RESEARCH REPORT SERIES
(Survey Methodology #2006-8)

Vignettes and Respondent Debriefings for Questionnaire Design and Evaluation

Elizabeth Martin

Director's Office
U.S. Bureau of Census
Washington, D.C. 20233

Report Issued: December 1, 2006

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Vignettes and Respondent Debriefing for Questionnaire Design and Evaluation

Elizabeth Martin
U. S. Bureau of the Census

8.1 INTRODUCTION

Recent decades have seen theoretical and empirical advances in understanding the cognitive sources of measurement errors introduced by failures of comprehension or retrieval. In this chapter we describe how two methods, vignettes and respondent debriefing questions, can be used to identify measurement problems and craft and test questionnaire designs to address them. The focus is on their application in field-based tests of interviewer-administered questionnaires, although they also are used in laboratory and qualitative studies (the latter use is discussed) and with other types of questionnaires, such as self-administered ones. The chapter draws on research (much of it hitherto unpublished) conducted for the redesign of several Census Bureau surveys as well as other studies. Section 8.2 describes the use of vignettes for questionnaire design, drawing on research on problems of comprehension in the Current Population Survey (CPS). Section 8.3 describes respondent debriefing questions, drawing on research undertaken to redesign instruments for the National Crime Victimization Survey (NCVS) to reduce recall and reporting problems.

8.2 VIGNETTES

Vignettes have a long history in qualitative and quantitative research on social judgments, going back (at least) to Piaget's (1932/1965) use of "story situations" to investigate moral reasoning in children. Piaget offers an important rationale for using vignettes, as well as the main methodological question about their validity: " . . .while pure observation is the only sure method, it allows for the acquisition of no more than a small number of fragmentary facts . . .. Let us therefore make the best of it and . . . analyse, not the child's actual decisions nor even his memory of his actions, but the way he evaluates a given piece of conduct . . ..We shall only be able to describe [it] . . . by means of a story, obviously a very indirect method. To ask a child to say what he thinks about actions that are merely told to him--can this have the least connection with child morality?" (Piaget, 1932/1965:112-113)

In other words, the use of vignettes permits an investigator to gather data that could not otherwise be collected at all, or only for a small number of cases, but the question of whether evaluations of hypothetical situations relate to judgments in real life remains an issue. Piaget himself adopts a pragmatic approach to validity when he states that " . . .any method that leads to constant results is interesting, and only the meaning of the results is a matter for discussion" (1932/1965:114).

Vignettes are brief stories or scenarios that describe hypothetical characters or situations to which a respondent is asked to react. Because they portray hypothetical situations, they offer a less threatening way to explore sensitive subjects (Finch, 1987). Their specificity allows contextual influences on judgments to be examined. To preserve realism, qualitative researchers may create vignettes based on actual situations reported to them. These are then used to stimulate open-ended discussions with respondents to explore their reasoning and judgments. Quantitative researchers more often construct vignettes by systematically manipulating features in different vignettes, which are administered in a controlled, experimental evaluation of the factors that affect respondents' judgments. This strategy was devised by Rossi and his colleagues, who labeled their vignettes *factorial objects* to capture this approach to vignette design, in which cells in a formal experimental design were represented by different vignettes (Rossi and Anderson, 1982). Respondents in quantitative or qualitative studies may be asked to perform a task, such as ranking, rating, or sorting vignettes into categories, or projecting themselves into a vignette situation, to imagine what a vignette character would or should do or feel.

Although vignettes typically contain detailed descriptive information, they may vary in degree of elaboration, as suggested by several illustrative vignettes used in studies of normative judgments:

(1) "Armed street holdup stealing $25 in cash" (from Rossi et al.'s 1974 study of the seriousness of crimes).
(2) "Cindy M., a freshman, often had occasion to talk to Gary T., a single 65-year-old professor. She went to his office after class. She seemed worried and asked him about grades. He remarked that she was making good progress in class. He reached out and straightened her hair. He said that he could substantially improve her grade if she cooperated" (from Rossi and Anderson's 1982 study of judgments of sexual harassment).

In both studies, vignette features [the amount stolen in (1), and social setting, prior relationship, male physical acts, and five other dimensions in (2)] were manipulated systematically in order to evaluate their effects on judgments.

The use of vignettes as a methodological tool for designing and evaluating questionnaires is more recent. [They were first used to evaluate alternative instruments for the redesign of the National Crime Victimization Survey (Biderman et al., 1986).] This paper describes the use of vignettes to:

- Explore a conceptual domain
- Test consistency of respondents' interpretations with survey definitions, diagnose question wording problems, and assess uniformity of meanings
- Evaluate the effects of alternative questionnaires on interpretations of survey concepts
- Analyze the dimensionality of survey concepts.

8.2.1 Exploring concepts

This use of vignettes is similar to substantive investigations of conceptual domains, although more focused on survey concepts and the implications for framing survey questions. A good example is Gerber's (1990, 1994) exploratory research employing ethnographic interviews to examine how people think about residence, the language they use to describe it, and the factors they take into account in deciding how to report it in surveys. Gerber initially conducted open-ended interviews with 25 informants, many with tenuous living situations that made residence determination difficult. (Respondents were recruited from a homeless shelter and a church.) She gathered descriptive information about informants' residence patterns and elicited the terms they used to describe them. The following excerpt from an interview illustrates the distinction one informant made between "living" and "staying" and several of the criteria (e.g., intentions, location of belongings) she invoked to explain her situation:

A: "I'm just a friend of hers. I lost my apartment in December . . .. That's why I said I'm staying there, cause I'm not living there. I'm doing everything I can to find a way out of there.

Q: So you're not living there . . ..

A: Well, you would say I'm living there, I been there since December, but I'm just saying it's not mine . . .But I live there I bathe there, I sleep there, I dress there, my clothes are there not everything I own. Most of my things I got out of storage and took to my mother's, but basically everything I have to live with since December is there. As a matter of fact, it's packed up at the door. Because I'm trying to get out . . .." (Gerber, 1990:15-16).

Gerber (1990) used the situations reported by the first set of informants to construct vignettes that were the focus of a second set of interviews. The excerpt above was simplified into the following vignette: "Mary asked her friend Helen if she could stay with her for a few days while she looked for a place of her own. It has been five months since then. Mary's suitcases are still packed, and are at the front door. Should Helen count Mary as usually living there?"

All the vignettes described ambiguous living situations, and were used to elicit informants' calculations of residence. According to Gerber, "In making the judgment about a complex or ambiguous case, informants revealed what elements of the situation were important to them and what sort of logic they followed in arriving at a decision. In the course of the interview, I would vary the circumstances somewhat in order to follow out these trains of logic . . .." (1990:5-6). For example, she probed to determine if informants' answers to the vignette above would change for stays of longer or shorter duration. As illustrated in the excerpt above, when life circumstances were complex, her informants used various criteria to determine where someone lives. Respondents' calculations may lead them to omit marginal residents who should be listed on a census roster or include those who should not be. Gerber (1994) interviewed additional respondents using an expanded set of vignettes in a follow-up study to identify appropriate terminology for census roster questions. Several features of her use of vignettes are worth noting.

First, vignettes were culled from ethnographic sources, in order to present respondents with living situations they might actually encounter. As Gerber notes, "By providing respondents with situations they recognize as 'real', we were able to tap into the expectations and reactions which they would have in similar social circumstances. This increases our confidence that the way respondents reasoned during our interviews is similar to the judgments they make in reporting rosters in survey situations" (1994:4).

