WHEN SHOULD WE ASK FOLLOW-UP QUESTIONS ABOUT ITEMS IN LISTS?[1]

John Bosley, Monica Dashen and Jean E. Fox
Bureau of Labor Statistics
John Bosley, Bureau of Labor Statistics, Room 4915, 2 Massachusetts Ave., N.E., Washington, DC 20212

Key Words: Response Error, Survey Methodology, and Respondent Burden.

INTRODUCTION

Asking screening questions, with follow-up questioning only when that is appropriate, is a widely used technique for increasing survey efficiency. The response to a screener question establishes whether or not a particular respondent has some attribute, or meets some condition, that makes it worthwhile or appropriate to ask additional follow-up questions about a particular topic. By excluding unqualified respondents from this additional questioning, survey length and resulting burden can be markedly decreased.

While screening for respondent qualification can clearly enhance efficiency, there may be some costs in data quality associated with the burden imposed by the follow-up questioning. Survey methodologists such as Fowler (1993) and Sudman and Bradburn (1982) have observed that respondents can easily recognize that they are asked these additional questions only if they answer the screener questions in a particular way. (The experts cited do not offer any empirical evidence that this negative response effect actually occurs; their observations are hypothetical--and plausible.) In this view, the data provided by follow-up questions may not be complete and accurate if respondents choose to answer screener questions so that follow-up is reduced. If this response strategy is adopted due to factors such as fatigue or boredom, data quality will suffer to an unknown extent.

These same experts suggest a way to reduce this hypothesized restriction of reporting, which we define as respondents' limiting the extent of follow-up questioning by using the screener questions to disqualify themselves from follow-up. They propose asking groups of screener questions before asking qualified respondents any follow-up questions for specific questions in that group. This may somewhat disguise any clear and direct association between screener response and the presence or absence of follow-up questions.

What do logic and previous research suggest as plausible conditions under which restriction of reporting should or should not occur? It seems likely that restriction of response should not occur in a survey with relatively few screening and follow-up questions. Many surveys contain scores or even hundreds of screeners with follow-up questions, and hence would be likely to induce reporting restriction. So in many cases the risk of reporting restriction may be substantial indeed.

The nature of the screeners and follow-up questions may also affect subsequent reporting restriction. If the screeners pertain to everyday, non-sensitive topics, respondents may be less likely to restrict reporting. Sensitive, personal questions may on the other hand induce much more reporting restriction.

A number of past studies found some empirical evidence that reporting restriction occurs, at least in some circumstances. Lehnen and Reiss (1978) found that the repetition of entire interviews over time, a panel survey, led to a decrease in crime incident reporting in successive waves of the National Crime Survey (NCS). They labeled this response effect "restriction of reporting," a label used in this paper for the same effect. The NCS used a screen-and-follow up approach to get details about reported crime victimization incidents. The qualifying event sought in the screener was a particular incident of crime victimization, and a positive response led to extensive follow-up questioning about that incident. Lehnen and Reiss (1978) found restriction of reporting effects across successive administrations of the NCS survey at six-month intervals. The NCS addressed a sensitive issue—crime victimization.

What does the literature say about reporting restriction within a single survey session? Recent research by Jensen and his associates (Jensen, Watanabe, & Richters, in press) employed a clever design to study within-interview fall off in reporting in a clinical-diagnostic context. They administered two modules of the Diagnostic Interview Schedule for Children, Version 2 (Shaffer, Fisher, Dulcan, Davies, Piacentini, Schwab-Stone, Lahey, Bourdon, Jensen, Bird, Canino, and Regier, 1996) to 88 pairs of parents and one of child of each pair. Half of each sample responded to an Attention Deficit/Hyperactivity

[1] We thank Scott Fricker, Charman Hayes, Gary Mitchell, and Bill Mockovak for help with various aspects of this project. The opinions expressed are those of the authors and do not necessarily reflect the views of the Bureau of Labor Statistics.

