A METHOD FOR IDENTIFYING COGNITIVE PROPERTIES OF SURVEY ITEMS

Barbara H. Forsyth and Michael L. Hubbard, Research Triangle Institute
P. O. Box 12194, Research Triangle Park, NC 27709-2194

Question wording can lead to error and bias in survey measurement (e.g., Forsyth & Lessler, 1991a; Groves, 1989; Turner & Martin, 1984; Sudman & Bradburn, 1974). This paper reports on research to develop and test a method for identifying survey items that are difficult for respondents to answer due to their cognitive demands. For example, items may be difficult to answer if the wording is difficult to understand, if response requires detailed memory recall, or if response categories fail to cover the range of respondent experience. Our aim is to design a taxonomy of item characteristics that can be used to identify potentially problematic survey items.

The research reported here was part of the methodological study of the National Household Survey on Drug Abuse (NHSDA) sponsored by the National Institute on Drug Abuse (NIDA). Major goals of the larger research project were (1) to identify potential sources of measurement error in the NHSDA, (2) to revise survey materials and survey procedures that seem to contribute to avoidable measurement error, and (3) to test revisions and identify improved strategies for measuring patterns of drug use. This paper focuses on reducing measurement errors that arise from the cognitive processes that respondents use when answering NHSDA survey items.

In 1974, Sudman and Bradburn reviewed research on survey response effects and concluded, "questionnaire construction and question formulation lie at the heart of the problem of response effects." Since then, survey methodologists have been applying theories and methods from cognitive science to study survey measurement error. The aim of this research is to enhance response accuracy by paying attention to the thought processes respondents use to answer survey items. In this arena, cognitive research methods are used to locate potentially problematic items and to revise them.

Many of the accepted cognitive research methods require extensive time commitments from respondents, interviewers and data analysts. In addition, the methods used and the nature of the data collected can vary widely between research facilities (e.g., Forsyth & Lessler, 1991b). As a result, it is often costly to collect data on respondents' thought processes, and it can be difficult to uncover general principles that describe sources of measurement error across diverse survey domains. This paper describes our attempts to develop and test a more cost effective method for studying the cognitive characteristics of survey items. Once validated, we hope that the method will provide a general language for describing item characteristics and their effects on response accuracy.

Our cognitive appraisal is designed to identify and highlight question features that may interfere with accurate reporting or response. We call our method the Cognitive Forms Appraisal coding scheme because the method relies on expert judgments to identify question features or question characteristics expected to affect measurement error and bias. We selected item features included in the coding scheme, based on a general model of survey response adapted from the model presented by Oksenberg & Cannell (1977). See Tourangeau & Rasinski, 1988 and Nathan et al., 1991 for similar cognitive frameworks).

We assume that respondents must complete five general cognitive tasks in order to respond to a questionnaire item. First, they must use **comprehension processes** to understand the question. Second, **interpretive processes** are used to construct a general representation of item task demands. This problem representation guides subsequent thought. For example, it may specify the kinds of information that the third set of **memory processes** must retrieve and compile toward the goal of answering the survey question. It may also specify goals relevant to the fourth set of **judgment processes**. Judgment processes use information retrieved from memory to forge assessments requested by survey items. These assessments are subjective evaluations or "feelings" rather than explicit, observable responses. Thus, a fifth set of **response generation** processes is needed to translate implicit judgments into overt responses that are acceptable under the survey instrument format.

The goal of our cognitive appraisal scheme is to develop a detailed view of item characteristics and their effects on response accuracy. Therefore, it is useful to partition each of the five general sets of cognitive processes into subsets that serve more local goals. For example, the general set of memory processes consists of (1) retrieval processes and (2)

information organization processes. Likewise, the general set of judgment processes consists of more specific integration and evaluation processes. We coded item characteristics relative to these more specific processes.

The full coding scheme is presented in Exhibit 1. The first set of columns in Exhibit 1 contains codes relevant to Comprehension, including instruction comprehension, question comprehension, and response comprehension. The columns contain additional subheadings that permit increasingly detailed descriptions of items and potential response errors. The codes under each of these subheadings indicate item features that either make written instructions misleading, make questions unclear, or that simply describe the cognitive processes that a respondent must implement to understand the survey materials. The remaining columns contain codes for Interpretation processes, Memory Judgment and Response Selection processes. It should be noted that some codes in Exhibit 1 can be used to represent hypotheses about survey response processes that might be explorerd using cognitive laboratory methods.

Illustrative Example Using Coding Scheme

We illustrate our coding scheme using Item C-8 from the Cigarettes section of the 1988 NIDA household survey questionnaire:

C-8 On the average, during most of this period when you smoked daily, about how many cigarettes did you smoke per day? (IF NEEDED, READ ANSWER CHOICES.)