Second, Gerber (1994) took care to use neutral vocabulary in the vignettes, to elicit the vocabulary respondents naturally use to describe residence situations. For example, vignette characters were described as sleeping in a certain place or spending time with a particular person, rather than as "living" or "staying" there. She also attempted to create an entirely neutral probe that would elicit residence terms without actually using any ("What would you call the time X spends with Y?"), but respondents did not understand the probe. Therefore, the common term "live" was introduced early in the interview, to train respondents in the task. Other, less common terms used in the census rostering process were introduced using structured probes (e.g. "Is X a member of Y's household?") later in the interview to avoid biasing answers.

Third, the ambiguity of the vignette situations stimulated respondents to think through and articulate the criteria they would apply to decide where a person should be considered to live. Altering the details of a vignette in unstructured follow up probes helped clarify respondents' reasoning, as illustrated in the following interview excerpt:

A: Well, it seemed to me that if you had said he ate his meals and slept there, then I would consider that he lived there.
Q: . . .if we said he eats at his wife's house, but he always sleeps at his mother's . . .
A: I'd say that's a weird arrangement.
Q: That's weird, but would you say that changed where he lived?
A: Well, if he slept at his mother's, I would consider that he lived at his mother's. On a permanent basis . . .if he just slept there occasionally, I would not consider that he lived there . . ." (Gerber, 1994:9).

By separating eating and sleeping (and other circumstantial details), Gerber was able to develop a more nuanced understanding of which factors influenced the answer given.

Fourth, the tasks involving vignettes were readily understood, even by respondents without much education or fluency in English. Respondents often treated the task as a puzzle or game, and only one interview (of 37) in Gerber's 1994 study had to be terminated because the respondent did not understand the task. However, focusing on hypothetical situations influenced responses to a subsequent request for factual information, which elicited a number of obviously fabricated answers.

8.2.2 Testing Interpretations of Question Intent and Diagnosing Question Wording Problems

It is well known that small changes in question wording can substantially affect responses (see e.g., Schuman and Presser, 1981), presumably by affecting respondents' interpretations of question meaning. Despite their sensitivity to wording changes, respondents commonly misinterpret the intended meanings of survey questions (Belson, 1981). Vignettes provide a tool for investigating the effects of question wording and context on interpretations of survey concepts, as illustrated by research conducted for the 1994 redesign of the Current Population Survey (CPS), the source of official U. S. unemployment estimates.

The CPS questions used ordinary words (such as "work" and "looking for work") with technical survey meanings that were not communicated to respondents. This situation arose because concepts had been refined over the years but the questionnaire had remained largely unchanged since the 1940s. (The pertinent questions about work were, "What were you doing most of *last week* [working, keeping house, going to school] or something else?" and, if the respondent did not report working, "Did you do any work at all *last week*, not counting work around the house? [*Note*: If farm or business operator in hh., ask about unpaid work.]" The question about looking for work asked, "Has NAME been looking for work during the past 4 weeks?" and [if yes] "What has . . . been doing in the last 4 weeks to find work?")

A series of vignettes (shown in Table 8.1) portraying irregular employment situations was administered to about 2,300 respondents in a computer assisted telephone debriefing interview (CATI) conducted in 1988, immediately after a final CPS interview (Campanelli, Martin, and Creighton, 1989). Also shown in Table 8.1 are interviewers' responses to the same vignettes, administered as part of a "knowledge of concepts" test conducted with the CPS field staff (Campanelli, Rothgeb, and Martin, 1989).

Table 8.1. CPS Respondents' and Interviewers' Classifications of Vignette Situations

Vignette	Percent "Yes"[a]	
	Respondents	Interviewers
I asked you a question about *working* last week. Now, I'm going to read a list of examples. After each example, please tell me whether or not you think the person should be reported as *working* last week.		
(1) Last week, Susan only did volunteer work at a local hospital. Do you think she should be reported as *working* last week?	38% (1,973)	4% (1,458)
(2) Last week, Amy spent 20 hours at home doing the accounting for her husband's business. She did not receive a paycheck. Do you think she should be reported as *working* last week?	50% (1,977)	83% (1,324)
(3) Sam spent 2 hours last week painting a friend's house and was given 20 dollars. Do you think he should be reported as *working* last week?	64% (1,976)	93% (1,395)
(4) Last week, Sarah cleaned and painted the back room of her house in preparation for setting up an antique shop there. Do you think she should be reported as *working* last week?	59% (1,949)	66% (1,348)
Please tell me whether or not each of the following activities should be reported as *looking for work*		
(5) During the past 4 weeks, George has occasionally looked at newspaper ads. He hasn't yet found any jobs in which he's interested. Do you think he should report that he is *looking for work*?	36% (1,122)	37% (1,413)

Source: Campanelli, Rothgeb, and Martin (1989).

[a] Correct answers are "yes" to vignettes 2, 3, and 4 and "no" to 1 and 5. Missing data are excluded from calculations. Ns are given in parentheses. Vignettes were asked in the order shown.

The results revealed common misunderstandings and suggested the intended meanings of "work" and "looking for work" were not communicated by the questions as worded. For example, only half of respondents correctly interpreted unpaid work in a family business (vignette 2) as work, with many interviewers (17%) also classifying it incorrectly. Results for vignettes 2 and 4 suggested that the phrase, "not counting work around the house" might have led respondents (and some interviewers) to exclude legitimate work activities that took place in the home.

The misinterpretations shown in Table 8.1 were likely to lead to reporting error, although the vignette results did not, in themselves, provide direct evidence about its magnitude. The vignettes were asked of people who had been asked a pertinent question in the main survey, for whom a vignette situation may or may not have been relevant. Reporting error would result if a respondent whose actual situation a vignette describes misinterprets how it should be reported.

However, the combined use of respondent and interviewer classifications of vignettes supported inferences about the likely impact of misinterpretations upon the data (as discussed by Campanelli, Martin, and Rothgeb, 1991). For example, the situation portrayed in vignette 5 is problematic because many (over a third) interviewers as well as respondents erroneously considered this passive job search to be "looking for work." Overly broad interpretations would be expected to lead respondents to erroneously report passive job searches and interviewers to erroneously accept them. This inference was consistent with evidence of high rates of misclassification of passive job searches (Fracasso, 1989; Martin, 1987). The results suggested that the wording of the question led both respondents and interviewers (including highly experienced ones) to misinterpret its intent. The problem could not be overcome by additional interviewer training or experience, but required rewording the question to better communicate its intent.

Discrepancies between intended and actual interpretations of work were more serious because they were correlated with age, with older respondents generally defining work much too narrowly and younger ones too broadly (Campanelli, Martin, and Creighton, 1989). The correlation with age was consistent with a suspected underreporting of teenage work activities by older proxy respondents (such as parents), and underreporting of work activities by retirees.

Thus, the vignettes confirmed that key questions did not adequately communicate the intended meaning of important concepts. The results supported the need to revise question wordings and helped identify likely sources of misunderstanding.

Several points should be noted about the use of vignettes in this application. First, administering the vignettes as part of a CATI supplement made it possible to tailor the vignettes to be asked of people who had also been asked the target question in the main survey. Contextual validity was preserved, since the vignettes captured respondents' interpretations in the context of an actual survey interview, rather than in a laboratory setting from which generalization is less certain.

Second, administering the vignettes to probability samples made it possible to generalize about differences between groups in their interpretations of key concepts. This is not possible with convenience samples.

Third, administering vignettes to both interviewers and respondents yielded more information than either study alone would have provided. Situations that were poorly understood by both seemed most vulnerable to reporting error and pointed to a need for questionnaire revision and improved interviewer training to address them.

Fourth, response rates for the supplement were high and item nonresponse rates for the vignettes were low (less than 3% per item) (Campanelli, Martin, Rothgeb, 1991). As in the qualitative study, the vignette task does not appear to be overly difficult for respondents or to lead to high rates of "don't know" answers in general samples.