Disorder (ADHD) module first, followed by a depression/dysthymia module. In the other half of each sample, the module order was reversed. This clinical diagnostic tool has three levels of inquiry—broad categorical questions, extensive detailed follow-up questions, and additional follow-up in cases of serious evidence of psychiatric disorder. Thus there are three possible measures that may show an order effect.

To briefly summarize an extensive analysis of their data by Jensen and his colleagues, they found evidence for what they term "diagnostic attenuation," an effect equivalent to our reporting restriction, as respondents were led deeper into the series of follow-up questions. The clearest indication of reporting restriction occurred in the few cases where follow-up questioning was most extensive; that is, for those cases where evidence of psychiatric disorder was clearest. In connection with their overall results, Jensen et al. state that they "estimate that order effects may account for fluctuations by a factor of one half or more around prevalence estimate base rates, simply by manipulation of the order of diagnostic modules." (Jensen et al. in press, pp. 12-13) In other words, the very rates at which psychiatric disorders are diagnosed in children and young people may be greatly increased or decreased depending on whether a set of questions with the potential of leading to any given diagnosis is asked early or late in a diagnostic interview!

Present Study

The present study was an effort to gather empirical support for this hypothesized reporting restriction effect, within a single administration of a survey interview, unlike the Lehnen and Reiss study which involved multiple interviews. It also had the goal of testing the degree to which the strategy suggested by Fowler (1993) and Sudman and Bradburn (1982) for reducing reporting restriction. That is to say, deferring follow-up questioning until a series of screeners have been asked, reduced any reporting restriction effect. The study compared this "List First" (LF) approach that asked a group of screeners before any follow-up questioning, with an "Item-by-Item" (II) pattern in which each positive screener response was followed immediately by a set of follow-up questions.

This study used a portion of an existing BLS survey in the research. The survey chosen, the Consumer Expenditure (CE) survey, collects information on consumer purchases from a national sample of households. The actual sample unit is a "consumer unit" (CU) which may be the entire household or some sub-unit of its members. The CE survey collects information about CU purchases quarterly for five consecutive quarters, in a panel design. The reference period is one to three months, depending on the serial position of the interview.

For research purposes, we used an abbreviated version of the CE in order to keep experimental sessions within practical boundaries. In order to have a within-interview basis of detecting reporting restriction, we collected validation data about consumer purchases in a self-report format before the simulated interviews, which asked about the same purchase categories.

The present study differs from Lehnen and Reiss (1978) and Jensen et al. (in press) by asking about consumer purchases, which for the most part are not sensitive topics. Another difference is that our experimental interview was shorter than the others. The entire CE interview can take over two hours to administer, which is why we thought it might trigger reporting restriction. We used only a few sections of the CE interview, spanning relatively few different categories for which screener questions are asked. Thus, this study attempts to see whether reporting restriction sets in relatively quickly with a fairly simple screener and follow-up formatted interview, lasting only about thirty minutes. These factors allow us to begin to understand the occurrence of response restriction in a range of situations, which is of considerable theoretical interest.

A more practical motive for this research was to provide guidance to a team that is converting the CE interview to a computer-based form. This team needed to know whether a LF format would be a better format for the CE interview than the currently used II format.

Experimental Overview

The study design sought to detect a reporting restriction effect in an abbreviated version of the CE interview, and to compare the degree of occurrence in two conditions. The first condition, which mirrors current practices for CE interviews in the field, always asks follow-up questions immediately after the respondent specifies a purchase. We called this the "Item-by-Item" (II) approach. In the second condition, respondents listed purchases of several items (within a category) before any follow-up questions about those purchases were asked. For example the interviewer read the list of all of the individual items the CE includes under "infant wear" and asked the respondent to respond with "Yes" to each item purchased, or "No" to any items not purchased. Only after all within-group purchases had been identified would any follow-up questions be asked. We term this approach "List First" (LF). As Fowler and others suggest, this condition was hypothesized to lessen any reporting restriction detected by the study.

