Journal of Official Statistics, Vol. 17, No. 4, 2001, pp. 457–477

Understanding the Cognitive Processes of Open-Ended Categorical Questions and Their Effects on Data Quality

Monica Dashen[1] *and Scott Fricker*[1]

Two studies investigated how people interpret open-ended categorical questions. The observed findings of both studies show that people do in fact misinterpret category titles and that they do so in systematic ways. The results of Study 1 indicate that people were most likely to give a false positive when they interpreted a category as including items that serve a particular goal. People were more likely to give correct (desired) responses when they thought in terms of varieties of items literally belonging to a category. Study 2 confirmed these findings. The use of supplemental instructions is recommended as a means to improve data quality in these questions.

Key words: Survey methodology; response errors.

1. Introduction

This article 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 and Schober 1992; Feldman 1992; Groves, Fultz, and Martin 1992; Martin and Polivka 1995; Schwarz 1990; Tourangeau and Rasinski 1988). Evidence suggests that people may use the accompanying list of members to clarify the intent of the question (e.g., Schwarz and 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.[2] Saving time contributes to accuracy because a respondent

[1] U.S. Bureau of Labor Statistics, Office of Survey Methods and Research, Suite 1950, 2 Massachusetts Ave., NE, Washington, DC 20212, U.S.A. Correspondence: Monica Dashen, e-mail: Dashen_M@bls.gov

[2] Many different surveys use categorical questions, including the Current Population Survey and Telephone Point of Purchase Survey (which are both sponsored by the U.S. Bureau of Labor Statistics), the Survey of Income and Program Participation (sponsored by the U.S. Census Bureau) and the National Health Interview Survey (sponsored by the U.S. National Center for Health Statistics).

Acknowledgments: Any opinions expressed in this article are those of the authors and do not constitute the policy of the Bureau of Labor Statistics. The authors are indebted to Karen Jackson, Randall Powers, Susan Schnipke, and Liz Shulman for their assistance in collecting and/or coding the data. We are also grateful to Fred Conrad, John Dixon, Jim Esposito, Bill Mockovak, Anne Polivka, Lance Rips, Brian Ross, Ernie Rothkopf, Clyde Tucker, Mick Couper and the two anonymous reviewers for their insightful comments on earlier drafts.

who must wade through scores of questions may tend to answer ''no'' more frequently simply to speed the interview (e.g., Lehnen and Reiss 1978; Sudman and Bradburn 1982). 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 (VCRs) are not in the same Telephone Point of Purchase Survey (TPOPS) category.[3] 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 (VCRs, 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 and 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 and Hippler 1991; Schwarz 1990, 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. Taking this line of reasoning one step further the list should be exhaustive (include all possible members), otherwise the respondents may not report them.

Despite the benefits of the list, there are some situations where respondents do not have the opportunity to see the list. For example, telephone respondents cannot see the list, but face-to-face respondents can (Groves and 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) have shown, respondents tend not to do.

Interestingly, even though face-to-face and self-administered survey designers have the option to show the list to the respondent, these designers may choose to omit the list as a way to decrease the time taken to complete the survey.[4] Hence, the issues regarding the designers' use of open-ended categorical questions in a survey apply to all modes of administration.

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

[3] The Telephone Point of Purchase Survey (TPOPS) uses categorical questions to collect data about consumer behavior – a way of computing the Consumer Price Index.

[4] One could even argue that even if there is a list shown in a face-to-face interview, respondents may not take the time to examine the list due to social pressure from the interviewer (e.g., the interviewer may have his or her pen posed ready to start recording).

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.[5]

The failure to infer the correct criterion in open-ended categorical questions becomes increasingly 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 telephone interview, with its more time pressured format than that of a face-to-face interview, makes people want to answer each question as quickly as possible; therefore respondents will likely not take the time to state how they decided (Schwarz, Strack, Hippler, and Bishop 1991; Dillman and Tarnai 1991; Dillman, Sangster, Tarnai, and Rockwood 1996; Rockwood, Sangster, and 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 and 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.

2. 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 theories – physical similarity, essence, and goal – for several reasons.[6] These theories are well documented in the literature and have received a lot of attention; they also offer identifiable and discernable 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.) 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 describes these theoretical predictions.