Fifth, the vignettes did not provide direct evidence about the magnitude of reporting errors. Rather, they provided feedback about misinterpretations of key survey concepts held by respondents who were asked a target question; misinterpretations may or may not result in error. One would expect a misinterpretation to result in error when the situation portrayed in the vignette applies, but this remains an inference.

8.2.3 Evaluating the Effects of Alternative Questionnaires on Interpretations

The CPS redesign also illustrates the use of vignettes to evaluate whether questionnaire revisions bring respondents' interpretations more in line with question intent. The CPS instrument went through several iterations of revision and testing, with a final split-sample comparison of old and proposed new questionnaires in a large national RDD sample in 1991. Vignettes were administered to one in 10 respondents after a final interview, in order to test whether interpretations were more standardized and consistent with CPS definitions under the new questionnaire. The last two columns of Table 8.2 present the 1991 vignette results by questionnaire version; selected 1988 results from Table 8.1 are included in the first column for comparison.

Table 8. 2. Classifications of Vignettes Following New and Old CPS Questionnaires in Two Surveys

Vignette	Percent "Yes"		
	Old q'aire 1988	Old q'aire 1991	New q'aire 1991
Earlier I asked you a question about working. Now I want you to tell me how you would answer that question for each of the persons in the following imaginary work situations. Would you report her/him as . . . *(Old qaire):* working last week, not counting work around the house? *(New qaire)*: working for pay (or profit) last week?*			
(1) Last week, Susan/Al put in 20 hours of volunteer service at a local hospital.	38%	37%[v]	4%
(2) Last week, Amy/Joe spent 20 hours at home doing the accounting for her husband's/his wife's business. She/he did not receive a paycheck.	50%	46% [v]	29%
(3) Sam/Diane spent 2 hours last week painting a friend's house and was given 20 dollars.	64%	61% [v]	71%
(4) Last week, Sarah/Jeff cleaned and painted the back room of her/his house in preparation for setting up an antique shop there.	59%	47%[y]	42%
Total N asked work vignettes	1,980	305	319

Source: Martin and Polivka (1995). Correct answers are "no" to 1 and "yes" to 2-4.

*The parenthetical "or profit" was used after vignettes 2 and 4, to which it was applicable.

[v] difference between versions in 1991 significant at p<.05

[y] difference between years (old q'aire only) significant at p<.05

Several methodological differences between the 1988 and 1991 vignette studies should be noted. Vignette 1 was reworded slightly to avoid using the word *work* and perhaps biasing respondents' classifications. (As it turns out, the same fraction classified vignette 1 as "work" in the first and second columns.) Second, the introduction was revised to repeat key portions of the target question, to ensure that classifications were contextualized by the question that respondents had been asked earlier. Similarly, the wording of follow-up questions more closely mirrored the target question to reinforce the effect of question wording. Third, another work vignette (not shown) was added before vignette 1 in 1991 while a vignette that had preceded vignette 4 in 1988 was dropped (the latter is discussed below). Vignette responses were probed ("Why would you consider/not consider that person to be working?"). Finally, the gender of the subject person was experimentally manipulated to examine whether men's and women's work activities were viewed differently. A random half of respondents received a vignette with a female name, and the other half received a male version (these results are discussed below). Results in Table 8.2 combine male and female versions[1].

Results in Table 8.2 support several conclusions about using vignettes to compare question wording and context effects. First, except for one vignette, the old questionnaire evoked the same classifications in both 1988 and 1991 surveys. Results in the first and second columns are similar despite the survey differences described above. The replicability of vignette results in independent surveys using the same questionnaire suggests they reliably capture the effects of questionnaire context and wording on interpretations. In addition, the response structure remained stable. Martin and Polivka (1995) fit log linear models to joint distributions of responses to the work vignettes in the 1988 and 1991 surveys (following the old questionnaire), and found that the same model described associations among items in both years. The one significant difference was a drop from 1988 to 1991 in "yes" responses to vignette 4, which may be due to a contrast effect. In 1988, vignette 4 immediately followed another that described donating blood for money, which few respondents considered work (it was dropped in 1991). The contrast to selling blood may have made setting up an antique shop seem more like "work." In addition to being sensitive to the context created by questions in the main survey, vignette responses also may be vulnerable to context effects created by the order in which they are asked.

Second, the questionnaire modifications partially succeeded in bringing respondents' interpretations more in line with CPS, but also may have led to some new misinterpretations. Responses to the "why" probe confirmed that the meaning of work in the old question was vague. Respondents gave more, and more various, reasons for their vignette responses, including the irrelevant consideration of location. In contrast, the revised wording focused their attention on payment. Comparison of the second and third columns in Table 8. 2 shows that the revised wording reduced positive responses to several vignettes (1 and 2) that did not involve payment, and broadened respondents' interpretations to more often include casual paid labor (vignette 3). Unfortunately, the narrower focus on pay led some respondents to rely exclusively on present payment, and rule out some legitimate work activities not yet yielding pay or profit, such as those in vignette 2.

The experimental manipulation of gender also suggested that the new questionnaire created a more gender-neutral interpretation of work. The meaning of work in the old questionnaire was vulnerable to gender bias. Vignette 1 (and other "helping" vignettes not shown) were more likely to be classified as work if the subject was female, and male respondents were more sensitive to gender than female respondents. In other words, "helping is 'women's work', if you ask men" (Martin, Hess, and Siegel, 1995:43). The focus of the revised

[1] Classifications of vignette 2 were significantly affected by subject gender. "Amy" was more likely than "Joe" to be considered as working (for the Amy version, 54% and 34% answered "yes" following old and new questionnaire versions, respectively).

question wording on "pay or profit" eliminated the effects of both respondent and subject gender on classification of "helping" vignettes (Martin, Hess, and Siegel, 1993). Thus, the new questionnaire elicited more gender-neutral interpretations of helping activities, as well as reducing the extent to which respondents thought they should be reported at all.

Did altered interpretations of "work" influence reporting under the new questionnaire? Evidence on this question is somewhat mixed. The expanded frame of reference indicated by results for vignette 3 should increase reporting of casual employment in the new questionnaire, and this prediction is borne out by evidence. A larger fraction of persons 16-19 years old (but not of older persons) were reported as working (Martin and Polivka, 1995) and there were significantly more reports of work activities involving a few hours. A slight gender bias due to underreporting of the number of female workers was eliminated. Thus, evidence from several sources suggested the new questionnaire was more inclusive of casual labor.

In another situation, the new questionnaire narrowed respondents' frame of reference too much, leading them to exclude unpaid work in a family business (vignette 2). Nevertheless, the new questionnaire elicited more, not fewer, reports of unpaid work in a family business (Polivka and Rothgeb, 1993), because a direct question about unpaid work in a family business was added. Respondents were no longer expected to understand they should report it in response to a general question about working. Thus, the questionnaire solution was to add a specific question, rather than try to improve respondents' understanding of a complex concept.

Several conclusions about the method are suggested by this research. First, vignette classifications are highly sensitive to questionnaire context. Even relatively small samples provide useful feedback about the effect of questionnaire revisions on respondent interpretations. This is especially useful when (as is true for CPS as well as many other surveys) really enormous samples are required to detect actual reporting differences for specific, relatively rare situations. This conclusion is also consistent with research on the NCVS (Biderman et al., 1986).

