Response to Issue Paper on CE Interview Structure

Frederick Conrad

Institute for Social Research, University of Michigan

1. Unnatural interview structure or unnatural categories?

The primary concern in the CEQ “interview structure” issue paper is that respondents are underreporting actual expenditures (articulated in the Gemini Project Vision Document, April 16, 2010, p. 2). The issue paper identifies one possible explanation for underreporting, namely that the order in which interviewers ask the questions cuts across the way respondents structure their memory for expenditures, interfering with recall of events they would ideally retrieve. While this may well account for part of the decline in expenditure reports my intuition is that underreporting is to a greater extent due to incompatibilities between the way respondents and the CE think about expenditure categories. More specifically, if the expenditure category in a CEQ item does not correspond to how the respondent classifies a particular expenditure, then the expenditure is unlikely to come to mind when probed with this category. Consider the CEQ category “Electrical personal care appliances” (Section 6b). It is plausible that for many respondents this category will not bring to mind the electric toothbrush they bought six weeks ago because they think of this as a *toothbrush* or *dental care product* (along with dental floss and tooth paste) but not an *electrical personal care appliance*. It’s not that they don’t know what “electrical personal care appliances” means and if asked whether an electric toothbrush falls into this category would likely say it does. But they don’t encode the purchase of their electric toothbrush as an instance of this category and so a question about this category will not bring the purchase to mind.

When I was at BLS, I conducted a laboratory study to test this idea with Norman Brown from the University of Alberta and Monica Dashen from BLS (Conrad, Brown & Dashen, 2004). In these experiments, we presented participants a series of ordinary nouns (e.g., “chair,” “rose,” “blood,” “Detroit,” etc.), one at a time, and instructed them to study the words to answer “some questions” about them later. The nature of these questions was intentionally left vague to encourage ordinary encoding processes. The questions they were subsequently asked concerned the frequency of study words that were either members of taxonomic categories such as “furniture,” “flowers,” “body parts,” “cities,” etc. or that contained certain properties, e.g., “wooden,” “red,” “large,” “slippery,” etc. The idea behind the experiments was that people spontaneously classify words – or events – as instances of certain “natural” categories but do not typically classify them into other types of “unnatural” categories. We reasoned that when presented with *chair*, people naturally think of the taxonomic category to which chairs belong – *furniture* – but not properties of chairs such as that they are often *wooden*. In the example above, the natural category for an electric toothbrush might be *dental care products* and an unnatural category might be the one used in the CEQ, *electric personal care products*.

In one of our experiments, we tested three groups of participants: the *taxonomic* group was asked for frequencies of taxonomic categories like furniture, the *implicit property* group was asked for frequencies of highly associated properties, and the *explicit property* group studied the words on which they would later be tested along with their properties (e.g., Wooden-Chair) and were then asked to report property frequencies.

We observed substantially more accurate frequency reports for the taxonomic than implicit property group; both groups underestimated actual frequency but the error was much larger for the implicit property group (see Figure 1). Participants who were asked about the frequency of taxonomic categories tended to use recall-and-count (or enumeration) strategies but those asked about properties were unable to recall and count, instead attempting to generate instances that contained the property and then checking whether they remember any of these words from the study phase of the experiment. This is a very ineffectual approach as it is unlikely they will generate all of the instances containing a particular property, especially for the higher frequency properties. Moreover, the taxonomic group also developed impressions of relative frequency (e.g., more pieces of furniture than animals) but the property group did not, severely restricting the range of estimation strategies they could use.

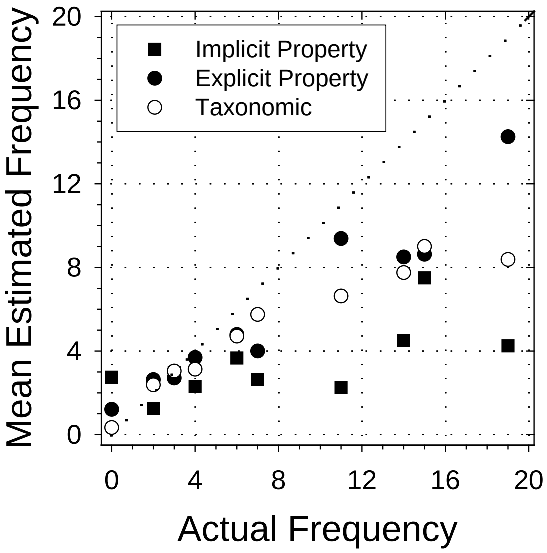

Figure 1. Actual versus reported frequency of implicit properties, explicit properties, and taxonomic categories, Conrad, Brown & Dashen, 2004.