As seen in Column 1 of Table 1, the physical similarity proponents argue that people strictly decide category membership based on an item's physical resemblance to other category members (e.g., Tversky and Gati 1978; Brooks, Norman and Allen 1991; Medin 1989; Nosofsky 1991). In recent years, considerable attention has been focussed on

[5] The authors will use the phrase ''open-ended categorical questions'' to describe those categorical questions not accompanied by a list.

[6] The processes involved in categorization is akin to those processes involved in formulating a criterion of inclusion.

Table 1. Summary of theoretical predictions

Explanation	Methods of interpretation for open-ended categorical questions		
	Physical similarity (*Column 1*)	Essence (*Column 2*)	Goal (*Column 3*)
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-piece bodice and skirt belong. For example, a *sundress*, *a T-shirt dress* and a *cocktail dress* are members, whereas a *matching top* and *skirt* is not a member.	Dresses that share the implied formality generally associated with a work place or special occasion belong.* For example, *gowns*, *business dresses*, and *cocktail dresses* are members, whereas *sun dresses* and *T-shirt dresses* are not members.	Dresses and accessories that serve the goal ''of getting dressed,'' belong. For example, *dresses*, *scarves*, *belts*, and *shoes* are members, whereas, *family photos* and *important documents* are not.

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

the importance of the essence (Table 1, Column 2). 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., Malt 1994; Medin and Ortony 1989; Putman 1975; 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, 1991; Lynch, Coley, and Medin 2000). 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 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.

In summary, we have cited three major methods of categorization from the field of psychology. Although this list is hardly exhaustive, these methods offer discernible predications and are applicable to consumer-oriented categories such as those used in the TPOPS.

Psychologists have also examined whether respondents use multiple methods to interpret a category title, or only one (Barsalou 1983; Murphy 1993; Ross and Murphy 1999). If people use only one method to interpret the category title, they tend to stick with that method for that category and justify all responses according to that method. As an example, when asked about *Coffee*, which is in fact one of the things we ask about, people may report all items that accompany coffee (e.g., *milk*, *cookies*, or *sugar*) and mention that these things go with coffee or accompany it. Here reporting both the items and the decisions behind the reports may very well be a form of goal-oriented thinking. In this case, the respondents only use a single method to interpret the category.

In contrast to the single method, the multi-method approach suggests that people use more than one interpretation of a category to formulate their responses. To illustrate this point, let us return to the *Coffee* category. People may say that *water* is a member of the *Coffee* category because it is necessary for the goal of making coffee. However, people may also include *coffee beans*, using the rationalization that these contain caffeine (rather than basing their decision on the fact that beans are necessary for making coffee). In both cases, people use both goal-oriented and essence thinking, to generate exemplars for a single category. The multi-method approach is consistent with the cross-classification finding (the tendency to use more than one method to generate instances of a particular category) reported by Ross and Murphy (1999).

3. Aims of the 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. If in fact the respondents do, then the further aim is to determine whether respondents formulate more than one criterion for each question.

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.

4. Overview of Studies

We conducted two studies to identify the types of criterion of inclusion used for open-ended categorical questions. These questions pertained to food, clothing, and computers, such as those employed by the Telephone Point of Purchase Survey (TPOPS). In Study 1, respondents were given category titles and asked to generate items they thought belonged in the categories and to give the reasons for their decisions. Two independent coders later classified the Study 1 respondents' answers into four groups of interpretations (e.g., ''to make'' goal-oriented thinking). For example, a respondent might have justified including *creamer* in the *Coffee* category by saying that she takes *creamer* with her coffee every morning. Such a response would have been interpreted as a ''goal'' response (e.g., to accompany one's coffee). There is the possibility of reasonable people interpreting these justifications differently. In other words, what seems like a ''goal'' justification to one person might seem like an ''essence'' justification to another.

To lessen any possible effects due to subjective interpretation, Study 2 differed from Study 1 in that Study 2 provided respondents with reasons. Thus, a respondent given the category *Coffee* might have been asked to name all the things used to serve the goal of ''to accompany'' coffee. Taken together, the results of the two studies help us understand how respondents are interpreting the open-ended TPOPS categories.

