How often, if at all, do you use (SECTION) to help you decide whether or not to purchase a product – Never, rarely, sometimes, often or always?

IF NEEDED:	When deciding whether to purchase a product do you use (SECTION) never, rarely, sometimes, often or always?

How about (SECTION)?

SECTION	Never	Rarely	Sometimes	Often	Always	NEVER SEEN	N/O
a. The nutrition panel that tells the amount of calories, protein, fat and such in a serving of the food?	1	2	3	4	5	6	8
b. The short phrases on the label like 'low-fat' or 'light?' Never, rarely, sometimes, often or always?	1	2	3	4	5	6	8
c. The list of ingredients?	1	2	3	4	5	6	8
d. The information about the number of servings in a package? Never, rarely, sometimes, often or always?	1	2	3	4	5	6	8
e. The information about the size of a serving?	1	2	3	4	5	6	8

21. Some foods have lower fat or lower calorie products which can be chosen instead of the regular product. The next few questions ask about the choices you may make for certain kinds of food.

HABIT	Never	Rarely	Sometimes	Often	Always	Don't Eat This
a. When you eat luncheon meats how often do you choose to eat lower fat luncheon meats such as turkey - - Never, rarely, sometimes, often, or always?	1	2	3	4	5	9
b. When you use milk, how often do you choose skim or 1% milk?	1	2	3	4	5	9
c. When you eat cheese, how often do you choose to eat special, low-fat cheeses? Never, rarely, sometimes, often or always?	1	2	3	4	5	9
d. When you eat frozen dairy desserts, how often do you choose to eat foods like ice milk, frozen yogurt or sherbet?	1	2	3	4	5	9
e. When you use salad dressing, how often do you choose low-calorie salad dressing? Never, rarely, sometimes, often or always?	1	2	3	4	5	9
g. When you eat a meat dish, how often do you eat fish or poultry as the meat dish?	1	2	3	4	5	9

22. When you look for nutrition information on the food label, would you say you never, rarely, sometimes, often or always look for information about (STATEMENT)?

IF NEEDED:	"Would you say you never, rarely, sometimes, often or always look for information about that?"

What about (NEXT STATEMENT)?

STATEMENT	Never	Rarely	Sometimes	Often	Always
a. **Calories?**	1	2	3	4	5
b. **Salt or Sodium?**	1	2	3	4	5
c. **Fat?**	1	2	3	4	5
d. **Cholesterol?**	1	2	3	4	5
e. **Vitamins or minerals?**	1	2	3	4	5
f. **Fiber?**	1	2	3	4	5
g. **Sugars?**	1	2	3	4	5

END TIME: _____(am / pm)

TURN THE PAGE!

5. Now I am going to read you some statements about diet and nutrition. Please tell me if you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion.

Statement	Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	N/O
a. I should use salt or sodium only in moderation. – Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	8
b. I should eat at least 5 servings of fruit and vegetables a day.	1	2	3	4	8
c. I should use sugars only in moderation.	1	2	3	4	8
d. I should eat a variety of foods.	1	2	3	4	8
e. I should maintain a healthy weight. – Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	8
f. I should choose a diet low in fat.	1	2	3	4	8
g. I should choose a diet low in cholesterol.	1	2	3	4	8
h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	8
i. I should eat at least two servings of dairy products a day.	1	2	3	4	8
j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	8

Standard, 4-point, fully labeled

Without show card

5. Now I am going to read you some statements about diet and nutrition. Choose an answer between 1 and 4 with 1 being "strongly disagree," 4 being "strongly agree" and 2 and 3 being somewhere in between. If you do not have an opinion about the statement, tell me "no opinion."

