Considerations Regarding the Use of Global Survey Questions

Prepared for the Consumer Expenditures Survey Methods Workshop, December 8-9, 2010

Paul Beatty
National Center for Health Statistics

For many survey purposes, desired data points can be generated by asking single questions. However, for more complex data needs, what we really want to know must be obtained indirectly or through multiple questions. For example, we may be interested in how many total calories respondents consume in a week, but the information it is unlikely to be known in that form. Data of any real validity would most likely be obtained by constructing it from smaller pieces of information, such as reports of foods that were eaten at particular times, or even estimates of calories consumed in shorter time periods.

Yet many behavioral questionnaires take a relatively head-on approach, asking for information that requires very complex recall and judgments, often covering broad subject matter. It is reasonable to ask whether such “global” questions can be adequately answered, even if rough estimates (as opposed to recalled information) would be acceptable for our data needs. Concerns about the accuracy of such responses have led to a number of studies considering whether “decomposition” into simpler questions offers improvements (e.g., Belli, Schwarz, Singer, and Talarico, 2000, and see also Tourangeau, Rips and Rasinski, 2000, pp. 95-96). The reverse consideration, that we should explore the feasibility of asking fewer questions encompassing more conceptual territory, seems to be rarely considered, at least from a data quality perspective. Methodologists are more likely to assume that we are asking too much within single questions to get valid data, than that we are inefficiently spreading our resources too thinly over many questions.

In that sense, the Consumer Expenditure Surveys program is investigating global questions in a different manner than is usually done: rather than asking whether we can improve validity by breaking down large questions, the consideration here is whether we can retain validity by consolidating them. Obviously there are very good reasons for such an investigation: if feasible, global questions have the potential to reduce costs and burden. But would such questions have acceptable levels of accuracy? The answer to this is not straightforward and depends a great deal upon the specific questions asked. However, in the course of this paper, I hope to lay the groundwork for several conclusions:

- Global questions of the sort being considered here are unlikely to yield responses that are literally correct—that is, they will probably be answered through broad estimation strategies. However:
    - The quality of responses to global questions could vary a great deal depending upon the question and the subject matter domain, and
    - Such responses could nevertheless be more accurate than data produced from more specific questions in the same domain—depending upon some key characteristics of those more specific questions and the response processes they invoke
- Some relatively straightforward laboratory investigations could provide a great deal of information about memory and response strategies, which could be informative about the relative advantages of global and specific questions
- Ideally, responses to the questions should be validated by external data, although such validation has methodological complications that should be addressed carefully.

I will lead to these conclusions in a somewhat roundabout way—first by reviewing some research I have done on related topics, then attempting to reconcile it with earlier research that might explain some findings. Then I will outline some steps forward for additional research related to the Consumer Expenditure Surveys.

The initial case against global questions

The potential downside of global questions became obvious in the course of conducting numerous evaluations in the cognitive laboratory, several of which involved questions about food consumption intended for the National Health Interview Survey (partially reported in Beatty, Fowler and Cosenza, 2006). The questions covered several different types of food (e.g., desserts, cereal, and others), but for purposes of illustration let's consider one particular example, regarding cheese consumption:

- During the past 30 days, how many times did you eat cheese, including cheese as snacks, and cheese in sandwiches, burgers, lasagna, pizza, or casseroles? Do not count cream cheese.

In cognitive laboratory testing, many participants freely admitted that their responses were very broad estimates or guesses, and when probed to think further, participants generally remembered additional instances of eating that had not figures into their original responses. Furthermore, breaking the question into components seemed to help: participants were much more confident in their answers when responding regarding particular sub-domains (e.g., cheese on sandwiches, cheese mixed into other foods, and cheese as snacks).

Subsequent split-ballot experiments compared responses to global questions with responses to multiple specific "decomposed" questions, based on sub-domains as described above:

- During the last 30 days, how many times have you eaten cheese on a sandwich, including burgers?
- During the last 30 days, how many times have you eaten cheese in lasagna, pizza, casseroles, or mixed in with other dishes?
- During the last 30 days, how many times have you eaten cheese as a snack or appetizer?