Second, research to date suggests that vignettes are reasonably robust measures of context and question-wording effects on interpretations. Despite survey differences, similar results were obtained in a replication of the vignettes in two independent surveys using the same questionnaire. Additional research is needed to establish the conditions under which vignettes reliably measure context and wording effects, but results such as those in Table 8.2 are promising. An apparent contrast effect induced by the order of the vignettes in one survey provides a caution that exact replication of a vignette series is necessary to ensure comparable results. Caution is also warranted in using vignette results to make improvements in items; improved items need to be evaluated *in situ*, in the context of a revised questionnaire.

Third, the open-ended "why" probes proved useful in understanding respondents' reasoning and interpreting wording effects.

Thus, revising questions based on vignette results and retesting the revised questions using the same vignettes can tell a questionnaire designer whether or not the questionnaire revisions have addressed the problems of interpretation satisfactorily. However, as illustrated by the CPS example, a questionnaire revision may correct one misunderstanding but create a new one. This reinforces the need for several rounds of testing, to ensure that problems have been addressed satisfactorily and that new problems have not arisen (see also Esposito, 2002; Esposito et al., 1992).

8.2.4. Exploring the Dimensionality of Survey Concepts

By analyzing the joint distribution of responses to a set of vignettes rather than analyzing them one at a time, a questionnaire designer can assess the global effects of a questionnaire revision on the inclusiveness or exclusiveness of the underlying concept.

For example, Martin, Campanelli, and Fay (1991) applied the Rasch measurement model to the joint distribution of the work vignettes to examine whether a latent dimension of meaning accounted for response patterns, or alternatively, whether respondents applied different criteria to classify each vignette situation, with no unifying concept[2]. Their analysis suggested that both were true to some extent. The data were consistent with a latent dimension of inclusiveness, with respondents who held a strict interpretation of work at one end and respondents willing to include marginal activities at the other. Beyond a propensity to be inclusive (or not), responses to three pairs of vignettes were associated, suggesting that respondents applied similar criteria or rules to classify each pair (e.g., vignettes 2 and 4 were associated, perhaps because some respondents ruled them both out because they took place at home). One vignette (about donating blood for money) could not be scaled with the others and therefore did not belong in the same conceptual domain; it did not partake of the meaning of "work." Alternative scorings of the vignette responses, as "yes/no" and as "correct/incorrect", led to the conclusion that the dimension underlying respondents' classifications was inclusiveness, not correctness. This implied that vignette series should include a balance of items with both "yes" and "no" correct answers, to avoid confounding correctness and inclusiveness.

Ideally, if the Rasch model had fit (with no additional between-item associations beyond those attributable to the underlying latent dimension), a very simple scale formed by summing the number of "yes" responses could have been used as a practical tool to evaluate the inclusiveness of respondents' interpretations of work. (See Duncan, 1984, for discussions of scaling and Rasch models.)

Martin and Polivka (1995) took this analysis one step further to assess the effects of the questionnaire revision on the structure of responses to the vignettes. A series of log linear analyses showed that the revision affected respondents' interpretations of particular types of situations, and hence responses to particular vignettes and associations among them, but did not globally broaden or narrow their interpretations of work.

Thus, modeling vignette responses can yield insight into the underlying dimensions of meaning evoked by a survey questionnaire, whether a questionnaire defines global or particular meanings, and whether respondents use rules or heuristics to judge situations. When the goal is specificity, the failure of the Rasch model to adequately describe vignette data may be taken as evidence of improvement. When a questionnaire designer intends to measure a global construct, abstracted or generalized from particular situations, it would be desirable to find that the Rasch model fits, and that respondents adopt global rather than specific rules for classifying different situations.

[2] The Rasch measurement model (Rasch, 1960/1980) treats each response as the product of two parameters: one unique to the item, and the other unique to the person. When items are scored "yes/no", the person parameter represents an individual's latent tendency to interpret the concept of work inclusively or restrictively. The item parameter represents how difficult or easy an item is; in other words, how congruent the activity described in the vignette is with respondents' underlying concept of "work." The Rasch measurement model is useful for analyzing the dimensions underlying a set of dichotomous items (see Duncan, 1984, for a detailed discussion). Other models, such as factor analysis, are appropriate when response categories can be considered to form an interval scale.

8.3. RESPONDENT DEBRIEFING QUESTIONS

A second method is similar to vignettes in being administered following a field interview, but encompasses more diverse types of questions. Typically, respondents are told that the main interview is complete and then asked general probing questions or standardized, retrospective questions about their experience of the interview, how they answered or interpreted questions, and other topics.

The respondent debriefing method reinvents and adapts probing techniques that have been used for decades to examine question meaning in survey pretests or the survey itself. In 1944, Cantril and Fried conducted an intensive study of 40 respondents, using follow-up probes to identify specific misunderstandings of poll questions. For example, answers to the question, "After the war is over, do you think people will have to work harder, about the same, or not so hard as before?" were probed by asking "When you said (harder/about the same/not so hard), were you thinking of people everywhere and in all walks of life–laborers, white-collar workers, farmers and business men-or did you have in mind one class or group of people in particular?" About half thought "people" meant everybody, a third interpreted the word as indicating a particular class, and a tenth didn't know what it referred to.

Belson (1981) relied on similar probes when he introduced the *question-testing method* using what he called *double interviews* to identify problems of question understanding. Target questions were embedded in a questionnaire administered in personal interviews. A second, intensive interview was conducted the following day by another, specially trained interviewer. Respondents were paid for the second interview, which began with informal conversation about the previous day's interview, then (for each of seven target questions) reminded respondents of the question and their answer, and interrogated them extensively about how they understood the question and arrived at their answer. Belson notes that "the intensive interviewer was responsible also for probing for a full reconstruction, for challenging inconsistencies between the indications of the present evidence and the answer actually given 'yesterday', and for keeping the respondent thinking of how she answered the question yesterday as distinct from her interpretation of it now" (1981:35-6). Belson evaluated specific misunderstandings of terms (e.g., "you," "usually") and developed hypotheses about sources of misunderstanding. He concluded, "There is simply no way in which standard piloting can be used reliably to reveal the many misunderstandings of respondents, many of them unsuspected by the respondent himself and not visible to the piloting interviewer . . .. Direct question testing is essential" (1981:390,397).

The various applications of respondent debriefing or special probes have several methodological features in common. First, question meaning (or other response issue) is evaluated in the context of a real survey interview, typically conducted in the field. Second, respondent debriefing questions or special probes are standardized and asked after the main interview is complete, to avoid influencing responses to survey questions. (Other uses of special probes (e.g., Schuman, 1966) employ them as part of the interview, immediately after a question has been asked.) Third, the method frankly enlists respondents' help in improving a survey by inviting them "to assume a new role: to become informants rather than respondents" (Oksenberg, Cannell, and Kalton (1991:357) by commenting and elaborating on their interview experiences. Fourth, studies that employ special probing methods conclude that misunderstandings are quite common, but that respondents (and interviewers) are largely unaware of them.

Respondent debriefing studies vary considerably in scope, in the amount of time that elapses between original and debriefing interviews, and whether the debriefing interview takes place in the lab or in the field. Respondent debriefing questions used in pretesting[3] may probe for:

- Interpretations of terminology, questions or instructions
- Subjective reactions or thoughts during questioning
- Direct measures of missed or misreported information

8.3.1. Meaning Probes

Probes to test interpretations of terminology or question intent are the most common form of debriefing question, and their usefulness for detecting misunderstandings is well documented (Belson, 1981; DeMaio, 1983a; DeMaio and Rothgeb, 1996; Oksenberg, Cannell, and Kalton, 1991; Schuman, 1966). An illustration is drawn from an evaluation (Von Thurn, 1996) of interpretations of the term *regular school* in the following question:

"Is . . .attending or enrolled in regular school? (Regular school includes elementary school, high school and schooling that leads to a college or professional school degree.)"