Statement	Strongly DISagree 1	2	3	Strongly Agree 4	N/O
a. I should use salt or sodium only in moderation. Choose an answer between 1 and 4, with 1 being strongly disagree, 4 being strongly agree, and 2 and 3 being something in between or tell me "no opinion."	1	2	3	4	8
b. I should eat at least 5 servings of fruit and vegetables a day.	1	2	3	4	8
c. I should use sugars only in moderation.	1	2	3	4	8
d. I should eat a variety of foods.	1	2	3	4	8
e. I should maintain a healthy weight. – Choose an answer between 1 and 4, with 1 being strongly disagree, 4 being strongly agree, and 2 and 3 being something in between or tell me "no opinion."	1	2	3	4	8
f. I should choose a diet low in fat.	1	2	3	4	8
g. I should choose a diet low in cholesterol.	1	2	3	4	8
h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. Choose an answer between 1 and 4, with 1 being strongly disagree, 4 being strongly agree, and 2 and 3 being something in between or tell me "no opinion."	1	2	3	4	8
i. I should eat at least two servings of dairy products a day.	1	2	3	4	8
j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. Choose an answer between 1 and 4, with 1 being strongly disagree, 4 being strongly agree, and 2 and 3 being something in between or tell me "no opinion."	1	2	3	4	8

Standard, 4-point, partially labeled

Without show card

5. Now I am going to read you some statements about diet and nutrition. Please tell me if you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion.

Statement	Strongly Disagree	Somewhat Disagree	Slightly Disagree	Slightly Agree	Somewhat Agree	Strongly Agree	N/O
a. I should use salt or sodium only in moderation. – Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
b. I should eat at least 3 servings of fruit and vegetables a day.	1	2	3	4	5	6	8
c. I should use sugars only in moderation.	1	2	3	4	5	6	8
d. I should eat a variety of foods.	1	2	3	4	5	6	8
e. I should maintain a healthy weight. -- Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
f. I should choose a diet low in fat.	1	2	3	4	5	6	8
g. I should choose a diet low in cholesterol.	1	2	3	4	5	6	8
h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8
i. I should eat at least two servings of dairy products a day.	1	2	3	4	5	6	8
j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. Do you strongly disagree, somewhat disagree, slightly disagree, slightly agree, somewhat agree, strongly agree, or have no opinion about that statement?	1	2	3	4	5	6	8

Standard, 6-point, fully labeled

Without show card

5. Now I am going to read you some statements about diet and nutrition. Choose an answer between 1 and 6 with 1 being "strongly disagree," 6 being "strongly agree" and 2, 3, 4 and 5 being somewhere in between. If you do not have an opinion about the statement, please tell "no opinion."

Statement	Strongly DISagree 1	2	3	4	5	Strongly Agree 6	N/O
a. I should use salt or sodium only in moderation. Choose an answer between 1 and 6, with 1 being strongly disagree, 6 being strongly agree, and 2, 3, 4, and 5 being something in between or tell me "no opinion".	1	2	3	4	5	6	8
b. I should eat at least 5 servings of fruit and vegetables a day.	1	2	3	4	5	6	8
c. I should use sugars only in moderation.	1	2	3	4	5	6	8
d. I should eat a variety of foods.	1	2	3	4	5	6	8
e. I should maintain a healthy weight. – Choose an answer between 1 and 6, with 1 being strongly disagree, 6 being strongly agree, and 2, 3, 4, and 5 being something in between or tell me "no opinion".	1	2	3	4	5	6	8
f. I should choose a diet low in fat.	1	2	3	4	5	6	8
g. I should choose a diet low in cholesterol.	1	2	3	4	5	6	8
h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. Choose an answer between 1 and 6, with 1 being strongly disagree, 6 being strongly agree, and 2, 3, 4, and 5 being something in between or tell me "no opinion".	1	2	3	4	5	6	8
i. I should eat at least two servings of dairy products a day.	1	2	3	4	5	6	8
j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. Choose an answer between 1 and 6, with 1 being strongly disagree, 6 being strongly agree, and 2, 3, 4, and 5 being something in between or tell me "no opinion".	1	2	3	4	5	6	8

Standard, 6-point, partially labeled

Without show card

5. Now I am going to read you some statements about diet and nutrition. Please tell me if you disagree, agree, or have no opinion.