To summarize, this study compares two interview approaches on measures of the degree to which each may lead to reporting restriction in an interview using a screener and follow-up format. We compared the two conditions using the following types of measures:

- the relative efficiency of the two approaches for inducing respondents to recall and accurately report the consumer purchases that their consumer unit actually made; we hypothesized that the LF approach would produce greater recall and accuracy;
- indicators of respondent cognitive load ("burden"), such as time to complete the interview and subjective measures such as irritation with the interview approach, boredom, inattention or other signs of fatigue, and the like. We hypothesized that such measures of cognitive effort would be lower for the LF than the II condition.

METHOD

Participants

Twenty-four participants (ten males and fourteen females) responded to an advertisement in a local newspaper and received $25.00 each in compensation for their participation. The participants' mean age was forty-seven, and their average educational level was sixteen years of schooling (or a college degree).

Procedure

There were three phases in this study. [2]

Phase 1: Collection of Item Purchases to Validate Responses in the Interview.

In this phase, all participants were asked to complete a recall task, which was designed to collect information on the participant's consumer unit's expenses. Participants were asked to report purchases of specific items during a three-month reference period within the following broad sections taken from the actual CE interview: Major Appliances, Smaller Household Appliances, Furniture and Housewares, (Adult) Clothing, and Infant Clothing.

Many of the purchases these sections cover are infrequently purchased, and since we needed enough baseline reports of purchases to permit the reporting restriction effect to emerge, we chose to augment these sections with some categories of frequently purchased items. These categories were (1) Books, (2) Gardening & Lawn Services, and (3) Automobile Repair & Maintenance (e.g., oil change).

As noted, participants were asked to record all their purchases for a three-month reference period, for each of the categories listed above. These tests were self-administered, using forms that cued recall at the individual item purchase level. For example, within Major Appliances, cues included "Electric Stove or Oven," "Refrigerator," etc. These cues were exactly the same as the item lists given in Phase 2. Participants were allowed as much time as needed to recall all their purchases related to the specific item cues provided. They were also asked to provide two items of identifying information, a brief description of the purchase such as a brand name, and the total cost. Collecting this supplementary information was intended to increase the certainty with which we could match baseline recall with interview report data. This phase was intended to collect accurate and exhaustive purchase information as a baseline.

Respondents were then scheduled to return three days later "for additional paper work." Respondents were not told anything of the nature of the next visit, to minimize their motivation to rehearse the purchases they had just recorded.

Phase 2: Interview Conditions.

At their second visit, respondents were randomly assigned to one of two conditions: (1) Item-by-Item (II) or (2) List First (LF). Interviewers asked respondents in both conditions to report all their purchases in the same reference period and same categories used in Phase 1. The interviewers used the actual CE paper-and-pencil survey forms and instructions to guide these interviews.

The difference between the Item-by-Item and List First conditions was the timing of the follow-up questions for each reported purchase. The Item-by-Item group was asked about the follow-up questions as soon as they reported an item purchased by someone in their consumer unit. In contrast, respondents in the List First condition were asked the follow-up questions only after having been asked about whether or not they had made any purchases within a group. (Note: The follow-up questions asked for information such as month of purchase, price, inclusion or exclusion of sales tax in price given, quantity (number) of items purchased, and for whom the item was purchased.) The mean size of the item group was four items. The group size varied from three to five items, according to how the current specifications for the CE Computer Assisted Personal Interview (CAPI) conversion allocate items to individual computer screens. It is important to note that this group was asked about the particulars after they had responded, either affirmatively or negatively, to a "computer screenful" of purchases.