We are particularly interested in their reasons for their decisions because different categorization methods may lead to the same response. For example, given the category *Bicycles and Accessories*, people might respond with anything that physically resembles a bicycle, such as a mountain bicycle, and thus follow the physical similarity principle. Alternatively, they might also think of all things needed to take on a bicycle trip and respond with a mountain bike and thus follow the goal-oriented principle. The end goal is the means by which respondents arrive at responses, not the responses themselves. This point is important since it is not the case that different interpretations necessarily lead to mutually exclusive and disjoint responses. The purpose of this study is not so much to determine how accurate people are in providing the desired responses but instead to determine the reasons behind any inaccuracies to help survey designers increase the accuracy of responses.

5. 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.

Obviously, the type of categories varies greatly in any survey. For example, consumer-oriented survey categories can range from the tangible (e.g., clothing) to the intangible (e.g., shoe repairs). For the purposes of this article, we focus primarily on tangible

categories because they are more concrete and often examined in the categorization literature (e.g., Kalish 1995).

5.1. Method

5.1.1. Participants

Twenty-two participants responded to an advertisement in a local newspaper and received 25.00 USD each in compensation for their participation. The participants' mean age was 49, and their average educational level was 16 years of schooling (or a college degree).

5.1.2. Materials

Each participant 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 (1) *Women's Outerwear*. 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.

Having respondents write down their responses (as in a self-administrated survey) enables those respondents to spend an unlimited amount of time generating items, and more importantly to describe why these items belong to the category. Any other mode of administration (e.g., face-to-face or telephone survey) would only serve to limit the amount of data collected because people may curtail their responses and justifications, as a means of speeding up the interview (Schwarz et al. 1991; Dillman and Tarnai 1991; Dillman et al. 1996; Rockwood et al. 1996). For this reason, this written task allows us to best address our goal of understanding how people formulate a criterion of inclusion for open-ended questions.

To further benefit our goal, the respondents were told to write down their justifications

immediately following the recording of the items, as required in a retrospective think-aloud task (Ericsson and Simon 1993). An advantage of a retrospective think-aloud task is that the thought process itself does not interfere with the process of the response formulation, as it might in concurrent think-aloud tasks where people are asked to think aloud while formulating their responses. A disadvantage of a retrospective think-aloud task is that people may forget what their thought processes were in formulating their answers due to the delay in completing the task. However, the current task accounts for this issue by minimizing the delay as much as possible.

5.1.3. Procedure

Respondents were instructed to interpret the open-ended categorical questions as hypothetical. One of the questions, for example, read: ''Hypothetically, if you were to have made a purchase from the category, *Coffee*, within the last two weeks, what items would you have purchased? Write down each purchase in the space provided below.'' Thus, respondents were not limited by their actual purchases in listing items within the categories. In addition, respondents were encouraged to say more than one item for each category. For each item generated, respondents were also asked to write down why they thought it was a member of the category. These responses are called ''justifications.''

As a means of emulating a telephone survey, respondents were instructed to complete the survey in a sequential order. The respondents were not allowed to skip ahead to questions, nor were they allowed to return to previously answered questions. While completing the task, the respondents were monitored by the experimenter to insure that they complied with the instructions. In keeping with the TPOPS methodology, respondents were not told that the same answer could not be used for two different categories.

5.2. Results and 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.

5.2.1. 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.

5.2.2. 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 on the TPOPS cue sheet), and (c) unintended exemplars

(reported by the respondent but not on the TPOPS cue sheet).[7] Using the intended exemplar and unintended exemplar counts, two proportional measures of performance were calculated: intended exemplar rate and unintended exemplar rate.[8]

There were a total of 692 items recorded for all twelve categories across all respondents. The average number of items recorded per category across respondents was 57.66 (692/12). Because some categories are more broadly defined than others, it is conceivable that people might have written down more items for one category than another category. The number of items per category ranged from the lowest – 40 items – assigned to the *Men's Sport Coats and Suits* category, to the highest – 93 items – assigned to the *Personal Computers and Peripheral Equipment* category.

5.2.3. 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 ''to 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.