5a. I should use salt or sodium only in moderation. – Do you disagree, agree, or have no opinion about that statement?

[1] disagree – *Go to 5a(1)*
[2] agree – *Skip to 5a(2)*
[3] no opinion – *Skip to 5b*

Branching, partially labeled

Without show card

5a(1). How much do you disagree with that statement? – Choose an answer between 1 and 4, with 1 being slightly disagree, 4 being strongly disagree, and 2 and 3 being something in between.

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5b*

5a(2). How much do you agree with that statement? – Choose an answer between 1 and 4, with 1 being slightly agree, 4 being strongly agree, and 2 and 3 being something in between.

[1] slightly
[2] ....
[3] ....
[4] strongly

5b. I should eat at least 5 servings of fruit and vegetables a day.

[1] disagree – *Go to 5b(1)*
[2] agree – *Skip to 5b(2)*
[3] no opinion – *Skip to 5c*

5b(1). How much do you disagree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5c*

5b(2). How much do you agree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

5c. I should use sugars only in moderation.

[1] disagree – *Go to 5c(1)*
[2] agree – *Skip to 5c(2)*
[3] no opinion – *Skip to 5d*

5c(1). How much do you disagree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5d*

5c(2). How much do you agree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

5d. I should eat a variety of foods.

[1] disagree – *Go to 5d(1)*
[2] agree – *Skip to 5d(2)*
[3] no opinion – *Skip to 5e*

5d(1). How much do you disagree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5e*

5d(2). How much do you agree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

5e. I should maintain a healthy weight. Do you disagree, agree, or have no opinion about that statement.

[1] disagree – *Go to 5e(1)*
[2] agree – *Skip to 5e(2)*
[3] no opinion – *Skip to 5f*

5e(1). How much do you disagree with that statement? – Choose an answer between 1 and 4, with 1 being slightly disagree, 4 being strongly disagree, and 2 and 3 being something in between.

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5f*

5e(2). How much do you agree with that statement? – Choose an answer between 1 and 4, with 1 being slightly agree, 4 being strongly agree, and 2 and 3 being something in between.

[1] slightly
[2] ....
[3] ....
[4] strongly

5f. I should choose a diet low in fat.

[1] disagree – *Go to 5f(1)*
[2] agree – *Skip to 5f(2)*
[3] no opinion – *Skip to 5g*

5f(1). How much do you disagree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5g*

5f(2). How much do you agree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

5g. I should choose a diet low in cholesterol.

[1] disagree – *Go to 5g(1)*
[2] agree – *Skip to 5g(2)*
[3] no opinion – *Skip to 5h*

5g(1). How much do you disagree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5h*

5g(2). How much do you agree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

5h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. – Do you disagree, agree, or have no opinion about that statement?

[1] disagree – *Go to 5h(1)*
[2] agree – *Skip to 5h(2)*
[3] no opinion – *Skip to 5i*

5h(1). How much do you disagree with that statement – Choose an answer between 1 and 4, with 1 being slightly disagree, 4 being strongly disagree, and 2 and 3 being something in between.

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5i*

5h(2). How much do you agree with that statement – Choose an answer between 1 and 4, with 1 being slightly agree, 4 being strongly agree, and 2 and 3 being something in between.

[1] slightly
[2] ....
[3] ....
[4] strongly

5i. I should eat at least two servings of dairy products a day.

[1] disagree – *Go to 5i(1)*
[2] agree – *Skip to 5i(2)*
[3] no opinion – *Skip to 5j*

5i(1). How much do you disagree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 5j*

5i(2). How much do you agree with that statement?

[1] slightly
[2] ....
[3] ....
[4] strongly

5j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. – Do you disagree, agree, or have no opinion about that statement?

[1] disagree – *Go to 5j(1)*
[2] agree – *Skip to 5j(2)*
[3] no opinion – *Skip to 6*

5j(1). How much do you disagree with that statement? – Choose an answer between 1 and 4 with 1 being slightly disagree, 4 being strongly disagree, and 2 and 3 being something in between.