Reviewers doubted that the term was meaningful to respondents, even with the parenthetical definition, which was judged likely to be interrupted by respondents or skipped by interviewers. To test interpretations, several open- and closed-ended debriefing questions were administered in the field after completion of the interview: "Earlier I asked if . . . is attending regular school. What does the phrase regular school mean to you in this question? Would a technical or vocational school be considered a regular school?”

Both open and closed probes confirmed that regular school was poorly understood, with closed probes providing more usable information. Responses to the open probe required coding, and many were too general or meaningless to be categorized as correct or incorrect (Von Thurn, 1996).

Oksenberg, Cannell, and Kalton (1991) found that comprehension probes, similar to the closed probes in the example above, were useful for revealing misinterpretations of key terms in survey questions. (An example from their research, following a question about consumption of red meat, was "Would you include things like bacon, hot dogs, or lunch meats as red meat?"). Similar probes have been asked to test interpretations of reference periods, such as "the past 12 months" or "last week" (see Campanelli, Martin, and Creighton, 1989; Hess and Singer, 1995; Moyer et al., 1997). Other types of probes also have proved useful for uncovering misinterpretations. Hess and Singer (1995) asked respondents in a field pretest to paraphrase survey questions about hunger (" . . .Could you tell me in your own words what that question means to you?"), and found that several complex questions were commonly misunderstood. However, Oksenberg, Cannell, and Kalton (1991) found that general "tell me more" probes and probes for direct reports of problems were not productive. (An example of the latter, following a question about illnesses "that kept you in bed for more than half of the day", was "How clear was it to you what to include as a half day in bed?"). Perhaps respondents were reluctant to admit or were unaware of their own misunderstandings. The authors noted that "respondents did not appear to doubt their own, often mistaken, interpretations" and concluded that "the particular

[3] Another type, not discussed here, is a more general “debriefing” about a prior interview, which is not a pretest or evaluation of survey questions, but may yield insights about questionnaire problems. For example, a large-scale reinterview was conducted after the 1980 census to learn more about the mail response process and perceptions of the census form (DeMaio, 1983b). In another example, Wobus and de la Puente (1995) conducted telephone debriefing interviews to learn respondents’ reactions to receiving both English and Spanish language forms in the mail as part of a census test.

strength of special probes lies in their ability to reveal problems that are not evident in interview behavior" (1991:358, 363). Other authors (DeMaio and Rothgeb, 1996; Morton-Williams and Sykes, 1984) reach similar conclusions. Research also shows that revising survey questions to correct problems revealed by special probes appeared to reduce misreporting (Fowler, 1992).

8.3.2 Thoughts and Subjective Reactions

Debriefing questions about respondents' thoughts or feelings have been used to address a variety of questionnaire issues.

Question sensitivity Miller and Davis (1994) conducted a field pretest of potentially sensitive questions with 29 mothers of children whose fathers lived elsewhere. Of particular concern were questions about whether the child's paternity had been established. After the pretest interview, respondents were asked, "Were there any questions in this interview that you felt uncomfortable answering?" and about one in five expressed discomfort with the paternity questions. Similar questions have been asked to assess discomfort with a request for social security numbers in a mailed census form (Bates, 1992) and sensitivity of long form census questions (Martin, 2001). Debriefing questions also may be asked about respondents' sensitivity to specific features of a survey's design. For example, in a debriefing conducted after adolescent respondents filled out a self-administered questionnaire that included sensitive questions, Hess et al. (1998) learned that respondents would be more concerned about the privacy of their answers if survey questions were printed where others might read them than if the questions were administered using an audio cassette tape player.

Confidence Debriefing questions that ask respondents about their certainty or confidence in their answers have thus far not provided useful information about questionnaires (see e.g., Oksenberg, Cannell, and Kalton, 1991). Respondents typically express high levels of confidence and certainty, which appear to have little relationship to the correctness of their interpretations of survey questions. Campanelli, Martin, and Rothgeb (1991) found no correlation between respondents' confidence and how well their classifications of various employment situations corresponded with CPS definitions. Moyer et al. (1997) found that respondents who misinterpreted the reference period for an income question were more confident of their answers than respondents who correctly interpreted the question (probably because they had reinterpreted the question to be one they could answer with confidence). Schaeffer and Dykema (Chapter 23, this volume) report that respondents' use of "doubt words" was related to accuracy.

Mental Processes Questionnaire pretests do not usually employ retrospective debriefing questions to assess respondents' mental processes while answering survey questions, and it is not clear how fruitful this might turn out to be as a pretesting method. Oksenberg, Cannell, and Kalton (1991) had limited success asking respondents how they came up with their answers to survey questions. It may be difficult for respondents to recall what they were thinking while answering a prior question, especially if other questions have intervened, and survey interviewers may not be skilled at asking what are in effect "think aloud" probes of the sort typically asked in cognitive interviews. On the other hand, Moyer et al. (1997) obtained useful information from a probe asking respondents how they came up with their answers to a survey question, as did Blair and Burton (1987) when investigating respondents' retrieval strategies (their debriefing question immediately followed the pertinent survey question, however).

An example of the use of debriefing questions to assess respondents' mental processes is drawn from research to redesign and test alternative screening questions for the National Crime Victimization Survey. The major goal of the redesign project was to reduce severe underreporting of victimizations that had been documented by record check studies (Biderman et al. 1986). The redesign of the screening questions was informed by a theory of cognitive barriers to recall and reporting of victimization incidents (see, e.g., Martin et al. 1986; Biderman, 1980a, b, 1981a, b; Sparks, 1982; Loftus and Marburger, 1983) and a screener designed to address the problems was tested against the standard screener in a split-panel CATI field experiment (Martin, Groves, Matlin, and Miller, 1986). Debriefing questions were asked after the interview to test whether the revised screener reduced hypothesized sources of retrieval and reporting problems, including failure of metamemory, recall interference, fatigue and negative response sets, mnemonic failure, and selective reporting.

Failure of metamemory Psychological research (e.g., Hart, 1965) had suggested that retrieval efforts are guided by an initial "feeling of knowing" that there is or is not something relevant available in memory to recall. If respondents conclude too quickly that they have nothing to report, they may fail to engage in a memory search. The experimental screener was designed to prevent respondents from committing themselves to a "nothing happened" response before screening. Debriefing item 1 (see Table 8.3, below) was intended to measure respondents' expectation, before screening started, that they would have something to report.

Recall interference Recall of one incident may block retrieval of additional incidents, because a respondent in effect keeps recalling the same incident (Roediger and Neely, 1982). Item 2 was intended to measure whether respondents experienced interference from a previously recalled incident.

Fatigue and negative response set Respondents may become annoyed and fatigued by a long list of screening questions to which the answer is almost invariably "no." The redesigned screener employed short cues and reminders rather than questions to try to avoid these problems. Items 3 and 4 measured the subjective burden of the alternative formats.

Mnemonic failure Because many crime experiences are not salient events in memory, their recall requires contextual cues to aid in retrieval. The new screener employed extensive cues, including reminders of non- stereotypical crimes, to improve the mnemonic properties of the screener and improve respondents' understanding of the crime scope of the survey[4]. Item 5 asked whether respondents were reminded of types of crime they hadn't thought of.

Selective reporting The new screener adopted a "broad net" approach (Biderman 1980b), encouraging respondents to report, and interviewers to accept, all incidents they thought of, even if they were unsure whether the incidents were covered by the survey. (Answers to followup

[4] To test understanding of the crime scope of the survey, respondents were read vignettes and asked "whether you think it is the kind of crime we are interested in, in this survey." Six vignettes portrayed situations vulnerable to misreporting, such as domestic violence ("Jean and her husband got into an argument. He slapped her hard across the face and chipped her tooth. Do you think we would want Jean to mention this incident to us when we asked her about crimes that happened to her?").

questions determined if an incident was in-scope or not.) Item 6 provided a measure of whether respondents were withholding information about potentially in-scope incidents. By telling respondents in advance they would not be asked to disclose anything about unreported incidents, we hoped to increase their willingness to acknowledge them. Respondents were also probed for reasons why an incident was not mentioned.