It becomes necessary at this point to review the rationale behind the coding scheme. After inspecting the justifications, the first author developed the above mentioned coding scheme. Using this coding scheme, two judges who were blind to the nature of the study classified all responses into four major groups. If the responses were unclassifiable or no reasoning was provided, the responses were classified into two additional groups (uncodable and unjustified). The response classifications of the two judges were correlated at .90, as a means of estimating reliability. There were a total of 692 reports. As an additional measure of agreement between the two judges a Kappa was computed to correct for chance between raters. The Kappa yielded an identical value ($K = .90$), as did the

[7] Examples of intended exemplars included (1) *raincoats* for *Women's Outerwear*, (2) *spreadsheets* for *Computer Software*; and (3) *decaffeinated coffee* for *Coffee*. Examples of unintended exemplars included: (1) *scarf* for *Women's Outerwear*, (2) *printers* for *Software* and (3) *sugar* for *Coffee*.

[8] The intended exemplar rate performance was defined as: p (intended exemplar) $= i/T$, where i is the number of intended exemplars between the exemplar generation task and the cue list and T is the total number of items reported in the category. The unintended exemplar rate performance was defined as: p (unintended exemplars) $= (u/T)$, where u is the number of unintended exemplars in the exemplar generation task and T is identical to that of the intended exemplar rate.

Pearson correlation.[9] In all cases, the second judge's codes were maintained for the justification analyses reported below.[10]

5.2.4. 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 the TPOPS category cue list.[11] 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.

5.2.5. 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. These rates are summarized in Table 2. 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). It should be noted that reporting unintended exemplars is not as serious as failing to report items in a category. If a respondent reports an unintended exemplar, it can be recoded into the appropriate category during the editing process. In contrast, because not all of the items belonging to a category will necessarily occur to the respondents, some expenditures may be under-estimated. As Table 2 shows, omissions and unintended exemplars did occur. (Note: the omissions are reflected in the somewhat low intended exemplar rates.) 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 ($F(2,173) = .94$, $p = .40$).

Table 2. Study 1: Mean intended exemplar and unintended exemplar rates by category type

Category type	Intended exemplar rate	Unintended exemplar rate
Food (6)	.51	.49
Clothing (4)	.45	.55
Computers (2)	.57	.43

Note: The numbers in the ( )'s refer to the number of categories.
Note: The mean intended exemplar rates are across all justification categories including uncodable and no justification.

[9] According to Fleiss (1981), values of Kappa above .75 represent excellent and values from .40 to .75 represent fair to good agreement above chance.
[10] The category type reliabilities between judges are as follows: (1) Food category $r = .89$, $K = .87$, and $N = 344$; (2) Clothing category $r = .93$, $K = .96$, and $N = 212$ and; (3) Computer category $r = .87$, $K = .87$, and $N = 136$.
[11] This analysis is important because it looks at overall performance and most closely resembles the actual TPOPS interview.

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 two sections address this question.

5.2.6. Prevalence of justifications

The analyses of respondents' justifications were designed to answer the following question: Are there some interpretations, which are more frequently used than others? The percentages of responses classified within each of the four general groups of interpretations (''to make'' goal, ''to accompany'' goal, ''literal,'' and ''essence'') were calculated by category type; these results, which are summarized in Table 3, help to answer the question.

In order to find out which method people resort to most frequently when they justify an item's membership of a particular category, we compared the frequencies of all four methods (as contained in the entry totals for each method in Column 7 of Table 3) using an adjusted probability level chi-squared analysis (Agresti 1990). (Note: the unjustifiable and uncodable responses were not considered in these analyses.) The results indicate that people had adopted the ''literal list'' method more often than the other methods, as signified by various adjusted probability chi-squared comparisons.[12]

In summary, the results indicate that people consistently followed a particular method of interpretation and rule out the possibility that people interpreted categories randomly or through guessing. These methods of interpretation have been documented by other researchers in the area of categorization (e.g., Barsalou 1991).

Interestingly, people did not adopt the ''physical similarity'' method (things that look alike go together) as a means of identifying what belonged in the category.[13] Instead, people adopted the ''literal'' method, which to the best of the authors' knowledge has not been documented in the psychological literature on categorization. A further understanding of underlying mechanisms of this method would be a productive avenue for future research.