[1] slightly
[2] ....
[3] ....
[4] strongly

→ *Skip to 6*

5j(2). How much do you agree with that statement? – Choose an answer between 1 and 4 with 1 being slightly agree, 4 being strongly agree, and 2 and 3 being something in between.

[1] slightly
[2] ....
[3] ....
[4] strongly

ATTACHMENT B

5. Now I am going to read you some statements about diet and nutrition. Please tell me if you disagree, agree, or have no opinion.

5a. I should use salt or sodium only in moderation. – Do you disagree, agree, or have no opinion about that statement?

[1] disagree – *Go to 5a(1)*
[2] agree – *Skip to 5a(2)*
[3] no opinion – *Skip to 5b*

5a(1). Do you slightly, somewhat, mostly, or strongly disagree?

[1] slightly
[2] somewhat *Skip to 5b*
[3] mostly
[4] strongly

Branching, fully labeled

Without show card

5a(2). Do you slightly, somewhat, mostly, or strongly agree?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5b. I should eat at least 5 servings of fruit and vegetables a day.

[1] disagree – *Go to 5b(1)*
[2] agree – *Skip to 5b(2)*
[3] no opinion – *Skip to 5c*

5b(1). How much do you disagree with that statement?

[1] slightly
[2] somewhat *Skip to 5c*
[3] mostly
[4] strongly

5b(2). How much do you agree with that statement?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5c. I should use sugars only in moderation.

[1] disagree – *Go to 5c(1)*
[2] agree – *Skip to 5c(2)*
[3] no opinion – *Skip to 5d*

5c(1). How much do you disagree with that statement?

[1] slightly
[2]somewhat *Skip to 5d*
[3] mostly
[4] strongly

5c(2). How much do you agree with that statement?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5d. I should eat a variety of foods.

[1] disagree – *Go to 5d(1)*
[2] agree – *Skip to 5d(2)*
[3] no opinion – *Skip to 5e*

5d(1). How much do you disagree with that statement?

[1] slightly
[2] somewhat *Skip to 5e*
[3] mostly
[4] strongly

5d(2). How much do you agree with that statement?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5e. I should maintain a healthy weight. – Do you disagree, agree or have no opinion about the statement?

[1] disagree – *Go to 5e(1)*
[2] agree – *Skip to 5e(2)*
[3] no opinion – *Skip to 5f*

5e(1). Do you slightly, somewhat, mostly, or strongly disagree?

[1] slightly
[2] somewhat *Skip to 5f*
[3] mostly
[4] strongly

5e(2). Do you slightly, somewhat, mostly, or strongly agree?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5f. I should choose a diet low in fat.

[1] disagree – *Go to 5f(1)*
[2] agree – *Skip to 5f(2)*
[3] no opinion – *Skip to 5g*

5f(1). How much do you disagree with that statement?

[1] slightly
[2] somewhat *Skip to 5g*
[3] mostly
[4] strongly

5f(2). How much do you agree with that statement?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5g. I should choose a diet low in cholesterol.

[1] disagree – *Go to 5g(1)*
[2] agree – *Skip to 5g(2)*
[3] no opinion – *Skip to 5h*

5g(1). How much do you disagree with that statement?

[1] slightly
[2] somewhat *Skip to 5h*
[3] mostly
[4] strongly

5g(2). How much do you agree with that statement?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5h. I should eat at least 6 servings of breads, cereals, rice or pasta a day. – Do you disagree, agree or have no opinion about the statement?

[1] disagree – *Go to 5h(1)*
[2] agree – *Skip to 5h(2)*
[3] no opinion – *Skip to 5i*

5h(1). Do you slightly, somewhat, mostly, or strongly disagree?

[1] slightly
[2] somewhat *Skip to 5i*
[3] mostly
[4] strongly

5h(2). Do you slightly, somewhat, mostly, or strongly agree?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5i. I should eat at least two servings of dairy products a day.

[1] disagree – *Go to 5i(1)*
[2] agree – *Skip to 5i(2)*
[3] no opinion – *Skip to 5j*

5i(1). How much do you disagree with that statement?