The specific questions matched as closely as possible the words in the global questions, except that they spread the material over multiple questions. Consistent with cognitive interview findings, the more specific questions yielded higher behavioral reports: global questions yielded a mean report of 13.9 times eating cheese, whereas multiple questions covering the same domain yielded reports of 19.0 times eating cheese, a difference that was significant at the 0.01 level (and similar patterns were observed for other sets of global and specific food consumption questions). Presumably, one of the dangers of global questions is that lead to incomplete reports, for several reasons: too much information may be presented at once, overwhelming working memory; and, respondents may spend less time recalling pertinent information. Thus a plausible argument could be made that the global questions produce under-reports, and the specific questions are more accurate.

Furthermore, behavior coding of the interviews revealed that several indicators of problems were more prevalent with the global questions than any of the more specific questions—and that respondent performance improved as they answered more questions (see Table 1, below).

Table 1: Behavior coding of global and specific questions regarding cheese consumption

	Global	Spec1	Spec2	Spec3
Inadequate initial response	15.9	9.9	8.3	3.1
Probes used during administration	13.7	7.8	6.3	2.1
Respondent requested help/repeat of question	19.1	15.1	3.1	2.1

(all expressed as %; most signif at p<.05)

Counter-arguments regarding global questions

However, specific questions have their disadvantages—most prominently time. In the study mentioned above, it took nearly twice as long on average to obtain answers to three specific questions as to one global question. Furthermore, some of the apparent advantages seen in behavior coding proved to be illusory. Whether we are asking one question or three, we are still generating one data point—thus, it might be more appropriate to consider the *cumulative* rates of "undesirable" behaviors needed to arrive at those points. When comparing global and specific questions in that manner, their performances are not significantly different. It takes about the same amount of probing to yield a response in either case.

The critical criterion is accuracy. If the specific questions are significantly closer to reality, *and* that precision is analytically significant, then the increased cost of multiple questions might be justifiable. On the other hand, if the global questions are more accurate, *or* any error they introduce is tolerable, then it would seem to be preferable to enjoy the efficiency that global questions offer. But to make that determination, we need more data on the relative accuracy of each.

Validating data from global and specific questions

In a subsequent study (Beatty and Maitland, 2008) we attempted to provide such validation data through another split-ballot design with a diary component. The experiment used the same food consumption questions, and also included an additional global question on physical activity (the alternative version was split into specific questions about walking, active chores, and more "deliberate" exercise). See the Appendix for the full texts of the questions. One important change from the versions of the questions presented previously was that it was necessary to change the reference period from 30 days to 3 days. Collecting validation data over a 30-day period was simply not possible.

As a first step, participants in the study agreed to complete a 3-day web diary that recorded all food eaten and physical activities done on a daily basis. The diary was a somewhat simplified version of a web diary that had been collaboratively designed by researchers at several federal health agencies, with input from the contractor (SRBI) who carried out data collection. It included examples of the sort of information desired, progress screens, and various prompts. For example, after respondents submitted the food diary for each day, a prompt asked "Have you included all snacks, all side dishes you had with meals, and everything you added to the foods you ate?" After the respondent submitted the physical activity diary a prompt asked, "Have you included any time you walked for at least 10 minutes, any sports or active games you played, and any time you worked out or exercised?" In addition to the prompts, the diary was designed to facilitate recall by asking for entries related to breakfast, lunch, dinner, and snacks.

Respondents who completed the diary on three consecutive days were invited to participate in a follow-up telephone interview. Respondents to this interview received one of two questionnaire versions. Each version contained questions about food consumption (regarding cheese, cereal, pasta and rice, desserts, and oils) and physical activity. The questionnaire versions were constructed so that all respondents received global questions for some domains, and multiple, specific questions for others. To summarize a very complicated process, results from the web diary were carefully coded as a validity check of responses to the global and specific survey questions.