One to five cigarettes a day.....1
About 1/2 pack a day (6-15 cigarettes).....2
About a pack a day (16-25 cigarettes).....3
About 1 1/2 packs a day (26-35 cigarettes)..4
About 2 packs or more a day (over 35 cigarettes).....5
NOT SURE.....94

When we reviewed the 1988 NHSDA questionnaire, we thought Item C-8 would be a tough item for respondents to answer. The codes assigned to Item C-8 suggest some reasons and helped us to select alternative wordings and formats that might improve response accuracy.

Item C-8 contains no instructions that are separate from the question content. Therefore, we selected none of the Instruction Comprehension Codes. Item C-8 has several characteristics that may interfere with Question Comprehension. There is technical term present, "average". "Average" is not explicitly defined, and therefore coded as undefined. We also coded the item for an vague or ambiguous term because in C-8, "average" may be given a technical interpretation or a more informal meaning such as "roughly" or "approximately". The question structure in item C-8 may also interfere with comprehension. The embedded clauses in C-8 led us to select the complex syntax code.

Item C-8 also demonstrates features that may interfere with Response Comprehension. The response categories refer to numbers of "packs" of cigarettes which may be a vague or ambiguous term. Different respondents may think of packs of cigarettes differently. For example, would 18 cigarettes a day be considered "about a pack?"

Furthermore, a single respondent's conception of "packs" of cigarettes may not correspond to the parenthetical category definitions. We coded the response categories as containing hidden definitions because the parenthetical definitions will not be read if the interviewer thinks they are not needed. The boundary problem code is used to indicate that respondents may have trouble establishing criteria that define and distinguish the response categories, particularly when interviewers choose not to read the parenthetical definitions. For example, in the absence of explicit instructions, where should a respondent draw the line between "about a pack a day" and "about 1 1/2 packs a day?"

Several features of Item C-8 may introduce error as respondents interpret the question reference period. We coded a carry-over reference period because the item refers to the reference period from the preceding item when it asks "During most of this period when you smoked daily...". We coded an ill- defined reference period because the question asks about "most" of that period. We selected the code indicating multiple interpretations because the reference period (the number of years the respondent smoked daily), is subject to individual interpretation. For example, some respondents may focus on their periods of heaviest smoking, while others may focus on the most recent months. The reference period has non-fixed boundaries because it refers to a time period that is respondent specific. It has unanchored boundaries because the item does not use marker events to set off the time covered by the reference period. Finally, the reference period length is tied to behavior because the reference period is defined according to the smoking habits of individual respondents.

The reference set may also be difficult to interpret. The question asks "about" how many

cigarettes," establishing a vague reference set, and "average number of cigarettes per day" is a complex reference set. Furthermore, under common interpretations of "average", the reference set implies a consistent pattern of behavior that may fail to characterize some respondents, making the response task more difficult and probably more error-prone.

We use the Task Definition codes to describe the task intended by researchers developing the questionnaire items. Thus, if respondents interpret Item C-8 as intended, they would recognize that the question-answering task involves the following subtasks: (1) Respondents must define the reference period and the reference set; (2) They must remember a set of episodes, their previous answer, and possibly some general information; (3) They must use these memories to make a judgment that corresponds to estimating an average; (4) They must generate a response.

We used additional codes to represent our hypotheses about how respondents complete the question-answering task for item C-8. Respondents may use a mixture of memory strategies for retrieving information. However, we expect that heuristic rules and inference will be prominent. We expect the information respondents retrieve from memory will consist of general self knowledge as well as sets of behavior episodes. We expect that respondents will use qualitative processes to formulate a subjective judgment. The item explicitly asks for a qualitative, ordinal response representing average smoking frequency. Nonetheless, we coded the judgement-response mapping as potentially incongruent because the qualitative categories that respondents spontaneously use may not coincide with the response categories available in Item C-8. We emphasize here that the Memory Process and Judgment codes represent hypotheses that might be validated by data from other cognitive methods such as think-aloud interviews or laboratory experiments.

General Coding Results

The 1988 NHSDA interview questionnaire consisted of seventeen sections. Items in eleven sections dealt with eleven specific substances (e.g., heroin, cocaine, tobacco). The remaining six sections asked for more general information about demographics, drug treatment experiences, and drug-related problems, among other things.

Three expert judges coded items in all seventeen sections of the questionnaire. There were several cases where the three judges disagreed about code assignments. Given the developmental status of the appraisal scheme, disagreements were resolved through discussion. Results presented here represent consensus codes recorded after resolving all disagreements among judges. This paper focuses on results from the eleven questionnaire sections dealing with specific substances.