5.2.7. Category accuracy and justifications

The analyses of category accuracy and respondents' justifications were designed to answer three 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 or 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

[12] The three relevant chi-squares are as follows: (a) Literal (36 percent) versus To Make (13 percent): $x^2(1) = 111.60$, $p < .001$; (b) Literal (36 percent) versus To Accompany (23 percent): $x^2(1) = 19.95$, $p < .001$; (c) Literal (36 percent) versus Essence (9 percent): $x^2(1) = 110.1$, $p < .001$.

[13] Although our justification-coding scheme accounted for the physical similarity interpretation, we did not mention the results of this interpretation because nobody used it to generate instances.

Table 3. Study 1: Percentage of responses in each scoring category by accuracy

Justifications	Categories						
	Food		*Clothing*		*Computer*		
	Intended exemplars (*Column 1*) percent	Unintended exemplars (*Column 2*) percent	Intended exemplars (*Column 3*) percent	Unintended exemplars (*Column 4*) percent	Intended exemplars (*Column 5*) percent	Unintended exemplars (*Column 6*) percent	*Totals* (*Column 7*) percent
To Make	2.7 [4]	19.9 [39]	1.3 [1]	3.0 [4]	6.3 [5]	15.8 [9]	9 [62]
To Accompany	5.4 [8]	38.8 [76]	2.5 [2]	28.0 [37]	22.8 [18]	29.8 [17]	23 [158]
Literal	62.2 [92]	13.3 [26]	60 [48]	22.0 [29]	51.9 [41]	21.1 [12]	36 [248]
Essence	4.7 [7]	6.1 [12]	20 [16]	16.7 [22]	2.5 [2]	7.0 [4]	9 [63]
No Justification	12.2 [18]	9.1 [18]	6.3 [5]	6.8 [9]	8.9 [7]	8.8 [5]	9 [62]
Uncodable	12.8 [19]	12.8 [25]	10 [8]	23.5 [31]	7.6 [6]	17.5 [10]	12 [99]
Totals	*100 [148]*	*100 [196]*	*100 [80]*	*100 [132]*	*100 [79]*	*100 [57]*	*100 [692]*

Note: The numbers in the brackets represent the count per cell or in some cases per column.

that a person is incorrect because he or she produced an unintended exemplar, what strategy did he or she use? (3) Do individuals often use more than one method of interpreting a single question? The percentages of responses classified within each of the four general groups of interpretations (''to make'' goal, ''to accompany'' goal, ''literal,'' and ''essence'') were calculated by intended exemplars, unintended exemplars, and category type; these results, which are summarized in Table 2, help to answer the first two questions.

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 3, Columns 1, 3, and 5) were compared across all four methods using an adjusted probability level chi-squared analysis (Agresti 1990). (Note: the unjustifiable and uncodable responses were not considered in these analyses.) The results indicate that when people adopted the literal list method, they were most likely to generate intended exemplars. (See Table 4.)

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 3, Columns 2, 4, and 6) were compared across all four methods using an adjusted probability level chi-squared analysis (Agresti 1990). (Note: the unjustifiable and uncodable responses were not considered in these analyses.) 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. (See Table 5.)

The goal of the third question is to find out whether people adopt only one interpretation for each question or more than one interpretation for each question during the exemplar generation task. That is to say, the intent of this analysis is to detect any cross-classification effects (the tendency to use more than one method to generate instances of a particular category) observed by Ross and Murphy (1999). To achieve this goal, we computed the number of respondents who used only one interpretation per category (single responses) and the number of respondents who used more than one interpretation per category (multiple responses).[14] The percentages of single and multiple responses across categories were compared using a chi-square. Respondents were more likely to stick to a single method of interpretation (64 percent or 346/540) than to use more than one method of interpretation when listing items for each category (36 percent or 194/540): $x^2(1) = 42.79$, $p < .001$.

This failure to observe a cross-classification effect seems at first glance to be inconsistent with Ross and Murphy's (1999) findings. However, the methodologies of the present and their studies differ significantly, and the inconsistency may well be explained by differences in methodology. First, they asked respondents to place items into categories,

[14] When we speak about single responses, we mean that an individual sticks to one interpretation for a particular category. This individual could use different interpretations for other categories. For example, a respondent may stick to the ''literal'' interpretation for generating items in the *Cereal* category, but that same respondent may use an ''accompany'' interpretation for the *Coffee* category. In contrast, when we speak about multiple responses, we mean that a person uses more than one interpretation for a category, but could use different interpretations for other categories.