The explicit property group performed much like the taxonomic group. Their frequency estimates were as accurate as were those for the taxonomic category group, and participants were able to recall and count items that shared a property much like they did for taxonomic categories. The fact that properties – not typically encoded with the event – can serve as effective retrieval cues when they are made an explicit part of to-be-recalled event suggests that, in principle, CE respondents might be sensitized to the CE category structure in a way that would lead them to think about an electric toothbrush as an electronic personal care product at the time of purchase. The problem is that the amount of training required to accomplish this for the many categories within each of the twenty sections of the questionnaire would be massive, perhaps unreasonably so. But over multiple waves of the survey, it might be possible to teach respondents to organize purchases according to CE categories at the time they engage in those purchases. The literature on “memory improvement”– which is essentially the activity CE would want respondents to engage in – instructs respondents to actively organize while filing experiences in long term memory. Consider this observation from Cermak (1975): “The person who organizes information as he learns it, who puts it into a proper niche in long term memory, is more likely to locate it when he wants it than is the person who does not take the time to organize as he learns” (p. 41).

It is conceivable that the longitudinal structure of the CE might make a “memory improvement” intervention possible. By training respondents in the first interview how to think about their purchases at the time of purchase, and reminding them of this in each interview, respondents may be more likely to report purchases they would overlook without this kind of training. The training might be focused on those CE categories which lab tests, as in the experiment described above, indicate are particularly unlikely to come to mind at the time of purchase. If this is effective, it could help ease the underreporting problems referred to in the interview structure issue paper.

The general lesson is that categories that make sense for theoretical or analytic reasons (as presumably is the case for the CE category structure) may not be in the forefront of respondents’ minds while they are purchasing products from those categories. If respondents do not classify an event as an instance of the appropriate CE category when it occurs, they will be unlikely to encode the category as part of the event; the result is that their recall for the event will be poor when a CE interviewer asks about that category. But calling their attention to an alternative categorization approach might help.

2. Interview structure and recall

The discussion so far suggests that aligning the way respondents and the survey sponsors encode and classify purchases is the critical step in improving recall. The implication is that there is little that can be done at the time of recall (the interview) if there is misalignment at the time of encoding. Nonetheless, there may be interview sequences that will help respondents recover expenditure memories they might not report when probed under the current approach with a CE category. One approach that might hold promise is to encourage respondents to recall the context in which they might have made purchases, e.g., online, on the telephone, by mail, in drive-through retail outlets, in convenience stores, in enclosed malls, etc., and if they can recall any purchases in a particular context the interviewer (or a coder, after the fact) classifies the individual purchases into CE categories. For example, a respondent might indicate she has in fact made numerous online purchases and when probed she reports books, computer hardware, cell phone service, plane tickets and cosmetic surgery. Note it is the purchase context and not the consumption context that matters.

The theoretical basis for this proposal is the *encoding specificity* notion (Tulving and Thomson, 1973; Watkins and Tulving, 1975). In essence Tulving and his colleagues demonstrated that recall is superior when the context at the time of recall matches that at the time of encoding. In fact, they demonstrated that under these conditions, recall is superior to recognition when the context does not match – and recall is almost never better than recognition! This had been demonstrated in many ways. Godden and Baddeley (1975) report a compelling demonstation in which deep-sea divers learn material under water and on land. When tested in the same context in which they learned the material their recall was superior to when the encoding and test contexts differed.

This idea has been applied to memory improvement in what is known as *context reinstatement*. The idea is that if people can bring to mind the circumstances in which they experienced an event then, to some degree, they can match the context when trying to recall aspects of the event. Context reinstatement is used to collect evidence from eyewitnesses of crimes and accidents, one method in a collection of memory enhancers known as “cognitive interviewing” – not to confused with the questionnaire pretesting method of the same name. In addition to context reinstatement, cognitive interviewing (CI) consists of varied recall order, changing physical perspective when recalling, and recalling everything that comes to mind. This combination of techniques has been shown in numerous studies to improve memory performance relative to free recall (see, for example, Bekerian & Dennett, 1993, Geiselman et al. 1986) but the combination of context reinstatement and recalling everything seems to be the locus of the “CI superiority” (Milne & Bull, 2002). Note that the method is not always entirely effective: in the Milne and Bull study, the combination of context reinstatement and recall everything led to better recall than any one technique alone but it did not lead to better recall than in a free recall control condition.

The main difference between CI context reinstatement and the approach I envision for the CE is that in the former, the eyewitness is asked about a particular event, such as a crime, and generates the context but in the latter the context – not a particular event – is provided to the respondent. It might be the case that the contexts need to be presented at a finer level of detail than in the examples above, e.g., online purchases that involved shipping a physical product versus online purchases of an online product or online purchases of an offline service. If there are many instances of a context, e.g., the respondent shops at several enclosed malls, then he or she might be asked to enumerate those malls and then be queried about his purchases at each of them.

