How Do people Interpret Open-ended Categorical Questions?

Monica Dashen and Scott Fricker

U.S. Bureau of Labor Statistics

2 Massachusetts Ave. N.E., Suite 1950 Washington D.C. 20212 USA

Key words: Survey methodology, open-ended questions and response errors

This paper explores the effects of respondent interpretations on data quality where the device is a categorical question. Categorical questions are an aggregate of questions that often have an accompanying list of category members (or response alternatives). For example, a marketing survey might inquire about *girls' clothing* and provide a list of members (e.g. *dresses*, *skirts*, *blouses*, *shirts*, and *pants*) from which the respondents can select their appropriate answers. It is commonly believed that how people interpret questions will influence how they arrive at their answers (e.g., Clark & Schober, 1992; Martin & Polivka, 1995; Tourangeau & Raskinki, 1988). Evidence suggests that people may use the accompanying list of members to clarify the intent of the question (e.g., Schwarz & Hippler, 1991). Yet there is little evidence to clarify how people interpret categorical questions when they have no accompanying list.

Categorical questions are often used in surveys because these questions save time and reduce respondent burden. Saving time contributes to accuracy because a respondent who must wade through scores of questions may tend to answer "no" more frequently than is not accurate simply to speed the interview (e.g., Lehnen & Reiss, 1978). Subsequently, reducing the number of questions asked ought to reduce the respondent's tendency to say "no."

Often items are aggregated into categories according to the needs of the data user, rather than of the respondent. For example, televisions and videocassette recorders (VCR's) are not in the same Telephone Point of Purchase Survey (TPOPS) category. The basis for this distinction is that televisions were originally purchased from a single outlet of television stores; thus, the distinction is historical in nature (e.g., Cage, 1996). Respondents who are not privy to this historical basis cannot reasonably be expected to understand or follow the distinction and thus might report all things related to televisions (VCR's, video tapes, video games, cable boxes, television stands, and so forth) in the television category. The histories governing the assignment of items to various categories can and do affect the ease with which respondents understand categories and can therefore affect the integrity of the resulting data.

A list often offsets any inadequacies of a categorical question because the respondent uses the list's contents to clarify the categories' contents (Schuman & Presser, 1981). Adding a list to a categorical question may reduce the respondents' uncertainty (and therefore improve data quality) because the respondents assume that if the item is not on the list, then it is not in the category (Schwarz, 1996). For example, the absence of *VCRs* as a response alternative for the category *Televisions* may signify to the respondent that *VCRs* are not a member of the category.

Despite the benefits of the list, there are some situations where respondents do not have the opportunity to see the list. Telephone respondents can not see the list, for example, but face-to-face respondents can (Groves & Kahn, 1979). One could argue, however, that the telephone interviewer has the option to recite a subset of the list's cues to help clarify the category title and its contents. Often, however, respondents do not learn about this subset unless they ask, something, as Schober and Conrad, 1996 and Conrad and Schober, 2000 have shown, respondents tend not to do.

The absence of the list makes the categorical question a good device for exploring how respondent interpretations affect data quality. Without the list, respondents are left to their own judgments and experiences in interpreting the categorical question. Respondents are also left confused as to whether the criterion they deduced for the inclusion of members is correct. If the criterion is wrong, respondents may

include incorrect members (false positives) and exclude correct members (omissions), which can affect data quality.

The failure to infer the correct criterion in open-ended categorical questions becomes increasingly more likely when respondents are not asked to mention the items they consider members of the category when responding to the question. Instead, all they have to do is reply "yes" or "no" when asked a categorical question. This format is problematic for two reasons. First, the "yes/no" format does not encourage respondents to ask what belongs in the category. If they simply say "yes" or "no," then they likely will not state how they decided. The second reason is that the phone interview, with its more time pressured format than that of a face-to-face interview, makes people want to answer each question as quickly possible; therefore respondents will likely not take the time to state how they decided (e.g.,Rockwood, Sangster, & Dillman, 1997).

To date, despite the prevalence and pitfalls of open-ended categorical questions little is known about how people respond to them. Many survey methods researchers show how people use a list to clarify the contents of a categorical question (e.g. Schwarz & Hippler, 1991). However, few survey methods researchers show how people understand such questions without the aid of a list. The present work fills this gap by focusing on how people formulate a criterion of inclusion for open-ended categories. To do so, we turn to the psychological literature where commonly researchers account for how people respond to open-ended categorical questions.