Based on hypotheses from earlier studies, we expected the data to show several things. First, we expected that global questions would generally undercount behaviors relative to the diary reports. Second, we expected that decomposed questions would lead to reports that were closer to diary reports than the global questions. That is, if X represents the diary report, D represents reports from decomposed questions, and G represents reports from global questions, we generally expected reports frequencies in the following pattern:

```
Low freq            G              D      X                        High freq
         |----------------------------------------|------------------------------------|
```

To evaluate the accuracy of the global and decomposed questions, we computed the relative bias for each, defined as (survey response – diary response)/diary response. This provides a standardized measure of the disagreement between diary and survey measures overall. P-values were computed by two-tailed t-tests, exploring the hypothesis that the mean responses from the diary and from the interview are equal.

In the case of the cheese questions, results were basically as expected: global questions strongly undercounted cheese relative to the diary; decomposed questions *over*-counted the diary, but not as strongly (p=.06). Decomposed questions similarly performed better when asking about cereal (although in this case, the *global* question significantly over-counted consumption); and although the bias of decomposed questions regarding physical activity was slightly lower than the bias from global questions, both significantly undercounted relative to the diary. However, for three other sets of questions, the *global* items were less biased. The difference was not significant for questions about pasta and dessert, but was for questions about oil.

Table 2: Bias for global and decomposed questions—original diary coding

Domain	Question type	Bias relative to diary (%)
Cheese	Global	-20.9***
	Decomposed	16.6*
Physical activity	Global	-19.5**
	Decomposed	-14.5**
Oil	Global	-16.9**
	Decomposed	- 25.1***
Cereal	Global	21.4***
	Decomposed	6.8
Pasta and rice	Global	1.3
	Decomposed	10.2
Dessert	Global	-9.9
	Decomposed	14.3

*p<.1, **p<.05, ***p<.01

Further muddling this picture is the fact that quantifying the coding diary data was complicated. Sometimes it was ambiguous whether a particular food or activity should be included, and other times it was unclear whether something should be counted one or more times (for example, two slices of pizza). Our initial coding scheme was liberal, accepting ambiguous items and counting them as many times as could be reasonably supported. However, given some concerns about the supportability of some of these decisions, we recoded data more conservatively, this time including only "clear cut" reports in the total counts.

An analysis using only the "clear cut" diary reports turns some of the original findings on their head. Global survey responses are not significantly different from the conservative diary estimates for three questions (cheese, pasta and dessert), whereas the multiple decomposed questions generate significant *over*-reports. And, although decomposed questions on physical activity and cereal did better than their global counterparts, the biases for global and decomposed questions were comparable (both were not significantly different from diary reports in the case of physical activity, and both were highly significant in the case of cereal).

Table 3: Bias for global and decomposed questions—conservative diary coding

Domain	Question type	Bias relative to diary (%)
Cheese	Global	-10.8
	Decomposed	22.5**
Physical activity	Global	-9.1
	Decomposed	-0.4
Cereal	Global	40.6***
	Decomposed	29.5***
Pasta and rice	Global	3.9
	Decomposed	16.7*
Dessert	Global	5.6
	Decomposed	28.7***

*p<.1, **p<.05, ***p<.01

Note: new coding rules did not affect questions about oil

Taken altogether, results are quite mixed. However, it is clear that the expected persistent advantage of multiple questions does not always materialize. In fact, global questions performed better than decomposed questions in a slight majority of the comparisons considered—but the results do not provide a clear slam-dunk either way. Six out of eleven global survey questions were not significantly biased compared to diary reports, but five were; conversely, five out of eleven decomposed question sets were not significantly biased compared to diary reports, but six were. In any case, support for our hypotheses was inconsistent.