Note that this has some of the character of the Event History Calendar (EHC) in that it recalled context stimulates subsequent recall. In the case of the EHC, the events that respondents retrieve from one life theme (e.g., employment) serve as context for the recall of events in another (e.g., residential moves) and so stimulate recall relative to performance without that context. Belli (1998) refers to this as "parallel retrieval" (p. 391) because the thematic narratives are parallel, i.e., run concurrently to one another. In the approach I am proposing the interviewer provides the context to respondents but the mechanism – recalled information providing context that can stimulate recall – is essentially the same. An important distinction is that the kinds of events whose recall seems to be promoted by the EHC approach tend to be extended over time – narratives – whereas most purchase events are more temporally discrete.

This approach might only be used to supplement the current CE interview procedure for those CE categories where underreporting seems most severe, i.e., the interviewer would ask only about purchase contexts in which the most products that are most underreported are purchased. These products and their associated purchase contexts could be established in a dedicated study that would be part of the research program I alluded to above. Whatever purchase contexts are ultimately presented to respondents, the interview would be structured by context, not category, and after respondents were sure they had nothing else to report for one context the interviewer would advance to the next, possibly iterating through a set of context instances provided by the respondent.

The approach I am proposing is not inherently ordered but the list of purchase contexts must be presented in some order, i.e., it is not structure-free. Certainly, if recalling an event in one purchase context reminds the respondent of an event in another purchase context, this should be accommodated – and I discuss some of the interface design issues that might help support this below. For example, if recalling an online expenditure for a book reminds the respondent of a book purchase in a brick and mortar bookstore, he or she should be able to report the related purchase at that point. My intuition is that in the interest of completeness, it would be wise to return to the original context – online purchases – before shifting to the context of the "reminded purchase" – physical stores – but this is an empirical question that can be evaluated as part of a comprehensive evaluation of the approach.

3. Flexible data entry

I cannot think of a good reason to impose a rigid order on data collection, although I suspect that most respondents will respond in the order the questions are asked. I am not terribly concerned about question order effects – differential context effects – as alluded to on p. 1 of the issue paper. Generally, these effects concern information activated by an earlier question that is incorporated into the answer for a later question (assimilation effects) but is only available if that earlier question is asked. But it's hard to see how asking about *home furnishings* before *clothing* for one respondent and in the reverse order for another introduces measurement error if the different orders result from different self-generated reminders during the recall process. So it makes sense to accommodate respondents' preference to report on one category – or context in the approach I have proposed – by allowing the interviewer to enter data in whatever order respondents happen to provide them.

The issue paper suggests that current CAPI software is not capable of supporting the kind of flexibility that would be required for interviewers to enter data in whatever order respondents produce it (p. 4-5). This may well be so, but CAPI software probably does not represent the state of the art in user interface design. Other interface design techniques may help with provide the flexibility to allow interviewers to shift capture unstructured recall as it occurs. A user interface that displays the questionnaire structure as an network would allow an interviewer to click or preferably touch a node corresponding to a section of the questionnaire and expand the section with a single action. This should allow rapid shifting from one section to another.

Multimodal interfaces may also help the interviewer keep up with a respondent's unstructured recall. For example, one can imagine that if a respondent recalls a purchase from a category (or context) about which the interviewer is not currently asking, rather than jumping to the appropriate section of the questionnaire the interviewer might create an on-screen note using a stylus or finger, analogous to the paper notes they apparently use in practice (issue paper, p. 2). The interface could be designed so that these input devices would "grab" control from the primary entry mode – presumably a keyboard, whether physical or virtual – and allow the interviewer to enter text into a "note field." The note field could be designed to promote integration with the rest of the questionnaire by allowing the interviewer to enter key attributes such as the category to which they believe the recalled expenditure belongs, the outlet at which it was purchased, when it was purchased, etc. Or one can imagine the interviewer speaking the note discretely while the respondent is engaged in recall and designing the system to both capture and recognize the speech, integrating it as it would the handwritten note.

Proponents of multimodal interfaces argue that they allow more natural interaction than is possible with a single input mode and a single display medium. For example, Oviatt and her colleagues (Oviatt and VanGent, 1996; Oviatt and Olson, 1994) observed that users of a multimodal dialog system switched between pen (stylus) and speech primarily to contrast a new discourse contribution to a previous one in resolving communication errors. Johnston, et al. (2002), reports on a system that helps users obtain information about the services and establishments in particular cities that presents the information graphically (maps), textually and using speech output; users enter their queries by speaking, pointing, writing, typing and gesturing with a stylus (e.g., circling an area on a map). Findings by Oviatt and her colleagues and Johnston and his suggest that users find value in multimodal options that could potentially improve system performance. Johnston (2007) argues that multimodal survey interfaces are attractive because they are inherently natural, and support the routine human practice of combining speech with pointing, gestures and facial expressions, and enable users to choose whatever combination of modes seems most appropriate at any moment.