Criterion for Inclusion of Responses in Open-Ended Categorical Questions

Although much research has been done on the topic of categorization, we have chosen to focus on three discernable and identifiable theories -- physical similarity, essence, and goal—that offer predictions as to how people formulate a criterion of inclusion for consumer-oriented categories such as those in the TPOPS. (In this work, we will focus on the clothing-, food-, and computer-related questions in the TPOPS.) Table 1 describes these theoretical predictions.

In our discussion of each theory, we will rely on the TPOPS *Women's Dresses* category to point out the different theoretical predictions. The TPOPS designers classify all types of dresses (e.g., *gowns*, *sun dresses*, and *business dresses*) as members of the category *Women's Dresses*, whereas all types of accessories (e.g., *scarves*, *hats*, *stockings*, and *belts*) are not classified as members.

Table 1. Summary of Theoretical Predictions.

Expla-nation	Methods of Interpretation for Open-Ended Categorical Questions		
	Physical Similarity	Essence	Goal
Definition of Processes	All items that look alike go together.	All items that share an inherent property go together.	All items that serve the purpose for the category go together.
Women's Dresses	Dresses that have a one-	Dresses that share the implied	Dresses and accessories

	bodice and skirt belong..	associated with a work place or special occasion belong.*	goal "of getting dressed," belong.

(*) Note: People will often use the phrase a " dress occasion" to describe the level of formality of an event.

The physical similarity proponents argue that people strictly decide category membership based on an item's physical resemblance to other category members (e.g., Medin, 1989). As a reaction against the physical similarity proponents, the essence advocates argue that people look beyond the surface of an entity and focus on an inherent property when assessing whether something is a member of a category (e.g.,Rips, 1989).

The essence interpretation differs from the physical similarity interpretation in that it is more restrictive in terms of what are acceptable candidates.

Under the essence interpretation, for example, the respondent would not include a *T-shirt dress* or a *sundress* in the *Women's Dresses* category because these dresses do not have the implied formality, even though they have a one-piece bodice and skirt, as required by the physical similarity interpretation.

Unlike the physical similarity and essence proponents, the goal advocates argue that all items that serve a *purpose* for the category go together (e.g. Barsalou, 1983). As can be seen in column 3 of Table 1, respondents interpret the *Women's Dresses* category as the act of getting dressed and go beyond the process of listing various types of dresses.

The goal-oriented interpretation differs from the physical similarity interpretation in that people do not restrict themselves to listing all things that resemble a dress. Similarly, the goal-oriented interpretation differs from the essence interpretation in that people who adopt the goal-oriented interpretation would not restrict themselves to simply listing formal dress wear.

Respondents may engage in many different types of goal-oriented thinking. Because the respondents in this work will be asked about items such as food, clothing and computers, it makes sense to focus on the two most likely types of goal-oriented thinking-- "to accompany" and "to make," -- that the respondents may adopt. Let us first consider the "to accompany" type of goal-oriented thinking. When asked about *Coffee* purchases, people might say *sugar*, *cookies*, *milk* and *spoon* and justify these expenses as things used in conjunction with coffee. Now let us consider the "to make" type of goal-oriented thinking. When asked about *Coffee* purchase, people might say filters, *coffee pot*, *water*, and *coffee grounds* and justify these items as things needed to make coffee.

Aim of Present Work

The specific aims of this work are several. One aim is to find out whether respondents systematically formulate a criterion of inclusion for open-ended categorical questions. Another aim is to identify ways to prevent errors before they occur. It is reasonable to assume that the closer the fit between a category name and description and the respondents' expectations, the lower the number of errors will be. For that reason, the present work seeks the most successful criterion for each categorical question and recommends that it be used as a lead-in that clarifies the intent of the question (e.g., Belson, 1984; Fowler, 1993). One could argue that repairing a category title is a more straightforward way of reducing the number of errors. However, survey designers must also be willing to reclassify items (i.e., add items to some categories and move items from one category to another). Given the competing needs of the data users, the likelihood of survey designers re-arranging the contents of the categories is small. Accordingly, the optimal solution is to provide a lead-in statement to clarify the intent of the question.

Study 1

The aim of Study 1 is to understand how people interpret open-ended TPOPS category questions pertaining to food, clothing and computers. In this study, respondents were asked to think of all relevant items in a category that they might buy and why they believed a particular item belonged in a given category. The justifications allow an exploration of the reasoning used to interpret the category title.