Reconciling these results with other research

Although our hypotheses were formed through observation of actual survey responses in the laboratory, several previous studies would not have supported them. For example, Belli et al. (2000) demonstrate that at least some of the time, global questions are superior to multiple questions because the latter produce over-reports due to double counting. Menon (1997) also demonstrated that decomposed questions are not necessarily better, and hypothesized that a critical distinction centers on the regularity of the behavior in question. The basic idea is that decomposing questions should have a positive effect for reports of irregular behaviors, where *counting* is the likely response strategy. Decomposed questions makes counting simpler because there are fewer events to remember, and each is likely to be more distinctive. However, if the questions ask about regular behaviors, decomposition can actually interfere with the response process: whereas a reasonable heuristic might be available for estimating the response to a global question, one might not be available to formulate answers to more specific questions.

Although these studies are very useful, I have reservations about their applicability to most behavioral surveys, for several reasons. First, some of the examples considered in both studies are rather far-removed from the typical survey topics. Menon, for example, evaluates questions on the frequency of drinking water from public fountains and similarly banal events. Second, it is easy to understand why some of the specific questions proposed in these studies are *more* challenging than their global counterparts. In another of Menon's examples, a global question about frequency of washing hair is broken down into specific questions about washing hair in the morning, before dates, before a party, and so on. For most people, the global question can be answered based on predictable patterns. The decomposed questions, rather than parsing the response task into more memorable components, arguably break the question into *less* memorable events. Similarly, Belli et al. decompose a question about frequency of phone calls into questions about *less* memorable behavior, i.e., making local and long distance calls. But decomposition does not *necessarily* make a question's frame of reference more esoteric. In some cases, it could parse a difficult question into manageable parts. Third, the response task regarding their global questions is often different than the global questions of concern to survey researchers, and certainly different than the CES. Whereas some of the global questions cited by Menon and others could be answered through episodic recall, I will suggest that most of the global questions we are concerned with could probably not be answered through literal remembering of events or amounts. Estimation that draws upon heuristics and semantic memory, and involving some degree of uncertainty (see Burton and Blair, 1991) will almost certainly be required.

Given that we are usually concerned with estimation processes rather than literal recall, it seems reasonable that response quality will largely be a function of whether question construction lines up optimally with the organization of respondent memories. In other words, there are likely to be questions where global question are better than their specific counterparts and vice-versa.

It might be possible to understand the data presented earlier in that light. Contrary to initial expectations, decomposed questions about desserts consistently did not perform as well as the global question they were drawn from. One possible explanation is that memories are readily accessible that make it possible to estimate patterns of "eating sweets," whereas the information needed to estimate eating particular types of sweets is not as readily accessible. It is also possible that the subcategories of sweets (ice cream and frozen yogurt; cookies, cakes, pies and brownies; candy and chocolate) are not clearly distinct, leading to cognitive burden and perhaps double-counting.

In contrast, a global question about physical activity combines so many activities (yard work, chores, formal exercise, and walking for functional purposes) that it is difficult to imagine a reasonable response strategy. Parsing the question seems logical, and indeed, the decomposed version performs consistently better (although not at statistically significant levels).

I will suggest that some global questions provide reasonable tasks and some do not, and that there is no simple verdict regarding the general suitability of global questions. Reasonable global questions, quite simply, are those for which there is a reasonable cognitive strategy for responding to them. It is logical that questions are likely to be more effective when they line up with specific ways that memories are encoded (see Tulving, 1972). If specific questions are constructed in a way that corresponds well with memory, then moving to global questions is likely to entail a sacrifice in the precision of estimates. However, it is also possible that specific questions are not optimally designed given the way respondent memories are organized. In that case, a well-designed global question could actually do a better job invoking a reasonable estimation strategy.

Toward future research and evaluation of CES measures

If some global questions perform better than others, data is needed to evaluate how well they match up with what respondents can actually report, as opposed to more specific counterpart questions. It is a relatively straightforward manner to collect such data through cognitive laboratory evaluations. One approach is to use probes or think-alouds to investigate how respondents answer global and specific questions. Resulting data can be used to evaluate which level of question specificity matches respondent memories best-- or, whether particular global questions push respondent estimation beyond what seems acceptable. It is also reasonable to expect that in some cases global questions could pose simpler tasks than more specific counterparts.