Table 8.3. Results of Debriefing Questions in an Experimental Comparison of Two Victimization Screener Designs

	Percent "yes"	
Debriefing question	Experimental (short cues) screener	Traditional (question-and-answer) design
1. At the beginning, before I asked you any questions, did you think you would have any crimes to report?	18%	15%
2. I (asked questions/gave examples) to help you remember crimes that might have happened to you. You told me about one incident. Did you find you were still thinking about that incident when we went back to the (examples/questions)? (Asked of respondents who reported one or more incidents)	57%	71%
3. While I was asking you the (questions about/examples of) crimes that might have happened to you, did you lose track or have a hard time concentrating?	12%	8%
4. Did you feel bored or impatient?	30%	25%
5. Were you reminded of types of crime you hadn't already thought of on your own?	41%	26%
6. Was there an incident you thought of that you didn't mention during the interview? I don't need details.	8.6%	4.6%
If yes: "Were you unsure whether it was the type of crime covered by the survey?"	20%	17%
"Was the incident a sensitive or embarrassing one?"	11%	33%
N	522	534

Source: Martin, Groves, Matlin, and Miller (1986).

The results suggested that respondents were willing and apparently able to report their subjective reactions and thought processes during screening. The cognitive properties of the two instruments differed substantially by some measures, and the differences were consistent with objective evidence about screener performance. Item 1 suggested that few respondents (less than 20% in both screeners) said they initially expected to have anything to report, which seemed to confirm the designers' concerns about possible failure of metamemory. (As it turned out, 37 and 48% of respondents in the traditional and experimental screeners, respectively, reported at least one incident.) According to item 2, experimental respondents were significantly less likely to persevere in thinking about an incident after the interviewer returned to the screening task, indicating less recall interference. This was consistent with the much higher rates of victimization reporting, especially of multiple incidents, produced by the experimental screener. However, most respondents in both screeners experienced interference from incidents previously recalled. Items 3 and 4 suggested the experimental screener was more cognitively burdensome, with experimental respondents significantly more likely to say they had a hard time concentrating or lost track during screening. This was consistent with its greater length (32 minutes, compared to 21 minutes for the traditional instrument). On the other hand, experimental respondents were much more likely to report being reminded of types of crime they hadn't thought of (item 5), suggesting the experimental screener was a more effective mnemonic aid, and consistent with this they reported many more victimizations. Finally, results of item 6 indicated less selective reporting in the new instrument, with fewer respondents reporting they thought of an incident but failed to mention it. This was consistent with the elevated reporting of both in-scope and out-of-scope incidents. Although fewer reports were withheld in the new instrument, a significantly larger fraction them pertained to "sensitive or embarrassing" incidents. Thus, the estimated fraction of respondents who withheld reports of sensitive incidents did not differ significantly between questionnaires.

Thus, debriefing questions derived from hypotheses about the cognitive sources of response errors appeared to yield meaningful information about the retrieval and reporting process. Although the results are suggestive, they do not demonstrate that the questions represent valid measures of the intermediate cognitive processes that were the intended target of the redesigned screening procedures. (For example, one might doubt whether respondents could accurately recall their expectations at the start of an interview, as item 1 in Table 8.3 asked them to do.)

Debriefing questions about respondents' mental processes have a natural connection with research on metacognition, or peoples' knowledge of (and, presumably, ability to report about) their own memories and cognitive processes (see, e.g., Koriat, Goldsmith, and Pansky, 2000, for a recent review). Cognitive psychologists have explored certain tasks (e.g., feeling-of-knowing and confidence judgments) similar to debriefing questions that are (or might be) asked of respondents. The experimental literature on metacognition may shed light on the validity of self-reports about cognitive processes, and help answer the question of what respondents can report about their cognitive states, especially their memory processes. Although a review of this literature is outside the scope of this chapter, it might be applied usefully to the design of respondent debriefing questions and other survey topics.

8.3.3. Direct Measures of Missed or Misreported Information

A third type of debriefing question probes for events or facts that a respondent failed to report or reported incorrectly in the main survey. In some cases, detailed debriefing questions may test whether questions in the main survey are eliciting reports consistent with survey definitions (e.g., Esposito, 2002; Fowler and Roman, 1992). For example, to evaluate how well a general question about income is performing, debriefing questions may probe whether a respondent reported net or gross income, and whether specific sources were included. If the question in the main survey is not obtaining the intended information, then the question wording or response categories might be revised, or a question added. Sometimes the debriefing question itself may be moved into the survey to improve measurement.

This type of debriefing question may also be used as a direct measure of underreporting. In the final split-panel comparison of new and old CPS instruments, a subsample of respondents who had not reported any work activities in the main survey was probed to determine if they had neglected to report a few hours of work (Martin and Polivka, 1995). The probe was: "In addition to people who have regular jobs, we are also interested in people who may work only a few hours per week. *Last week* did NAME do any work at all, even for as little as one hour?"

Followup probes ("What kind of work did NAME do?" and "Did NAME get paid for the work?") were asked to screen out reports that were not legitimate work activities (about 80% of responses were bona fide). Between 2 and 3% of respondents who had not reported working in the survey did report bona fide work activities in debriefing, with no overall questionnaire difference. The age bias in underreporting casual labor was reduced but not eliminated in the new questionnaire. A problem with the missed work probe was that large samples were required to detect meaningful differences, so most of the questionnaire differences were not statistically reliable.

Item 6 in Table 8.3 also permitted direct examination of victimization underreporting, although it had the disadvantage that the characteristics of the unreported incidents were unknown. Because it was unknown whether the unreported incidents were in-scope, the information could not be used to estimate the fraction of in-scope victimizations missed.

8.4. CONCLUSIONS

In the past, it has been necessary to approach questionnaire design and revision as a process of redesigning a "black box" whose output is evaluated but whose inner workings are poorly understood and often produce puzzling results. Respondent debriefing and vignettes do not eliminate all the surprises involved in questionnaire design and pretesting, but they can help a designer better understand and predict the nature and underpinnings of questionnaire effects. By shedding empirical light on the inner workings of a survey instrument, these methods help demystify the questionnaire design process and take us a step toward placing the design of survey measurements on a firmer (dare I say scientific?) footing. Below I summarize the advantages and disadvantages of vignettes and respondent debriefing questions and compare them with some other questionnaire pretesting methods.

An advantage of both vignettes and respondent debriefing questions is that they reveal hidden problems of meaning that respondents and interviewers may be unaware of, and that do not necessarily result in interviewing difficulties. This advantage is shared by cognitive interviewing, but not by pretesting methods that do not probe respondents' interpretations, such as behavior coding. A second advantage is that respondents appear able to step into the informant role and perform the tasks, and even appear willing to disclose their less-than-complete reporting of sensitive facts, as in the case of the NCVS debriefing question.

The methods are flexible and may be used in exploratory, qualitative studies, laboratory investigations and experiments, small field pretests, large-scale pilot studies, ongoing surveys, and split-panel field experiments. Indeed, the same set of debriefing and vignette questions may be carried forward from one stage of pretesting to the next to provide systematic, comparable measures at each stage. Behavior coding can be applied in both field and laboratory settings, but cognitive interviewing probably cannot be.