The principal dependent measure of interest is answer conversions, which occur when a respondent changes his or her answer from "Yes, I made the purchase" to "No, I did not make the purchase" between the Phase 1 reports and the subsequent interview. Such changes indicate false denials and support the

[2] Phase 1 was administered first. Approximately three days later, Phases 2 and 3 were both administered one immediately after the other. Phase 1 occurred in Session 1 and Phases 2 and 3 occurred in Session 2.

conjecture that the respondent is omitting mention of a purchase in order to avoid being subjected to additional follow-up questions.

While the respondents were interviewed, we collected data on verbal and non-verbal behavior, which indicates the burden each interview places on the respondent. These additional measures are intended to augment the principal measure -- answer conversion. For example, the number and types of complaints about the repetitiveness of the questions or time taken to complete the interview would be verbal indicators of interest. In addition, yawning, checking the clock or other less obvious signs of inattentiveness (e.g., such as failure to answer the question and thus requiring the question to be asked more than once) would also be non-verbal behaviors of interest. In summary, the verbal and non-verbal measures work together to assess the burden each interview places on the respondent.

Phase 3: Respondent Debriefing and Ratings Task

Immediately following the interview, respondents performed a rating task to assess the perceived task difficulty of being interviewed. For example, people were asked to rate on a five-point scale how difficult it was for them to recall item detail. Two rating scales asked directly about subjective feelings of boredom and interest. These questions were: (1) "During this interview it was hard to stay interested and pay attention (where '1' is never hard to stay interested and '5' is always hard to stay interested)," and (2) "I was bored listening to the interviewer's questions and giving my answers (where '1' is never bored and '5' is always bored)."

An additional rating scale was used to assess how frustrated people were by the task at hand. Respondents were asked the following question: "Looking back over the whole interview, I would rate my feelings about it, (where '1' is frustrating and '5' is satisfying)."

RESULTS

The first section below describes the likelihood of reporting restriction occurring in the Item-by-Item and List First conditions. The second section describes indicators of respondent burden (i.e., interview time, ratings, and observable behaviors).

Reporting Restriction

We first wanted to determine whether respondents restricted their reporting in Session 2. The measure we used was the percent of items reported in Session 1, but not in Session 2. To do this, we first compared the purchases reported in Sessions 1 and 2. All items (from both sessions) were classified into one of the following three categories: (1) reported in Session 1 only, (2) reported in both Sessions or; (3) reported in Session 2 only.

The mean number of items reported for each condition are shown in Tables 1 and 2 for Sessions 1 and 2 respectively. The column labeled "Both Sessions" represents the items reported in both sessions, and is, therefore, the same in Tables 1 and 2. The total number of items reported in Session 1 equals the number reported in Session 1 only plus those reported in both sessions, and likewise for items in Session 2.

Table 1. Mean number of items reported in Session 1

Condition	Session 1 Only	Both Sessions	Total Session 1
Item-by-Item	5.9	11.9	17.8
List First	6.5	13.1	19.6

Table 2. Mean number of items reported in Session 2

Condition	Session 2 Only	Both Sessions	Total Session 2
Item-by-Item	6.5	11.9	18.4
List First	6.4	13.1	19.5

We compared the total number of items reported in each condition across sessions. There were no differences in the number of items reported between the two sessions ($F(1, 21) = 0.04$, $p = 0.85$). There were also no differences in the number of items reported in each condition ($F(1, 21) = 0.11$, $p = 0.75$). Further, there was no interaction between condition and session $F(1, 21) = 0.10$, $p = 0.76$). Neither the condition nor the session affected the number of items reported.

We then calculated the percentage of items reported in Session 1 but not in Session 2. For example, for the Item-by-Item condition, it would be (Session 1 Only)/(Total Session 1) or 5.9/17.8 = 31%. There were no differences in the percentages for the two conditions ($t(21) = 0.66$, $p = 0.52$). Thus, the method used to administer the survey did not induce reporting restriction, as measured by the percentage of items left off of the survey in Session 2 (31% for Item-by-Item, 35% for List First).