Ultimately, validation data is highly desirable to evaluate the accuracy of global or specific questions, but collecting such data can be impractical and challenging. The research cited earlier is partially a cautionary tale of coding perils. When using diaries, great attention to coding procedures is necessary to ensure that diary data will be comparable to survey responses. It is also important to validate questions using the reference period that will actually be used in survey questions. Moving from a 30-day to a 3-day reference period was an unfortunate compromise that could have significantly altered the response task (i.e., whereas estimation is almost certainly required for a 30 day reference period, a counting strategy is plausible over three days). This compromise could have affected the generalizability of findings.

Another hypothesis that emerged from the study described earlier is that patterns of bias could be related to frequency of behaviors. In data not presented here, we found that the bias from multiple decomposed questions was relatively constant, whereas the bias of global questions (toward underestimation) seemed to increase the more frequent the behavior. The data are hardly definitive, but suggest the possibility that biases from global questions may be stronger for higher frequency behaviors. This could suggest that there is greater benefit to keeping multiple questions for such behaviors. This is certainly worth exploring in future research.

Global questions, especially those covering wide conceptual terrain and long reference periods, are unlikely to provide answers that are literally correct in terms of recall of specific information. They are more likely to invoke estimation strategies, but this does not necessarily mean that the information produced is of lesser quality. But it could mean that, compared to specific questions, respondents will rely on broader estimation strategies that could produce less precise data. This possibility should be explored on a question by question basis, as the quality of global questions can vary greatly.

References

Beatty, P., Cosenza, C., and Fowler, F. (2006). "Experiments on the Structure and Specificity of Complex Survey Questions." Paper presented at the 2006 AAPOR Conference held in Montreal, Quebec, Canada. Proceedings of the Section on Survey Research Methods, American Statistical Association, 2006.

Beatty, P., and Maitland, A. (2008). "The Accuracy of Decomposed vs. Global Behavioral Frequency Questions." Paper presented at the American Association for Public Opinion Research Conference held in New Orleans, LA, May 14-18, 2008.

Belli, R.F., Schwarz, N, Singer, E, and Talarico, J. (2000). "Decomposition Can Harm the Accuracy of Behavioural Frequency Reports." Applied Cognitive Psychology, 14, 295-308.

Burton, S. and Blair, E. (1991). "Task Conditions, Resposne Formulation Processes, and Response Accuracy for Behavioral Frequency Reports in Surveys." Public Opinion Quarterly, 55, 50-79.

Menon, G. (1997). "Are the Parts Better than the Whole? The Effects of Decompositional Questions on Judgments of Frequent Behaviors." Journal of Marketing Research, 34, 335-346

Tourangeau, R, Rips, L.J., and Raskinski, K. (2000). The Psychology of Survey Response. Cambridge, UK: Cambridge University Press.

Tulving, E. (1972). "Episodic and Semantic Memory." In E. Tulving and W. Donaldson, eds., Organization of Memory. New York: Academic Press.

Appendix: Global and Decomposed Questions

Physical Activity—Global

The first question is about moderate or strenuous physical activities you may have done at home or in your leisure time. By moderate or strenuous, we mean physical activities that lasted 10 minutes or longer, and caused at least some increase in heart rate or breathing. Please do not include physical activities done in any job for pay.

From [START DAY] to [END DAY], how much time did you spend doing moderate or strenuous physical activities, including yard work or other chores, walking for exercise or to get somewhere, or other exercise such as running, cycling, working out in a gym, or playing sports?

Physical Activity-- Decomposed

The next few questions are about moderate or strenuous physical activities you may have done at home or in your leisure time. By moderate or strenuous, we mean physical activities that lasted 10 minutes or longer, and caused at least some increase in heart rate or breathing. Please do not include physical activities done in any job for pay.

DQ1A. From [START DAY] to [END DAY], how much time did you spend doing moderate or strenuous yard work or other chores?

DQ1B. From [START DAY] to [END DAY], how much time did you do moderate or strenuous walking, either for exercise or to get somewhere?