The methods yield useful information even when administered to relatively small samples. Their efficiency is increased because a respondent does not need to have experienced a specific situation in order to interpret how it should be reported. In contrast, very large samples are needed to measure actual reporting differences, especially for uncommon situations. Large samples also may be required for debriefing questions about actual events or behavior, such as missed work.

The methods are cheap when administered as supplements to an ongoing survey or a field pretest, because they involve no separate field contact. Most of the debriefing interviews (including vignettes) reported in this paper took no more than 5-8 minutes of interviewing time. This cost advantage may be shared by cognitive interviewing, which can provide useful results with a small number of interviews, but not by behavior coding, which requires labor intensive coding.

When administered as a survey supplement, both methods preserve contextual validity because respondents are asked to interpret a term or concept, classify a vignette, and so on, in the context of the actual survey. This is less true of laboratory methods, such as cognitive interviewing.

When administered to probability samples, results are generalizable to the survey population, and group differences in question interpretation may be assessed meaningfully. Other pretesting methods, including cognitive interviewing and (in applications to date) behavior coding, are not generalizable because they are not sample-based.

Respondent debriefing questions and vignettes also share several disadvantages. With the exception of probes for missed or misreported information, they do not provide direct evidence of reporting error. They provide indirect evidence about questionnaire performance that is useful in conjunction with other performance indicators, such as reporting differences. They are most useful when their design is informed by substantive and methodological knowledge and theory. In the examples discussed in this chapter, advance knowledge was provided by prior ethnographic interviews, as in Gerber's research, by prior investigations of reporting problems, as in the CPS example, and by hypotheses derived from cognitive literature on recall, as in the NCVS example. In contrast, cognitive interviewing provides more opportunity for an interviewer to design probes flexibly and explore problems that emerge during an interview.

Although their results are plausible and consistent with other evidence, respondent debriefing and vignettes have not been rigorously evaluated. Evidence to date suggests that vignettes are sensitive and relatively robust measures of the context and wording effects of the particular questionnaire they follow, but more research is needed to evaluate this key assumption. And, although it is reasonable to assume that misinterpretations (as measured by responses to debriefing questions or vignettes) are indicative of measurement error, the connection is indirect rather than direct, and needs additional investigation.

Issues involved in the design and implementation of vignettes and debriefings need further exploration. Evidence suggests that some types of debriefing questions are more meaningful and valid than others, and both vignettes and debriefing questions are likely affected by their own sources of error and bias. Research to date suggests that questions about respondent certainty or confidence elicit exaggerated reports that shed little light on question misunderstandings, because respondents seldom doubt their own idiosyncratic interpretations. Responses appear to be sensitive to vignette order, suggesting the need for careful replication when comparisons are made between surveys. The validity of debriefing questions about mental processes during an interview is subject to the same limitations that self-reports about cognitive processes are generally subject to, and this type of question has not been much explored in pretesting. The effects of delay between a survey question and a retrospective debriefing question have not been addressed, but ephemeral thoughts or reactions are likely to be quickly forgotten. Subjective reactions and interpretations may also be affected by questions that intervene between target questions and debriefing questions or vignettes.

Each method also offers some advantages not shared by the other. Because vignettes are posed in hypothetical terms removed from a respondent's own situation, this method is well suited for exploring sensitive or stigmatizing subjects (this type of application was not illustrated here). Vignettes offer advantages for exploring how respondents arrive at complex judgments that are influenced by social context, because situational factors can be varied among vignettes in qualitative or experimental research. Probes to determine why respondents classify vignettes as they do shed light on their reasoning. Vignettes can be used to examine particular problematic situations, test the match between respondents' interpretations and survey definitions, and assess the degree of standardization of meaning. Administering them to interviewers helps identify concepts and situations that are poorly understood and require additional training.

Respondent debriefing questions are more direct measures than vignettes and can provide information about a greater range of response problems, including direct measures of reporting error, although this requires larger samples than are needed for comprehension probes. Vignettes and respondent debriefing questions are useful in three phases of questionnaire development and pretesting. First, vignettes are useful for exploring respondents' understandings of terms and concepts before designing a questionnaire, and can help designers design better questions and avoid wording pitfalls. A second use is to identify or verify problems of interpretation of existing questions in a survey. The statistical modeling of vignette responses may yield insights about changes or inconsistencies in underlying survey concepts, marrying methodological and substantive purposes. Respondent debriefing questions can elicit information about subjective reactions to the questions and feedback about unreported or misreported information. Information about which words and phrases are misunderstood and which types of situations are misreported can help the questionnaire designer address the problems by rewording questions, adding instructions or examples, and so on. (Alternatively, a designer might revise survey definitions to bring them in line with respondents' interpretations.)

In a third phase, vignettes and debriefing questions can be used to evaluate alternative questionnaires. Performance measures based on vignettes and debriefing questions can be used (along with actual reporting differences) to select the questionnaire which is best understood, least cognitively burdensome, or which yields other measurable improvements. By using these methods through a program of iterative design and pretesting, it is possible to gain much richer knowledge about the performance of questions and the nature of the errors affecting survey measurement of a phenomenon.

ACKNOWLEDGMENTS

This chapter reports the results of research and analysis undertaken by U. S. Census Bureau staff. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessary those of the U. S. Census Bureau. I am grateful to the colleagues and friends who were collaborators on some of the research reported here, including Al Biderman (now deceased), Pam Campanelli, Bob Fay, Bob Groves, Jennifer Hess, Jay Matlin, Anne Polivka, Stanley Presser, Jennifer Rothgeb, Donny Rothwell (now deceased), and Paul Siegel.

REFERENCES

Bates, N. 1992, *The Simplified Questionnaire Test (SQT): Results from the Debriefing Interviews*. Census Bureau. Unpublished report, August 18.

Belson, W. A. 1981, *The Design and Understanding of Survey Questions*. London: Gower.

Biderman, A. D. 1980a, *Report of the Workshop on Applying Cognitive Psychology to Recall Problems of the National Crime Survey*. Washington DC, Sept. 17-18, 1980.

Biderman, A. D. 1980b, "Crime-circumscribed versus Broader-net Screening Approaches." Item 361 in Crime Survey Research Consortium Teleconference, 4 June 1981.

Biderman, A. D. 1981a, "Cue Specificity, Time Reference, Mnemonics, and Semantics." Item 581 in Crime Survey Research Consortium Teleconference, 7 Aug. 1980.

Biderman, A. D. 1981b, "Further Reflections on the Recollection Problem." Item 633 in Crime Survey Research Consortium Teleconference, 21 Sept. 1981.

Biderman, A. D., Cantor, C., Lynch, J. P., and Martin, E. 1986, *Final Report of the National Crime Survey Redesign Program*. Washington DC: Bureau of Social Science Research.

Blair, E. and Burton, S. 1987, "Cognitive Processes Used by Survey Respondents to Answer Behavioral Frequency Questions," *Journal of Consumer Research* 14: 280-288.

Campanelli, P.C., Martin, E. A., and Creighton, K. 1989, "Respondents' Understanding of Labor Force Concepts: Insights from Debriefing Studies." *Proceedings* of the Fifth Annual Research Conference. Washington, DC: Bureau of the Census.

Campanelli, P., Martin, E. A., and Rothgeb, J. M. 1991, "The Use of Respondent and Interviewer Debriefing Studies as a Way to Study Response Error in Survey Data." *The Statistician* 40: 253-264.

Campanelli, P. C., Rothgeb, J. M., and Martin, E. A. 1989, "The Role of Respondent Comprehension and Interviewer Knowledge in CPS Labor Force Classification." *Proceedings* of the Section on Survey Research Methods, American Statistical Association.