4. Conversational interviewing

The issue paper speaks about conversational interviewing at several points, in particular the work Michael Schober and I have carried out to explore the costs and benefits of allowing interviewers to assure respondents understand questions as intended by using whatever words they deem necessary to explain the survey concepts (e.g., Schober & Conrad, 1997; Conrad & Schober, 2000). There may be a role for this approach in the CE but I believe the way it has been described somewhat mischaracterizes this role. While the technique encourages interviewers to use whatever words seem appropriate for the respondent and the situation, we have said very little about question order. The point of allowing interviewers to choose their words was to help them explain the intentions behind survey concepts; some respondents will require different wording to reach the same interpretation and the same wording may lead to different interpretations given respondents' backgrounds and beliefs and assumptions. In all of our studies, conversational interviewing led to substantially more accurate responses than interviews that strictly scripted the questions and prevented clarification, i.e., standardized interviews. But the questions were always read in a fixed order. The approach is simply not about flexible interview order.

As pointed out in the issue paper, conversational interviewing takes more time than strictly standardized interviews. The reason is that it allows interviewers to clarify question meaning, either when respondents express confusion or interviewers believe the respondent needs clarification, and clarification takes time not. So it's really a matter of the number of words spoken in the interview, primarily by the interviewer. There could in fact be a temporal cost in adopting the purchase-context-instead-of-purchase-category approach I have proposed above. First if respondents recall events that they don't recall under the conventional approach, this will take more time. In addition, there might be a cost in following respondents' unstructured recall – even with appropriately designed user interfaces – because it is likely some of the expenditures that the respondent is reminded will be out of scope for the CE and so in effect a waste of time. As with conversational interviewing, the tradeoff that practitioners must weigh is improved data quality versus increased cost.

5. Research program

At several points in the above discussion, I have suggested that decision require empirically based guidance. What follows is the start of a list of research issues that will need to be addressed if this approach is to be workable:

1. Does context reinstatement, i.e., asking about any expenditures in general types of purchase contexts instead of product categories, help respondents recover purchase events that do not come to mind when they are probed with CE categories?
2. What CE products are most underestimated under the current approach? In what contexts are they most likely to be purchased?
3. What are the temporal costs of following respondents' unstructured recall? Can interviewers do this effectively in real time with training and a good interface design without affecting respondents' recall.?
4. User interfaces to promote flexible data entry in noisy tasks: what are effective combinations of input devices/modes for an unstructured verbal task?

References

Bekerian D.A. & Dennett, J.L. (1993(. The cognitive interview technique: Reviving the issues. *Applied Cognitive Psychology, 7*, 275–297.

Belli, R. F. (1998). The structure of autobiographical memory and the event history calendar: Potential improvements in the quality of retrospective reports in surveys. Me*mory, 6*, 383-406.

Conrad, F.G., Brown, N.R. & Dashen, M. (2003). Estimating the frequency of events from unnatural categories. *Memory and Cognition, 31*, 552-562

Geiselman, R.E., Fisher, R.P., MacKinnon, D.P., & Holland, H.L. (1986). Enhancement of eyewitness memory with the cognitive interview. *American Journal of Psychology, 99*, 385–401.

Godden, D.R. & Baddeley, A.D. (1975). Context dependent memory in two natural environments: On land and under water. *British Journal of Psychology, 66,* 325-331.

Johnston, M. (2008). Automating the survey interview with dynamic multimodal interfaces. In Conrad, F.G. & Schober, M.F. (Eds.), *Envisioning the survey interview of the future (137-160).* New York: Wiley & Sons.

Johnston, M., Bangalore, S., Vasireddy. G., Stent, A., Ehlen. P., Walker, M., Whittaker, S., & Maloor, P. (2002). MATCH: An architecture for multimodal dialog systems. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Milne, R. and Bull, R. (2002). Back to Basics: A Componential Analysis of the Original Cognitive Interview Mnemonics with Three Age Groups. *Applied Cognitive Psychology, 16*, 743–753.

Oviatt, S., & Olsen, E. (1994). Integration themes in multimodal human-computer interaction. *Proc. of the lntl Conf. on Spoken Language Processing, 2*, 551-554.

Oviatt, S. L., & vanGent, R. (1996). Error resolution during multimodal human-computer interaction. *Proc. of the Intl. Conf. on Spoken Language Processing*.

Tulving E, Thomson DM. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 353–370.

Watkins, M. J. and Tulving, E. (1975). Episodic memory: When recognition fails. *Journal of Experimental Psychology: General, 104*, 5-29.