Table 4. Study 1: Chi-squared comparisons of justifications for intended exemplars by category type

Justification comparisons	Food categories	Clothing categories	Computer categories
Literal vs. To Accompany	$x^2(1) = 70.6$, $p < .001$	$x^2(1) = 42.32$, $p < .001$	$x^2(1) = 1.29$, $p = .26$
Literal vs. Essence	$x^2(1) = 72.98$, $p < .001$	$x^2(1) = 16.00$, $p < .001$	$x^2(1) = 35.37$, $p < .001$
Literal vs. To Make	$x^2(1) = 80.67$, $p < .001$	$x^2(1) = 45.08$, $p < .001$	$x^2(1) = 28.17$, $p < .001$
To Accompany vs. Essence	$x^2(1) = .07$, $p = .80$	$x^2(1) = 10.89$, $p < .001$	$x^2(1) = 12.80$, $p < .001$
Essence vs. To Make	$x^2(1) = .82$, $p = .37$	$x^2(1) = 13.24$, $p < .001$	$x^2(1) = 1.29$, $p = .26$
To Accompany vs. To Make	$x^2(1) = 1.33$, $p = .25$	$x^2(1) = .33$, $p = .57$	$x^2(1) = 7.34$, $p < .007$**

Note: Because the frequencies of intended exemplars were compared across all four justifications using a chi-squared analysis, we computed an adjusted p-value using the Bonferroni method. The adjusted p-value of .002 was computed by dividing the probability level of .05 by 18 (which is the number of justification comparisons (6) times the number of category types (3)). After computing the adjusted p-value, we then compared it to each of the significant non-adjusted p-values in the table above. As it turns out, even though the justifications were compared in multiple ways, their p-values are still well below the adjusted value of .002 – with one exception. This exception was the ''To Accompany versus To Make'' justification comparison for the *Computer* category (as denoted by the double asterisks in the table above).

whereas the present work presented respondents with categories and asked the respondents to supply items. Second and more important is the fact that the categories in Ross and Murphy's study were more natural, whereas the present study employed more unusual (or data-user oriented) categories such as those found in TPOPS.[15]

6. 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 by asking people to generate items using one of four interpretations. 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,'' then the findings of Study 1 are considered validated.

[15] Though the present work was not explicitly designed to test for Part-Whole effects, we performed a chi-squared analysis patterned after Schuman and Presser (1981) to detect any such effects in two category combinations involving Men's Clothing and Computers. (See Strack 1992 and Schwarz 1996 for more details about the Part-Whole effect.) A Part-Whole effect was not observed in either combination. This finding is not surprising given that the target categories were not adjacent to each other, as they would be in most studies that examine this effect. Instead, there were intervening categories between the target categories. Conceivably, the larger the number of intervening categories, the less likely it is that people will remember their answers to the first question; therefore they might not adopt a different interpretation for the second category, as suggested by Part-Whole advocates.

Table 5. Study 1: Chi-squared comparisons of justifications for unintended exemplars by category type

Justification comparisons	Food categories	Clothing categories	Computer categories
Literal vs. To Accompany	$x^2(1) = 24.51$, $p < .001$	$x^2(1) = .970$, $p = .33$	$x^2(1) = .862$, $p = .35$
Literal vs. Essence	$x^2(1) = 5.16$, $p < .05$**	$x^2(1) = .961$, $p = .33$	$x^2(1) = 4.00$, $p < .05$**
Literal vs. To Make	$x^2(1) = 2.60$, $p = .11$	$x^2(1) = 18.94$, $p < .001$	$x^2(1) = .429$, $p = .52$
To Accompany vs. Essence	$x^2(1) = 46.55$, $p < .001$	$x^2(1) = 3.8$, $p < .05$**	$x^2(1) = 8.05$, $p < .005$**
Essence vs. To Make	$x^2(1) = 14.30$, $p < .001$	$x^2(1) = 12.50$, $p < .001$	$x^2(1) = 1.92$, $p = .17$
To Accompany vs. To Make	$x^2(1) = 11.90$, $p < .001$	$x^2(1) = 26.56$, $p < .001$	$x^2(1) = 2.46$, $p = .12$

Note: Because the frequencies of unintended exemplars were compared across all four justifications using a chi-squared analysis, we computed an adjusted *p*-value using the Bonferroni method. The adjusted *p*-value of .002 was computed by dividing the probability level of .05 by 18 (which is the number of justification comparisons (6) times the number of category types (3)). After computing the adjusted *p*-value, we then compared the value to each of the significant nonadjusted *p*-values in the table above. As it turns out, even though the justifications were compared in multiple ways, their *p*-values are still well below the adjusted value of .002 – with four exceptions. These exceptions are denoted by a double asterisk (**) in the table above.