[1] slightly
[2] somewhat *Skip to 5j*
[3] mostly
[4] strongly

5i(2). How much do you agree with that statement?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

5j. I should eat 2 to 3 servings a day from the meat, poultry, fish, dry beans or eggs group. – Do you disagree, agree, or have no opinion about that statement?

[1] disagree – *Go to 5j(1)*
[2] agree – *Skip to 5j(2)*
[3] no opinion – *Skip to 6*

5j(1). Do you slightly, somewhat, mostly, or strongly disagree?

[1] slightly
[2] somewhat *Skip to 6*
[3] mostly
[4] strongly

5j(2). Do you slightly, somewhat, mostly, or strongly agree?

[1] slightly
[2] somewhat
[3] mostly
[4] strongly

GO TO THE NEXT PAGE

Attachment C

BEHAVIOR CODING - Definition of Codes

Interviewer - Question Codes:

Remember, these codes only apply to the `question-part' of what the interviewer reads. For example, if the interviewer reads "How often do you or does someone else wash fresh fruits and vegetables before you eat them - Never, rarely, sometimes or often?" These codes only pertain to the part read before the dash, "How often do you or does someone else wash fresh fruits and vegetables before you eat them?"

Code ES: Exact question reading or only Slightly changed

Use code "ES" if the interviewer reads the question exactly as worded, or if the only deviation from an exact reading is omitting an article (i.e., the, an, a), changing a "that" to a "this," or adding transitional words, such as "and" or "now" at the beginning of an item can be considered an exact reading. Omitting the words "Look at category A on your card" also constitutes an exact reading. In other words minor, slight changes that do not alter the meaning of the question should be coded as an ES. If the interviewer stumbles while asking the question, but starts over and reads the question as worded, code this as ES. If the interviewer stumbles over a word and mispronounces it, code this as an ES as long as one could understand what the interviewer was trying to say. If the interviewer is interrupted while reading the question but finishes the question, code this as an ES also.

Code OS: Omit Stem of question (applies only to questions in matrix format)

Use code OS in the cases when the questions are formatted in a matrix, and the interviewer does not read the stem of the question for the second, third, fourth, etc., item in the matrix. The stem can be something simple like "What about..." or "How about" or it can be something more involved such as with Q16b "When you buy food, how important is...." These `stems' should be read between each item in the matrix. The point of this code is to determine how hard the procedures are for asking questions in a matrix format. These codes should only apply to items other than the first item in a matrix. Note that Q6 stem says "As Needed." You should still code this item as if the stem should be read every time. Thus if the stem "Which has more fat..." is not read, it should be coded OS.

Code MC: Major Change in Question Reading

Use code "M" if the interviewer changes item wording in a way that alters the intended meaning of the question. The omission of one or more key words, or entire phrases, are sufficient conditions for the use of this code. For example, if the interviewer omits the last part of the question "to help you decide whether or not to purchase a product" in Q17a or Q20a code this as a major change. Or for

example, if the interviewer omits the two words "a day" in any of the questions 6b, 6h, 6i, 6j, 9b, 9h, 9i, or 9j this should be coded as a MC.

If the meaning of the question has been altered as a result of a change in wording--no matter how minor the change may seem--code this behavior as a major change. It is sometimes difficult to differentiate between a slight change (coded as an ES) and a major change (coded as an MC). If the meaning of the question can be interpreted differently (even though it doesn't appear that it necessarily was), code this as a major change.

Whenever you use code MC, for major change, describe the change in the "comment" field. Write down exactly how the question was changed. Record what words were omitted, what words were changed, etc.

Code Om: Omit question (applies only to branching versions)

In the branching version of the questionnaire, the disagree/agree questions are split into two steps. The first step asks whether they disagree or agree with the statement. The second step asks how much they disagree, or how much they agree depending on their answer in the first step. Once respondents get used to the two step procedure, it may be that they answer "strongly agree" at the first step, rather than answering "agree" at the first step, and "strongly" at the second step. In the case when they do answer in one step, AND the interviewer does not read the second step question at all, use this code, Om. If the respondent answers in one step, but the interviewer still reads the second step question, this code should NOT be used. Instead, use ES.