Respondent Burden

We also looked at several measures of respondent burden. First, we evaluated the more objective measure of interview time (for Session 2). Second, we considered the subjective measures of the ratings from Phase 3 and of the frequencies of behaviors indicative of boredom.

For session time, we calculated both the total time for Session 2 and the time per item (total time

divided by the number of items). There was no difference in the total interview time across the two conditions (t (20) = 0.65, p = 0.52). The mean time was 27.2 minutes for Item-by-Item and 24.3 minutes for List First. There was also no difference in the time per item (t (19) = 0.61, p = 0.55). The mean time per item was 1.5 minutes for Item-by-Item and 2.0 minutes for List First. Thus, the interview condition did not affect the interview time.

The first subjective measures of burden are the ratings the respondents provided at the end of the interview. The means for the relevant ratings are shown in Table 3 below.

Table 3. Mean boredom, interest and boredom ratings

Condition	Boredom	Sustained Interest	Frustra-tion
Item-by-Item	1.9	2.0	4.4
List First	1.8	1.8	3.8

There was no difference in boredom between the two conditions (t (22) = 0.27, p = 0.79), and the ratings indicate minimal boredom in both conditions. Those in the Item-by-Item condition were not any more bored than those in the List First condition. This finding does not support our prediction that participants would find the Item-by-Item condition more boring.

There was also no difference between conditions in sustained interest (t (22) = 0.38, p = 0.71). Participants in both conditions were able to stay interested in the survey fairly well. Again, we did not find a difference where we had expected to.

There was a significant difference in ratings of frustration between conditions (t (22) = 2.17, p = 0.04). Those in the List First condition were more frustrated than those in the Item-by-Item condition, but both ratings show minimal frustration. These results contradict our expectations that those in the List First condition would be less frustrated.

In evaluating behaviors, we developed a list of behaviors (e.g., yawning, playing with objects, or verbal comments) that we monitored to assess boredom in Session 2. We recorded the time from the beginning of the interview to the onset of the first behavior. We also counted the total number of behaviors throughout the interview. For the following analyses, we only considered the twenty-two participants who exhibited at least one of the behaviors we were interested in.

There were no differences in the time to the first behavior (t (20) = 0.48, p = 0.63. The mean time was 4.8 minutes for Item-by-Item and 5.7 minutes for List First. Participants in both groups exhibited their first behavior approximately one-fifth of the way through the interview. There was also no difference in the number of behaviors between conditions (t (20) = 1.04, p = 0.31). There were on average 12.0 behaviors for Item-by-Item and 8.2 behaviors for List First.

In summary, most of the measures indicate no differences between the two conditions. The only exception was in the ratings of frustration, which indicate that people in the List First condition were more frustrated than those in the Item-by-Item condition.

CONCLUSIONS

The data reported in this study address a theoretical issue central to survey methods: When should survey designers ask follow-up questions about items presented in lists? We constructed two conditions for our study-- Item-by-Item and List First -- to address this question. The principal difference between these two conditions is the timing of the follow-up questions. Four dependent measures were of interest: (1) answer conversions (which occur when a respondent changes his or her answer from "Yes, I made the purchase" to "No, I did not make the purchase."), (2) time taken to complete the interview, (3) ratings, and (4) behaviors. Because one of our measures, answer conversions, required that we know what they actually bought, validation measures were employed.

No evidence supported the conjecture that people in the Item-by-Item condition would consider the interview more burdensome than those people in the List First condition. For one, the failure to find any differences between conditions for answer conversion data suggests that the respondents did not restrict their reports during the interview. (Note: We consider reporting restriction to be indicative of respondent burden.) Secondly, the failure to find any differences between the two conditions (Item-by-Item and List First) for either the time taken to complete an interview or the observable behavioral data (e.g., complaints or yawns) suggests that the respondents did not find either condition to be more burdensome than the other. This finding is consistent with Silberstein and Jacobs' (1989) results about the interview portion of the CE. They examined four interviews, each of which had the same respondents. They found that the level of reports was fairly consistent across each of the quarterly interviews for most sections of the interview. That is to say, people behaved the same way at the end as they did at the beginning. Thus, these results provide further support that people did not restrict their reports during the interviews.