Method

Twenty-two participants received a booklet containing instructions and twelve category titles. The instructions, located on the first page of the booklet, pertained to all twelve categories. The remaining pages consisted of category titles. Each category question was on a separate page with ample space for participants to write down all relevant purchases and justifications (why the participants believed an item belonged in the categories) for those purchases. (Note: Participants were instructed to provide a justification for each item.) The participants were required to generate example purchases for the following categories: (a) *Bread*, (b) *Breakfast Cereal*, (c) *Coffee*, (d) *Cookies*, (e) *Lettuce*, (f) *Potatoes*, (g) *Computer Software*, (h) *Personal Computers & Peripheral Equipment*, (i) *Men's Suits and Sport Coats*, (j) *Men's Outerwear*, (k) *Women's Dresses*, and (l) *Women's Outerwear*. Respondents were instructed to interpret the open-ended categorical questions as hypothetical. Though all participants saw the same set of category titles, no two people saw the same order of category titles in the booklet. (Note: these categories were patterned after the TPOPS questions.) With one exception (the *Personal Computers & Peripheral Equipment* category), all categories are designed in such a way that only literal instantiations will belong in the category. For example, the *Coffee* category consists of items such as *decaffeinated coffee* and *flavored coffee*. Similarly, the *Women's Dresses* category consists of such items as *sun dresses, evening dresses*, and *bridal dresses*. The *Personal Computers & Peripheral Equipment* category consists of the computer in its entirety (literal instantiations), but it also includes items that accompany computers and are not, strictly speaking, computers. For example, *modems, speakers, printers*, and other peripheral devices are included in that category. In this respect, the *Personal Computers & Peripheral Equipment* category differs from the other two mentioned above. Including *printers* in the *Personal Computers & Peripheral Equipment* category is akin to including *coffee filters* in the *Coffee* category or *slips* in the *Women's Dresses* category.

Results/Discussion

The discussion of the data analysis has been broken down into two sections. Section one describes the scoring procedure. Section two discusses the results of the exemplar generation task.

Description of Scoring Procedure.

The scoring procedure for the exemplar generation task (in which people were asked to say what belongs in a particular category and why they think it belongs) was two-fold. First, the fictitious purchases listed were scored as intended or unintended reports based on whether they correspond to the intentions of the designers of the TPOPS survey. Second, the open-ended justifications were collected and classified into various categories for further analyses. These two procedures are further discussed in the following two sections: (1) scoring of listed fictitious purchases and (2) scoring of justifications.

Scoring of Listed Fictitious Purchases.

For each participant, the items or fictitious purchases reported for each category were classified into three mutually exclusive categories: (a) intended exemplars (present on the TPOPS cue sheet and reported by the respondents), (b) intended but not mentioned exemplars (items reported only in the TPOPS cue sheet) and, (c) unintended exemplars (reported by the respondent but not on the TPOPS cue sheet). Using the intended exemplar and unintended exemplar counts, two proportional measures of performance were calculated: intended exemplar rate and unintended exemplar rate.

Scoring of Justifications. Responses to the question, "Why do you think this item is a member of the category?" in the exemplar generation task were classified into one of four major groups: (1) literal, (2) to make, (3) to accompany, and (4) essence. First, the "literal" group involved those participants who interpreted (or justified) the category titles in a literal and narrow manner. In doing so, respondents tended to comment on the fact that it is an instantiation of a category (e.g., "it is a type of lettuce"). Second, the "to make" group involved those respondents who justified their responses as things that were either used to make something or used as an ingredient in something (e.g., "water is used to make coffee;" "potatoes are used to make potato salad"). A justification coded as "to make" is related to the goal-oriented interpretation of the category. Third, the "accompany" group involved those participants who said that the item was used to accompany something (e.g., "cream is used to flavor my coffee;" "sour cream is a topping for potatoes"). A justification coded as "to accompany" is related to the goal-oriented interpretation of the category. Fourth, the "essence" group involved those participants who said that the item contained some sort of underlying property of the category (e.g., "gloves provide warmth;" "coffee contains caffeine which is a 'pick me up'"). A justification pertaining to the essence of the category is related to the essence interpretation of the category.

Exemplar Generation Task Performance.

Analysis of the exemplar generation task performance also involved two steps. First, preliminary analyses involving accuracy were conducted to find out just how good people were at generating items that are on the TPOPS category cue list. Second, analyses were conducted to find out just how people interpret the category title and whether their interpretations are random or based on some systematic misinterpretation.

Category Accuracy. The present analysis was performed to address the following question: How difficult is it for respondents to interpret the questions correctly? As mentioned previously, two indices were calculated: (1) intended exemplar rates and (2) unintended exemplar rates. For the sake of clarity and brevity, the questions have been collapsed into three different types (food, clothing and computer).