Code Ot: Other

Use code O for situations that do not fit into the above mentioned categories. You should also include an explanation for why you used this code in the "comment" field.

A specific time to use the Ot code is when the interviewer verifies the respondents answer to a question. For example, in the disagree/agree branching questions, if the respondent gives the answer to the second step question at the first step, and the interviewer verifies the second step answer instead of asking the question, code this as Ot and record in the comment box "verified answer."

Use code O whenever it is impossible to determine from the audiotape what an interviewer has said or done. In most cases, when this code is used, it will be because of silent verifies or because the tape recording is of poor quality. In other cases, it may be necessary to use this code because of background noise.

Interviewer - Scale Codes:

Remember, these codes only apply to the `scale-part' of what the interviewer reads. For example, if the interviewer reads "How often do you or does someone else wash fresh fruits and vegetables before you eat them - Never, rarely, sometimes or often?" These codes only pertain to the part read after the dash, "Never, rarely, sometimes, or often?"

The `scale-part' of the question covers the part of the question that has the answer categories the respondent is to choose from. This include the words "Do you disagree, agree or have no opinion about that statement" in the first step of the branching questions. It also includes the words that refer to numerical scales, such as "Choose an answer between 1 and 4 with 1 being strongly disagree, 4 being strongly agree, and 2 and 3 being something in between or tell me no opinion"

The point of these codes is to find out how much of the scale and how often the scale is actually read in order to get an answer from the respondent. Note that not all questions include the scale wording in the question itself. For those items, the code will always be NS (No Scale). The coding sheets are formatted as to indicate whether or not the NS code is applicable to a question. If it is applicable, all other `interviewer-scale' codes are blacked out. If NS is not applicable, it is blacked out.

Code RA: Read All of the scale

Use this code when the scale is included in the question wording, and the interviewer reads the entire scale before the respondent answers. There are some words that are not part of the scale per se, but are included in the "scale-part" of the question. For example, in the disagree/agree branching questions, the scale-part of the question reads "Do you disagree, agree, or have no opinion about that statement?" The words "about that statement" are not considered to be part of the scale. Thus, if the interviewer does not read those words, but reads the rest of the scale, the code should be RA.

Code RP: Read Part of the scale

Use this code when the interviewer only reads some portion of the scale-part of the question. For example, in the disagree/agree standard questions the scale part of the question reads "Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree or have no opinion about that statement?" If the interviewer only reads "Do you strongly disagree, somewhat disagree..." before the respondent answers, it should be coded as RP.

Whenever you use the RP code, you also need to record in the comment field exactly how much of the scale was read. Record the last two words of the scale read to the respondent followed by the word <end>. For example, for the previous example, the following should be included in the comment field: "somewhat disagree <end>."

Code RN: Read None of the scale

Use this code when the interviewer reads none of the scale-part of the question. This also includes occasions when the only part of the scale-part of the question that the interviewer reads is "Choose an answer..." or "Would you say..." If these phrases are all that is read of the scale-part of the question, the code should be RN.

Code NS: No Scale included in question

As noted above, this code can only be used in the instances when the scale or response options are not included in the question wording. The questions to which this applies have been pre-marked on the coding sheets by blacking out all `interviewer-scale' codes other than NS. If possible, you should circle the NS code on your sheet. However, if you forget to do so, you do not have to go back and circle it later as an edit.

B. Respondent Codes

Coding respondent behavior is important for determining if respondents are having difficulty understanding the meaning of questions and for identifying sensitive questions. In this particular interview, respondent codes are also important in determining how respondents use the response scales. They will tell us whether respondents more often provide numerical responses or verbal response. In the case of the branching versions of the questionnaire, the respondents codes will also tell us how often respondents provide their answer in one step rather than two steps.