However, there is one type of expenditure where the reporting restriction effect was evident in Silberstein and Jacobs' analysis. The reporting level for clothing did decrease from the first interview to the last.

This finding contradicts the findings in the present work because we did not find any evidence of this reporting restriction effect in the clothing portion of our interview. There are several possible reasons for this inconsistency between studies. The first reason for this inconsistency may be the different methodologies. We examined restricted reporting in the context of a single interview, whereas Silberstein and Jacobs examined it in the context of multiple interviews. An alternative reason for the inconsistency may be in due part to the small sample size of the present work. Perhaps, the people in our sample simply did not buy that much clothing and therefore did not have a need to restrict their reports. If this explanation is indeed the case, a larger survey size might show people restricting their responses.

Taking the findings of the present work and those of Silberstein & Jacobs (1989), the failure to find restricted reporting may be attributed to the fact that the CE may be a relatively easy-to-answer survey. The CE interview may not be as lengthy or tedious as other surveys where the restricted reporting phenomenon is observed (Lehnen & Reiss, 1978; Jensen et al., in press). In addition, the purchase-oriented questions in the CE may not be as sensitive as those crime or mental disorder questions used in other surveys where the restricted-response effect was observed (Lehnen & Reiss, 1978; Jensen et al., in press). Clearly, there is a strong need for additional work in this area to understand further the parameters that lead to restricted reporting.

It is curious to note that the List First group rated the interview as more frustrating than the Item-by-Item group. It should be noted that both rating scores were below average. This finding indicates that while there was a difference between conditions, the respondents did not feel particularly frustrated in either group. Despite the fact that the ratings were fairly positive, the finding that one survey style is substantially more frustrating than the other is interesting and therefore warrants some speculation. The differences in frustration ratings may reflect the respondents' belief that the List First interview format is more frustrating than the Item-by-Item format because the former disrupts the flow of conversation. In fact, interviewer observations indicated that people felt that they had much more to say about a particular item before proceeding to the next item than they were allowed to say in the List First condition. Though this conjecture is pure speculation, it may also be worth pursuing in future work.

In summary, the study found few differences in the measures we used between the Item-by-Item and List First conditions. The length of the interview (about a half-hour) and the nature of the survey (purchases) did not burden the respondents to such a degree that they restricted their reporting. This topic clearly needs more research to identify those factors, which lead to reporting restriction, so we can work to minimize the effects in future surveys.

REFERENCES

Fowler, F.J. (1993). Survey Research Methods. Newbury Park, CA: Sage Publications.

Jensen P.S., Watanabe H.K., and Richters J.R. (in press). Who's on First? Testing for Order Effects in Structured Interviews Using a Counterbalanced Experimental Design. Journal of Abnormal Child Psychology.

Lehnen, R.G. and Reiss, A.J. (1978). Response Effects in the National Crime Survey. Victimology. 3, (1-2) pp.110-160.

Shaffer, D., Fisher, P., Dulcan, M., Davies, J., Piacentini, J., Schwab-Stone, M., Lahey, B., Bourdon, K., Jensen, P., Bird, H., Canino, G., and Regier, D. (1996). The Second Version of the NIMH Diagnostic Interview Schedule for Children (DISC – 2). Journal of the American Academy of Child and Adolescent Psychiatry, 35, 865-877.

Silberstein, A.S. and Jacobs, C. (1989). Symptoms of Repeated Interview Effects in the Consumer Expenditure Interview Survey. In Kasprzyk, Ducan, Kalton & Singh (Eds.). Panel Surveys. pp. 289-303.

Sudman, S. and Bradburn, N. (1982). Asking Questions: A Practical Guide to Questionnaire Design. San Francisco, CA: Jossey-Bass Publishers.