Cantril, H. and Fried, E. 1944, "The Meaning of Questions." In Cantril, H. et al. (eds.) *Gauging Public Opinion*. Princeton: Princeton University Press.

DeMaio, T. J. 1983a, *Approaches to Developing Questionnaires: Statistical Policy Working Paper 10*. Subcommittee on Questionnaire Design, Federal Committee on Statistical Methodology. Washington DC: Office of Management and Budget.

DeMaio, T. J. 1983b, *Results of the 1980 Applied Behavior Analysis Survey or What People Do With Their Census Forms*. Preliminary Evaluation Results Memorandum No. 61. Census Bureau. October 26, 1983.

DeMaio, T. J. and Rothgeb, J. M. 1996, "Cognitive Interviewing Techniques: In the Lab and in the Field." In N. Schwarz and S. Sudman (eds.) *Answering Questions*. San Francisco: Jossey Bass.

Duncan, O. D. 1984, "Rasch Measurement in Survey Research: Further Examples and Discussion." In C. F. Turner and E. Martin (eds.), *Surveying Subjective Phenomena*, Vol. 2. New York: Russell Sage Foundation.

Esposito, J. 2002, "Iterative, Multiple-Method Questionnaire Evaluation Research: A Case Study." Paper presented at the International Conference on Questionnaire Development, Evaluation, and Testing, Charleston, SC, November 15-17, 2002.

Esposito, J. L., Rothgeb, J., Polivka, A. E., Hess, J., and Campanelli, P. C. 1992, "Methodologies for Evaluating Survey Questions: Some Lessons from the Redesign of the Current Population Survey," Paper presented at the International Conference on Social Science Methodology, Trento, Italy, June 22-26, 1992.

Finch, J. 1987, "Research Note: The Vignette Technique in Survey Research." *Sociology* 21:105-114.

Fowler, F. J. 1992, "How Unclear Terms Affect Survey Data." *Public Opinion Quarterly* 56:218-31.

Fowler, F. J., and Roman, A. M. 1992, *A Study of Approaches to Survey Question Evaluation.* Report produced for the U. S. Census Bureau, Feb. 25, 1992.

Fracasso, M. P. 1989, "Reliability and Validity of Response Categories for Open-Ended Questions in the Current Population Survey," *Proceedings* of the Section on Survey Research Methods, American Statistical Association.

Gerber, E. 1990, *Calculating Residence: A Cognitive Approach to Household Membership Judgements among Low Income Blacks*. Unpublished Census Bureau report.

Gerber, E. 1994, *The Language of Residence: Respondent Understandings and Census Rules. Final Report of the Cognitive Study of Living Situations*. Center for Survey Methods Research, U. S. Census Bureau.

Hart, J. T. 1965, "Memory and the Feeling-of-knowing Experience." *Journal of Educational Psychology* 56:208-216.

Hess, J., Rothgeb, J., Zukerberg, A., Richter, K., Le Nenestrel, S., Moore, K., and Terry, E. 1998, "Teens Talk: Are Adolescents Willing and Able to Answer Survey Questions?" *Proceedings* of the Section on Survey Research Methods, American Statistical Association.

Hess, J. and Singer, E. 1995, "The Role of Respondent Debriefing Questions in Questionnaire Development." *Proceedings* of the Section on Survey Research Methods, American Statistical Association: 1075-1080.

Koriat, A., Goldsmith, M., and Pansky, A. 2000, "Toward a Psychology of Memory Accuracy." *Annual Review of Psychology* 51: 481-537.

Loftus, E. F. and Marburger, W. 1983, “Since the Eruption of Mt. St. Helens, Has Anyone Beaten You Up? Improving the Accuracy of Retrospective Reports with Landmark Events,” *Memory and Cognition* 114-120.

Martin, E. A. 1987, "Some Conceptual Problems in the Current Population Survey." *Proceedings* of the Section on Survey Research Methods, American Statistical Association.

Martin, E. 2001, “Privacy Concerns and the Census Long Form: Some Evidence from Census 2000,” *Proceedings* of the Section on Survey Research Methods, American Statistical Association.

Martin, E. A., Campanelli, P.C., and Fay, R.E. 1991, "An Application of Rasch Analysis to Questionnaire Design: Using Vignettes to Study the Meaning of 'Work' in the Current Population Survey." *The Statistician* 40:265-276.

Martin, E. A., Groves, R., Matlin, J., and Miller, C. 1986, *Report on the Development of Alternative Screening Procedures for the National Crime Survey*. Washington DC: Bureau of Social Science Research.

Martin, E. A., Hess, J., and Siegel, P.M. 1993, "An Empirical Examination of the Meaning of Work." Unpublished paper, Census Bureau, November 6, 1993.

Martin, E. A., Hess, J., and Siegel, P.M. 1995, "Some Effects of Gender on the Meaning of ‘Work': An Empirical Examination." In R. L. and I. H. Simpson (eds.) *Research in the Sociology of Work, Vol. 5*. Greenwich Conn.: JAI Press.

Martin, E. and Polivka, A. E. 1995, "Diagnostics for Redesigning Questionnaires: Measuring Work in the Current Population Survey." *Public Opinion Quarterly* 59: 547-567.

Miller, E. R. and Davis, W. L. 1994, *Findings from the Cognitive and Field Interview Research on Questions about "Proof of Paternity."* Unpublished report. Census Bureau, March 18, 1994.

Morton-Williams, J. and Sykes, W. 1984, "The Use of Interaction Coding and Follow-Up Interviews to Investigate Comprehension of Survey Questions." *Journal of Market Research Society* 2:109-127.

Moyer, L. H., Fansler, N. E., Lee, M. A., and Von Thurn, D. 1997, "How Do People Answer Income Questions?" *Proceedings* of the Section on Survey Research Methods, American Statistical Association.

Oksenberg, L., Cannell, C., and Kalton, G. 1991, "New Strategies for Pretesting Survey Questions*."* *Journal of Official Statistics* 7: 349-365.

Piaget, J. 1932/1965, *The Moral Judgment of the Child.* New York: The Free Press.

Polivka, A. E. and Rothgeb, J. M. 1993, "Redesigning the Questionnaire for the Current Population Survey," paper prepared for presentation at the annual meeting of the American Economics Association, Anaheim, CA, January 1993.

Rasch, G. 1960/1980, *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Roediger, H. L. and Neely, J. H. 1982, "Retrieval Blocks in Episodic and Semantic Memory." *Canadian Journal of Psychology* 36:213-242.

Rossi, P. H. and Anderson, A. B. 1982, "The Factorial Survey Approach: An Introduction," Chapter 1 in Rossi, P. H., and Nock, S. L. (eds.), *Measuring Social Judgments: The Factorial Survey Approach*. Beverly Hills CA: Sage Publications.

Rossi, P. H., Waite, E., Bose, C. E., and Berk, R. E. 1974, "The Seriousness of Crimes: Normative Structure and Individual Differences," *American Sociological Review* 39: 224-237.

Schuman, H. 1966, "The Random Probe: A Technique for Evaluating the Validity of Closed Questions." *American Sociological Review* 31:218-222.

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

Sparks, R. 1982, "First Shot at a 'Short-Cue' Screener." Item 794 in Crime Survey Research Consortium Teleconference, 10 May 1982.

Von Thurn, D. 1996, *Report of the October 1995 CPS School Enrollment Debriefing Questionnaire.* Memorandum from D. Von Thurn to J. Day, Nov. 5, 1996. U. S. Census Bureau.

Wobus, P. and de la Puente, M. 1995, "Results from Telephone Debriefing Interviews: The Census Bureau's Spanish Forms Availability Test." *Proceedings* of the Section on Survey Research Methods, American Statistical Association: 1040-1045.