To capture the information particular to this interview, the respondent codes are split into those for numerical answers (AN, IN, QN, PN) and those for verbal answers (AV, IV, QV, PV). In most instances respondents will only answer with a numeric answer OR a verbal answer. However, for some versions of the questionnaire (B, Bc, D, Dc, F, Fc) it will be possible for respondents to give BOTH a numeric and verbal answer. <u>In those instances, you should use codes for both numeric and verbal answers.</u>

The codes you will be using to code respondent behavior are described below.

Again, note that only the <u>first</u> exchange between interviewer and respondent is coded in this study. However, there is one exception to this rule. You may run into situations in which the respondents first comment after the interviewer reads the question is a question to the interviewer about whether the question just asked is the same as something asked earlier. In these instances, and only these instances, the respondents first response after the question is read is NOT to be considered part of the first exchange. Rather, you should code the respondents actual answer. For example:

I: I should use salt or sodium only in moderation - Do you strongly disagree, somewhat disagree, somewhat agree, strongly agree, or have no opinion about that statement?

R: Didn't you already ask me that question?

I: As I said in the beginning of the interview, some of the questions may seem similar to one another, but that it is just part of the design processes.

R: Oh, yea I remember. Strongly agree.

In this exchange, both interviewer codes pertain to the first interviewer line, but the respondent codes should pertain to the second respondent line. Note that if the interviewer must repeat the question before the respondent answers the question, code the interviewer codes from the second reading of the question.

Code AV: Adequate Verbal answer

Use code AV if the respondent provides a verbal answer that meets the objective of the question and fits one of the response scale options for the question.

In instances when there is a series of questions (such as questions in the matrices) and the respondent answers "same answer" for anything other than the first question in the matrix, code this as AV. This should be an AV code even when the previous answer to which the respondent is referring was numeric.

For this interview a "don't know" or "no opinion" response is always considered an adequate verbal response, and should be coded as such -- even in instances when these are not options explicitly included on the questionnaire.

Code AN: Adequate Numeric answer

Use code AN if the respondent provides a numeric answer that meets the objective of the question and fits one of the response scale options for the question.

In instances when there is a series of questions (such as questions in the matrices) and the respondent answers "same answer" for anything other than the first question in the matrix, code this as AV. This should be an AV code even when the previous answer to which the respondent is referring was numeric.

Code IV: Inadequate Verbal answer

Use code IV if the respondent provides an answer that does not match, or cannot reasonably be classified, into one of the available codes. For example, if for questionnaire version C (a verbal branching version) the respondent answers the question about how much he/she agrees with the statement as a "a little bit" this should be coded as IV because the only valid response choices are "slightly, somewhat, mostly, or strongly."

Whenever you use this code, you should record in the comment field exactly what the response was. So in the above example you would record "a little bit" in the comment field.

Code IN: Inadequate Numeric answer

Use code IN if the respondent provides an answer that does not match, or cannot reasonably be classified, into one of the available codes. For example, if for questionnaire version D (a numeric branching version) the respondent answers the question about how much he/she agrees with the statement as a "5" this should be coded as IN because the only valid response choices are "1 - 4."

Whenever you use this code, you should record in the comment field exactly what the response was. So in the above example you would record "responded with a 5" in the comment field.

Code QV: Qualified Verbal answer

Use code QV if the respondent appears uncertain about the verbal answer he/she has provided and qualifies that answer in some way. For example, "For the most part, I think that is important" would be coded as a qualified verbal answer.

Also use this code when a respondent says "I don't know," and then gives a verbal answer anyway. A qualified answer expresses uncertainty or imprecision.

Code QN: Qualified Numeric answer

Use code QN if the respondent appears uncertain about the numeric answer he/she has provided and qualifies that answer in some way. For example, "Uhm, I guess I would say a 4" would be coded as a qualified numeric answer.

Also use this code when a respondent says "I don't know," and then gives a numeric answer anyway. A qualified answer expresses uncertainty or imprecision.