DQ1C. From [START DAY] to [END DAY], how much time did you spend doing other moderate or strenuous exercise such as running, cycling, working out in a gym, or playing sports?

Cereal—Global

GQ2. How many times did you eat hot cereal such as oatmeal or grits, or cold cereal such as Wheaties or Cheerios, either as part of a meal or as a snack? [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY']

Cereal—Decomposed

DQ2A. From [START DAY] to [END DAY], how many times did you eat hot cereals such as oatmeal or grits either as part of a meal or as a snack?

DQ2B. How many times did you eat cold cereals such as Wheaties or Cheerios either as part of a meal or as a snack? [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY]']

Pasta and Rice—Global

GQ3. How many times did you eat pasta or rice? Include spaghetti, noodles, macaroni and cheese, pasta salad, any other kind of hot or cold pasta, and either white or brown rice. [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY']]

Pasta and Rice—Decomposed

DQ3A. How many times did you eat any kind of pasta, such as spaghetti, noodles, macaroni and cheese, pasta salad, or any other kind of hot or cold pasta? [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY']]

DQ3B. How many times did you eat any kind of rice, either white or brown rice? [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY']

Cheese—Global

GQ5. How many times did you eat cheese, including cheese as snacks or appetizers, and cheese in sandwiches, burgers, lasagna, pizza, or mixed in with some other dish? Do not count cream cheese. [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY'] [IF ASKED, CHEESED DIP CAN BE INCLUDED AS A SNACK OR APPETIZER]

Cheese—Decomposed

The next few questions are about any cheese you may have eaten from [START DAY] to [END DAY]. Please do not count any cream cheese you may have eaten.

DQ5A. From [START DAY] to [END DAY], how many times have you eaten cheese on a sandwich, including burgers?

DQ5B. How many times have you eaten cheese as snacks or appetizers? [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY']] [IF ASKED, CHEESED DIP CAN BE INCLUDED AS A SNACK OR APPETIZER]

DQ5C. How many times have you eaten cheese in lasagna, pizza, or mixed in with some other dish? Please do not include in your answer anything you reported in the previous two questions. [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY]']

Dessert—Global

GQ6. The next question asks about dessert foods, including ice cream, candy, chocolate, cookies, cakes and pies, and other sweet bakery items you might eat at breakfast or as a snack like doughnuts, Pop tarts, Danishes, and muffins. Please include anything that was low-fat or fat-free, but do NOT include sugar-free items.

How many times did you eat these foods? [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY]']

Dessert-- Decomposed

The next few questions are about dessert foods, including sweet baked goods you may have eaten at breakfast or as a snack from [START DAY] to [END DAY]. Please include anything that was low-fat or fat-free, but do NOT include sugar-free items.

DQ6A. During the period of [START DAY] to [END DAY], how many times did you eat ice cream, sorbet, or frozen yogurt?

DQ6B. How many times did you eat cookies, cakes, pies, or brownies? [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY]']

DQ6C. How many times did you eat candy or chocolate? Please do not include any chocolate that was an ingredient in something you've already told me about. [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY]']

DQ6D. How many times did you eat donuts, Danishes, Pop Tarts, or muffins? [IF ASKED FOR TIME PERIOD: 'FROM [START DAY] TO [END DAY]']

Oil—Global

GQ7. The next question is about oils used with food. You should include things like vegetable oil or olive oil, but not butter or margarine. From [START DAY] to [END DAY], how many times have you used oils to cook food, or added oils to foods like salad, pasta, or bread?

Oil—Decomposed

DQ7A. The next few questions are about oils used with food. You should include things like vegetable oil or olive oil, but not butter or margarine. From [START DAY] to [END DAY], how many times have you used oils to cook food?

DQ7B. From [START DAY] to [END DAY], how many times have you added oil to salads, such as oil and vinegar?

DQ7C. From [START DAY] to [END DAY], how many times have you added oils to other foods like pasta or bread?