6.1. Method

6.1.1. Participants

Forty-five paid volunteer participants (fifteen participants per condition) responded to an advertisement in a local newspaper and received 25.00 USD in compensation for their participation. The participants' mean age was 44, and the average educational level was 15.67 years of schooling.

6.1.2. Procedure

Study 2 involved an exemplar generation task. This task was identical to that of Study 1 with one exception. Respondents were asked to adopt one of four interpretations in this task. The instructions defined and provided an example of each of the four possible interpretations.[16] The four interpretations were as follows:

(1) **Accompany-Goal** group: Participants were asked to generate a list of ''things that people might purchase to accompany'' the category title items. For example, if the category was *Coffee*, respondents could be expected to list items like *milk*, *creamer*, *sugar*, *spoons*, *cookies*, *cakes*, *donuts*, or any other items that might conceivably accompany coffee.[17]

[16] To make the instructions more concrete to the reader, we provide examples of what the respondents actually said. However, the reader should be aware that the actual respondents were given hypothetical examples (e.g., telephones and soup).

[17] Respondents also might reasonably have listed *coffee filters* in this condition despite the fact that the next condition would be a more reasonable place for *filters*.

(2) **Make-Goal** group: Participants were asked to generate a list of ''things that people might purchase to make or use'' the category title items. For example, for the *Coffee* category such items as *filters*, *coffee grounds* or *beans*, *coffee makers*, and *water* might have been listed.
(3) **Essence** group: Participants were asked to think of the ''essence'' of the category title and to write down that essence. The ''essence'' is defined as an underlying characteristic of the category and represents the primary function or intended use of the category items. For example, the ''essence'' of the category *Men's Outerwear* might be ''keeping people warm'' or, more broadly, ''protecting people from the elements.'' Thus, a participant describing *Men's Outerwear* might have listed three possible essences but would not have listed *parkas*, *sweaters*, or other garments.
(4) **Literal** group: Participants were asked to list types of category purchases (or those that are fairly characteristic of the category). The participants were instructed to avoid product names except for those product names that have become entirely or nearly synonymous with a type of product (e.g., Band-Aid or Kleenex). For example, a participant listing items for the *Coffee* category might list *cappuccino*, *decaffeinated coffee*, *instant coffee*, *espresso*, and *ground coffee* but was asked not to list *Taster's Choice* or *Folger's* as these are brand names.

6.2. *Results and discussion*

6.2.1. Validation of responses

To find out whether the findings of Study 2 validate those of Study 1, we compared the items generated under the same condition in Study 2 and justification in Study 1. The results of this analysis are shown in Table 6 below.

Three points about the analyses are worth mentioning. First, we collapsed across categories and grouped the items into category types. Second, people in Study 2 generated far more items than did people in Study 1. Third, we excluded duplicate responses so that if two people mentioned item X, that item was counted only once. Both intended and unintended exemplars were combined in this analysis.

The findings in Table 6 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 percent). 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 65 percent in the

Table 6. Percentage of responses in justification categories of Study 1 that were also in Study 2

Justifications	Food percent	Clothing percent	Computers percent
Goal (To Make)	88 (35/40)	100 (9/9)	79 (11/14)
Goal (To Accompany)	89 (31/35)	85 (17/20)	73 (16/22)
Essence	65 (11/17)	100 (11/11)	67 (4/6)
Literal	88 (28/32)	88 (22/25)	77 (20/26)

food 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 65 percent 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.

7. 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 or 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 the 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.