Across the eleven drug sections, we computed the proportion of items receiving each appraisal code. The most frequent problem codes pertained to question and response comprehension, reference set interpretation, and reference period interpretation. Detailed results for these code categories are presented in Exhibit 2. Results for the question and response comprehension codes suggest at least two clusters of potential problem items. The first cluster consists of items using vague or ambiguous terminology to define response categories. The second cluster consists of items using response categories with hidden definitions that may not be read to respondents.

The results in Exhibit 2 suggest a third cluster of items using vague or technical terminology. In a questionnaire on substance use it is difficult to avoid using terminology that may sound technical to at least some portion of the household population. Some examples of terms coded as technical include "amphetamine", "sedative", and "prescription medication." Technical terminology may not be problematic if interview materials contain adequate definitions. It is more difficult to compensate for the use of vague or ambiguous terms that may be interpreted differently by different respondents. Thus, the question comprehension code results point to some instances where revised terminology or enhanced definitions might be developed.

Results related to reference period interpretation results suggest a fourth cluster of potential problem items that use unanchored or unfixed reference periods.

Think-Aloud Validation

We conducted a small study, using "think aloud" interviews to examine the validity of the cognitive appraisal coding scheme. Six "think aloud" interview respondents were recruited to represent different subpopulations within the general household population, including a mix of ages, races, education levels, and established history of substance use. Participants were instructed to report any thoughts or response strategies they were aware of as they answered items from the 1988 NHSDA interview questionnaire. The "think aloud" interview responses suggested that three general problems characterized several respondents' answers: (1) There were differences between respondents in how they

interpreted terminology identified as vague or ambiguous by the coding method; (2) Respondent reports suggested they had difficulties consistently defining and anchoring question time frames; (3) Respondents' answers suggested that item formats, using hidden questions and hidden category definitions introduced response inaccuracies. Thus, the problems identified based on "think aloud" interview protocols were similar to those identified by the coding results. The similarity provides preliminary evidence about coding method validity.

Conclusions

The appraisal results summarized here were used as one basis for identifying method improvements to test under more formal laboratory and field test conditions. Based in part on these appraisal results, we developed three sets of improvements. First, we used laboratory and field test procedures to investigate test decomposition approaches for defining technical terminology and complex reference sets. Second, we used laboratory and field test methods to test procedures for anchoring reference periods. Third, we used field test methods to test experimental questionnaire materials that eliminated hidden questions by using branching instructions and skip patterns. As reported in other papers in this volume, the experimental and field test results suggested that our appraisal methodology made an important contribution to identifying sources of response inconsistencies, response biases, and response variability.

We realize that additional research is necessary before we can use this coding scheme as a general purpose tool for analyzing survey items. We are currently working to clarify, refine, and trim our coding categories; collapsing some while expanding upon others. In addition, we need research to provide valid tests of the coding scheme once it has been refined. Although further development and testing is necessary, we believe that we have begun to develop a cost-effective method for systematizing expert evaluations and for identifying and cataloging critical aspects of questionnaire wording and format.

References

Ericsson, K.A. & Simon, H.A. (1980). Verbal reports as data. Psychological Review, 87, 215-251.

Ericsson, K.A. & Simon, H.A. (1978). Retrospective verbal reports as data. CIP Working Paper No. 388, Department of Psychology, Camegie-Mellon University.

Forsyth, B.H. & Lessler, J.T. (1991a). Cognitive laboratory methods: A taxonomy. In P..P. Biemer, R. M. Groves, L. E. Lyberg, N.A. Mathiowetz & S. Sudman (Eds.) Measurement Errors in Surveys. Wiley: New York.

Forsyth, B.H. & Lessler, J.T. (November, 1991b). A taxonomy of cognitive laboratory methods. Paper presented to the Workshop on Cognition and Survey Methodology, University of Utrecht, The Netherlands.

Groves, R.M. (1989). Survey Errors and Survey Costs. Wiley: New York.

Nathan G., Sirken, M., Willis, G. & Esposito, J. (November, 1990). Laboratory experiments on the cognitive aspects of sensitive questions. Paper presented to the International Conference on Measurement Errors in Surveys, Tucson, Arizona.

Oksenberg, L. and Cannell, C.F. (1977). Some factors underlying the validity of response in self report. International Statistical Bulletin, 48, 324-346.

Sudman, S. and Bradburn, N.M. (1974). Effects of time and memory factors on responses in survey. Journal of the American Statistical Association, 68, 805-815.

Turner, C.F. & Martin, E. (1984). Surveying Subjective Phenomena. Sage: New York.

Tourangeau, R. & Rasinski, K.A. (1988). Cognitive processes underlying context effects in attitude measurement. Psychological Bulletin, 103, 299-314.