If the respondents had understood the question perfectly, they would have been expected to report all (or almost all) the correct items for each category without erroneously reporting false positives (unintended exemplars). Omissions and unintended exemplars did occur. No differences in intended exemplar rates (.45 -.57) were observed among category types ($F(2,173)=.94$, $p=.40$). Similarly, there were no differences in unintended exemplar rates (.43-.55) among category types ($F(2,173)=.94$, $p=.40$).

In summary, the low intended exemplar rates and high unintended exemplar rates indicate that people did not understand the open-ended categorical questions as intended. This finding raises a follow-up question: Are the omissions and unintended exemplars random, or are they based on systematic misinterpretation? In other words, are survey respondents consistently or frequently using some rationale to guide their responses? The next section addresses this question.

Category Accuracy & Justifications. The analyses of category accuracy and respondents' justifications were designed to answer two questions intended to give a general idea of how people interpreted category titles and arrived at their reports, both correct and false: (1) Is there a particular interpretation that tends to lead people in the right direction and produces significantly more intended exemplars for all or some of the questions than do other interpretations? Given that a person is correct what strategy does he/ she use? (2) Is there a particular interpretation for each category type that tends to lead people astray and produce more unintended exemplars for all or some of the categories than do other interpretations? Given that a person is incorrect because he/she produced an unintended exemplar what strategy did he/she use?

The goal of the first question was to find out which method people resort to most frequently when they produce the correct answer (or intended exemplar). The frequencies of intended exemplars (as contained in

the entries for each category of Table 2a) were compared across all four methods using an adjusted probability level chi-square analysis (Agresti, 1990).

Table 2a: Intended Exemplar Percentages by Justification

		Categories	
	Food	*Clothing*	*Computer*
	Intended	Intended	Intended
	Exemplars	Exemplars	Exemplars
Justifications	*(column 1)*	*(column 2)*	*(column 3)*
To Make	2.7% [4]	1.3% [1]	6.3% [5]
To Accomp.	5.4% [8]	2.5% [2]	22.8% [18]
Literal	62.2% [92]	60.0% [48]	51.9% [41]
Essence	4.7% [7]	20.0% [16]	2.5% [2]
No Just.	12.2 %[18]	6.3%[5]	8.9%[7]
Uncodable	12.8% [1]	10.0%[8]	7.6%[6]
Totals	100 %[148]	100%[80]	100%[79]

The results indicate that when people adopted the literal list method, they were most likely to generate intended exemplars. The reliable chi-squares ($p < .001$) comparing the literal and other interpretations ranged from 16.00 to 80.67. (The literal and accompany comparison for the computer category did not yield a reliable difference.)

The goal of the second question was to find out which method people resort to most frequently when they produce unintended exemplars. The frequencies of unintended exemplars (as contained in the entries for each category of Table 2b) were compared across all four methods using an adjusted probability level chi-square analysis (Agresti, 1990).

Table 2b: Unintended Exemplar Percentages by Justification

		Categories	
	Food	*Clothing*	*Computer*
	Unintended	Unintended	Unintended
	Exemplars	Exemplars	Exemplars
Justifications	*(column 1)*	*(column 2)*	*(column 3)*
To Make	19.9%[39]	3.0% [4]	15.8% [9]
To Accomp.	38.8%[76]	28.0% [37]	29.8%[17]
Literal	13.3%[26]	22.0%[29]	21.1%[12]
Essence	6.1%[12]	16.7% [22]	7.0% [4]
No Justif.	9.1%[18]	6.8%[9]	8.8%[5]
Uncodable	12.8%[25]	23.5%[31]	17.5%[10]
Totals	100.0%[196]	100.0%[132]	100.0%[57]

The observed results indicate that the number of unintended exemplars was higher for the "to accompany" and "literal" methods than for the "essence" and "to make" methods with the exception of the food categories.

Study 2

The observed results from Study 1 suggest that people systematically formulate a criterion of inclusion for open-ended categorical questions. Study 2 was designed to confirm and extend the findings of Study 1.

Method

Forty-five people generated items using one of four interpretations: essence, to make goal, to accompany goal and literal. Following this exercise, the findings were then compared to Study 1 for validation purposes. If the item generated under the same interpretation in Study 2 was identical to that item in Study 1 then it was considered validated. For example, if *creamer* is reported in Study 2 under the "accompany" condition, and also reported in Study 1 and justified as "to accompany" than the findings of Study 1 were considered validated.