Code PV: Prior Verbal response

This code is only applicable to the branching versions of the questionnaire (C, Cc, D, Dc), and can only be used for second step of the disagree/agree questions.

Use this code in the second step of the disagree/agree branching questions when the respondent provides a verbal answer for this question, but in response to the first step of the branching question. For example, this code can apply to the case when, in response to the question (Q9c) "I should use sugars only in moderation - Do you disagree, agree or have no opinion about that statement?" the respondent answers "Strongly Agree." The

appropriate response for this first step of the question is either disagree, agree or no opinion. However, the respondent has told us the response to this first step question (he/she agrees) plus told us how much he/she agrees which is actually a response to the second step question. So the respondent code for the second step question should be PV because the answer to this question was given in the prior question, and it was a verbal response.

Code PN: Prior Numeric response

This code is only applicable to the branching versions of the questionnaire (C, Cc, D, Dc), and can only be used for second step of the disagree/agree questions.

Use this code in the second step of the disagree/agree branching questions when the respondent provides a numeric answer for this question, but in response to the first step of the branching question. For example, this code can apply to the case when, in response to the question (Q9c) "I should use sugars only in moderation - Do you disagree, agree or have no opinion about that statement?" the respondent answers "Slightly agree, number 1." The appropriate response for this first step of the question is either disagree, agree or no opinion. However, the respondent has told us the response to this first step question (he/she agrees) plus told us how much he/she agrees which is actually a response to the second step question. So the respondent code for the second step question should be PN and PV because the answer to this question was given in the prior question, and it included both a numeric and a verbal response.

Code PI: Prior Inadequate response

This code is only applicable to the branching versions of the questionnaire (C, Cc, D, Dc), and can only be used for the disagree/agree questions.

Use this code in the second step of the disagree/agree branching questions when the respondent provides an inadequate answer for this question but in response to the first step of the branching question. For example, this code can apply to the case when, in response to the question (Q9c) "I should use sugars only in moderation - Do you disagree, agree or have no opinion about that statement?" the respondent answers "Agree, Number 5." The appropriate response for this first step of the question is either disagree, agree or no opinion. However, the respondent has told us the response to this first step question (he/she agrees) plus told us how much he/she agrees (a 5) which is actually a response to the second step question. In this case, though, a 5 is not a valid answer, so the respondent code for the second step question should be PI because the answer to this question was given in the prior question, and it was an inadequate response.

Another more likely example would be for the respondent to answer "slightly agree, number 4" in response to the first step question. He/She has told us the answer to the first step question (agree), and has given us a verbal answer to the second step question (slightly agree), but the verbal answer does not correspond to the numeric answer also

given. Thus we are unable to tell whether the respondent meant "slightly" which corresponds to a "1," or "strongly" which corresponds to a "4." Thus the code must be PI, because the answer to the second step question was given in the prior question, but it was an inadequate answer.

Code NA: No Answer

This code should really only come up very rarely, if at all. You should use this code in instances when the respondent does not provide any answer to a question. "Don't Knows" and "No Opinions" are adequate verbal (AV) answers and should NOT be coded here.

Code RC: Request for Clarification

Use code RC whenever the respondent asks the interviewer to clarify the meaning of a particular question or concept. If possible, try to distinguish a request for clarification from a request to have the question repeated that is due to a hearing impairment or to surrounding noise. If hearing impairment or noise is the reason for the request, use code Ot, Other, and write in the comment field that the request was for a repeat of the question due to noise or hearing impairment.

If the respondent makes it clear why he/she is asking for clarification on a particular question (e.g. difficult word, concept, or reference period), please write this reason down in the comment field. If you don't know the reason, write down the respondent's exact wording. For example, if in response to the question "I should eat a variety of foods" the respondent asks "what do you mean by variety?" you should record in the comment field something along the lines of "unsure what is meant by variety."

Code Ot: Other respondent behavior

Use this code when the respondent behavior does not fit into any of the other respondent codes. If you use this code be sure to write in the comment field a clear description of what occurred.