EXHIBIT 1: Cognitive Appraisal Coding Scheme

COMPREHENSION

INSTRUCTIONS

MISLEADING INSTR

Conflicting Instructions (1)

Inaccurate Instructions (2)

UNCLEAR INSTR

Complex Syntax (3)

Unclear Examples (4)

Unclear Layout (5)

Hidden Instruction (6)

QUESTIONS

TECHNICAL TERM CODES

Present (7)

Undefined (8)

Ambiguous (9)

Vague (10)

QUESTION STRUCTURE

Hidden Question (11)

Unclear Goal (12)

Implicit Assumption (13)

Q/A Mismatch (14)

Complex Syntax (15)

Several Questions (16)

Several Definitions (17)

Violates Conversational Conventions (18)

RESPONSES

RESPONSE TERMINOLOGY

Ambiguous Categories (19)

Vague Terms (20)

Complex Syntax: Resp (21)

Hidden Definitions (22)

RESPONSE STRUCTURE

Boundary Problems (23)

Categories Not Mutually Exclusive (24)

Categories Not Exhaustive (25)

Non-Dominant Ordering (26)

DEFINITION OF COGNITIVE TASK

REFERENCE PERIOD

Unanchored Boundary (27)

Non-Fixed Boundaries (28)

Ref Period Change (29)

Ill-Defined Ref Period (30)

Carry-Over Ref Period Def (31)

Embedded Reference Period (32)

Undefined Ref Period (33)

Ref Period Length Problem (34)

Multiple Interpretation of Ref Period Possible (35)

REF PERIOD DESCRIPTION

Lifetime (36)

12 Months (37)

30 Days (38)

Tied to Behavior/Prev Q (39)

REFERENCE SET

Consistent Pattern of Beh implicit (40)

Vague Ref Set (41)

Complex Ref Set (42)

REFERENCE SET CHANGES

Domain Change (43)

Level Change (44)

Abrupt: Level + Domain (45)

Carry-Over Ref Set (46)

REFERENCE SET LEVEL

Basic (47)

Subordinate (48)

Superordinate (49)

Multilevel (50)

TASK DEFINITION

Estab Ref Set Boundary (51)

Est Ref Period Boundary (52)

Remember Episode (53)

Remember Set of Episodes (54)

Remember General Info (55)

Remember Previous Answer (56)

Determine +/- Ocurrence (57)

Determine +/- Match (58)

Determine Date/Onset (59)

Determine Age (60)

Estimate Duration (61)

Estimate Average (62)

Estimate Total (63)

Complex Estimation (64)

Recognize/Answer Hidden Question (65)

Generate Response (66)

INFORMATION RETRIEVAL

MNEMONIC PROCESSES

Recall (67)

Recognition (68)

Heuristic/Inference (69)

Mixed Above (70)

MEMORY CONTENT

General Self Knowledge (71)

General World Knowledge (72)

Specific Beh(or Try) (73)

Class of Behaviors (74)

Affect/Attitude (75)

Time Point/Interval (76)

PROBLEMS

High Detail (77)

Low Detail (78)

Unexpected Detail (79)

Shift:Psych Ref Period (80)

JUDGMENT

INFO INTEGRATION

Count (81)

Qualitative Judgment (82)

Quantitative Judgment (83)

INFO EVALUATION

Accuracy Eval Possble (84)

Sensitive Behavior (85)

Sensitive Attitude (86)

Sensitive (general) (87)

Socially Undesirable (88)

CONSEQUENCE EVAL

Safety Consequences (89)

Legal Consequences (90)

Social Consequences (91)

Behavioral Consqeuces (92)

RESPONSE GENERATION / SELECTION

RESPONSE DESCRIPTN

Yes/No (93)

Qualitative:Categor (94)

Qualitative:Ordinal (95)

Quantitative:Count (96)

Quantitive:Complex (97)

uration (98)

TimePoint/Most Rece (99)

Age (100)

INFORMATION/ RESPONSE CONGRUENCE

Congruent (101)

Incongruent (102)

Exhibit 2. Detailed Coding Results for Selected Problem Codes

Codes	Proportion of Items Coded
Question Comprehension	
Technical term(s)	.70
Vague term(s)	.38
Hidden Question	.57
Question-answer mismatch	.23
Unclear goal	.18
Implicit assumption	.06
Complex syntax	.56
Several questions	.14
Response Comprehension	
Vague or ambiguous terms	.32
Boundary problems	.43
Hidden definitions	.38
Complex syntax	.23
Non exclusive	.28
Non exhaustive	.08
Reference Period	
Unanchored	.45
Nonfixed boundary	.85
Ill-defined	.08
Carry-over	.04
Period change	.26
Undefined	.01