Results

The findings indicate generally high agreement between the items generated in Study 1 and the items generated in Study 2 under the same condition. The clothing category led the way with perfect agreement (100%). For example, respondents in Study 1 listed scarves in the category *Women's Outerwear* and justified that item as "accompanying women's coats." In Study 2, respondents who were told to list items that accompany other items did indeed list scarves. The lowest level of agreement was 67% in the computer category for the essence justification. That is to say, respondents in Study 2 who were instructed to list items according to their essence managed to come up with only 67% of the items generated by their Study 1 counterparts and justified according to their essence. > In sum, the findings suggest that the justifications in > Study 1 were fairly reliable and did provide insight into how people > interpreted the category title.

Conclusions

The exemplar generation and justification data reported in these studies address a theoretical issue central to survey methods: How do respondents interpret open-ended categorical questions? We constructed two studies to address precisely this question. Both studies asked respondents to generate items for given categories. In Study 1, the respondents were given a category title (e.g., *Coffee*) and asked to list items they thought belonged in the category and concurrently to justify those items. For example, a respondent could say that *cream* belongs in the *Coffee* category and justify that response by saying that she always takes *cream* with her coffee. (Such a justification would be classified as "to accompany" since the "respondent's justification was that the item is one he/she uses to accompany coffee.) Study 2 built on the results of Study 1. In Study 2, respondents were again given an open-ended categorical question. This time, however, they were also provided with a justification method. For example, a respondent might be given the category *Coffee* and asked to identify all things that are used "to accompany" coffee.

The respondents' answers in Study 1 followed predictable patterns; the justifications for incorrect responses fell into a few fairly well-defined groups and were not randomly errant responses. This finding is particularly encouraging in light of the alternative. Random unintended exemplars would suggest that survey designers can do little to predict and account for respondents' reactions. The fact that the respondents followed similar patterns suggests that it is possible to understand these responses (as the present work seeks to do). In addition, such an understanding will, in turn, allow survey designers to incorporate those methods in an effort to reduce the number of unintended exemplars and increase the number of intended exemplars.

References

Agresti, A. (1990). Categorical Data Analysis, New York: John Wiley & Sons, Inc.

Barsalou, L.W. (1983). Ad-Hoc Categories. Memory & Cognition, 11, 211-227.

Belson, W. (1984). Design and Understanding of Survey Questions. Gower Publishing, England.

Cage, Robert (1996). New Methodology For Selecting CPI Outlet Samples.Monthly Labor Review, 118(12), Pp. 49-83.

Clark, H. & M.F. Schober (1992). Asking Questions and Influencing Answers. In J.M. Tanur (Ed.) Questions About Questions. (Pp. 15-48). New York: Russel Sage Foundation.

Conrad, F.C and Schober, M.F. (2000). Clarifying Question Meaning in a Household Telephone Survey. Public Opinion Quarterly. 61(1) 1-28.

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

Groves, R.M., & Kahn, R.L, (1979). Surveys by Telephone: A National Comparison with Personal Interviews. New York: Academic Press.

Lehnen, R.G. & Reiss, A.J. (1978). Response Effects in the National Crime Survey. Victimology. 3, Pp.110-160.

Martin, E. & Polivka, A. (1995) Diagnostics for Redesigning Questionnaires. Public Opinion Quarterly. 59(4) 547-567.

Medin, D.L.(1989). Concepts and Conceptual Structure. American Psychologist. 44(12) 1469-81.

Rips, L. J. (1989). Similarity, Typicality, and Categorization. In S. Vosniadou & A. Ortony (Eds.), Similarity and Analogical Reasoning. (Pp.21-59). Cambridge: Cambridge University Press.

Rockwood, T.H., Sangster, R.L, & Dillman, D. A.(1997). The Effect of Response Categories on Questionnaire Answers: Context and Mode Effects. Sociological Methods and Research. 26(1) 118-140.

Schober, M.F. & Conrad, F.C. (1997). Does Conversational Interviewing Reduce Survey Measurement Error? Public Opinion Quarterly. 61(4) 576-602.

Schuman, H. & Presser, S. (1981). Questions and Answers in Attitude Surveys. New York: Academic Press.

Schwarz, N. (1996). Cognition and Communication: Judgmental Biases, Research Methods, and the Logic of Conversation. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Schwarz, N. & Hippler, H.J. (1991). Response Alternatives: Impact of their Choice and Presentation Order. In Biemer, P., Groves, R., Lyberg, L.E., Mathiowetz, N.A., & Sudman, S. (Eds.), (pp. 41-56). Measurement Errors In Surveys. New York, John Wiley & Sons Inc.

Tourangeau, R. & Rasinki, K.A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. Psychological Bulletin, 103, 299-314.