Study 1 respondents provided intended exemplars, usually when they adopted the literal strategy. However, the success of the literal strategy depended upon a category being narrowly defined and aptly named. While this strategy worked well, there were times when category names were broader than the list of correct cues would seem to warrant. For example, the label *Men's Outerwear*, intended to describe winter and rain jackets, led respondents to list all manners of clothing that were worn and not hidden (i.e., everything except underwear). With one exception, other means of justification seem ill suited to respondent accuracy. The ''to accompany'' goal method seemed well suited for the *Personal Computers* and *Peripheral Equipment* category. This method's suitability is probably due to the fact that many of the category members did in fact accompany a computer in the respondents' homes. For example, many of the respondents probably own printers with their computers.

The present work has important practical implications. Conscientious survey designers are naturally interested in obtaining predictable, or ''correct,'' responses. The results of the present work strongly suggest that category titles alone do not necessarily attain

that end. What may be an obvious interpretation to designers may not be obvious to respondents. Perhaps more important is the fact that two respondents, faced only with a category title, may use different interpretations and therefore provide different answers. This differential interpretation of the same question will cause measurement error (Fowler 1993). In other words, unless category titles are supplemented with some sort of instructions about how they should be interpreted, survey designers are almost certain not to get consistent results. Survey designers could provide a lead-in statement to assist the respondents in understanding a category's desired method of interpretation. Such a statement naturally would have to take into account the nature of the category. For example, a category defined simply by instantiations of the title (e.g., *Bread*) would be better served by an instruction telling respondents to think of literal types or varieties of the title. On the other hand, a category that the TPOPS survey designers have constructed to include more than simple instantiations of the title (e.g., the computer category, which includes components) requires broader instructions about how to interpret the category. However, such instructions can be problematic by leading to larger numbers of ''false positives'' as the leeway for responses is increased. In other words, any instruction that tells people to do more than simply list items that fit the category title literally necessarily invites a certain amount of interpretation, which may differ from the designers' intent.

While the focus of this work is on consumer-oriented categories, lead-in statement usage should be readily adaptable to a significant number of other types of open-ended categorical questions. For example, if a question is narrowly defined, it stands to reason that a lead-in sentence should involve a literal interpretation. In addition, the methodology of the present work may provide a vehicle by which to develop lead-in sentences for other types of categories. That is to say, measuring both the items people include in categories and the reasons for their decisions may provide insight into the optimal interpretation of the category (the interpretation that leads to the largest number of correct reports).

Another practical solution for the false positive problem might be to ask respondents whether they bought any *Men's Outerwear*, for example, and to follow-up with, ''What did you buy?'' This follow-up question avoids the presentation of a lengthy list, while providing a check on whether the purchase should be assigned to the category. This follow-up question also offers a practical solution for telephone survey designers who often use a ''yes/no'' response format to avoid a lengthy interview at the expense of not providing a check as to whether the purchase is a member of the category. Though the follow-up question increases interview time, it provides a check on whether the purchase is a member of the category. Obviously, however, this solution does not address the problem of ''omissions,'' which may be larger than the problem of ''false positives.''

On a theoretical note, the finding that people tend toward the literal method when responding might be attributable to any number of factors. For example, one explanation is that people simply ascribe literal interpretations to language as a general position unless otherwise indicated from context. As a general rule, for example, a person says, ''I bought a stereo,'' only if he or she in fact bought a stereo. A person who bought a compact disc or a pair of headphones would not identify the purchase as a stereo. Another explanation may be that the question structure can affect how people interpret the question. If the question is fairly narrow in nature and definition, as in *Coffee*, people may narrowly

interpret this case in a literal sense and only say types of coffee. In contrast, if the question is broad in nature and definition, as in *Men's Outerwear*, people may interpret in a more general way.

Recent experiences might well influence interpretations of category titles and instructions (see Baddeley 1990, for a review on the recency effect). For example, a person may supply *cream* as a response to the *Coffee* question because he or she just bought a cup of coffee and added some cream to it.

7.1. Future research

The present work is a step toward understanding how people interpret open-ended categorical questions. Future work aimed at examining how the three factors – survey goals, respondent expertise, and survey mode of administration – influence respondents' interpretations would be a fruitful avenue. First, the interpretation of a given category will depend on the perceived goals of the survey (Schwarz 1996). If *Coffee* were presented in the context of questions about substance abuse, for example, *creamer* would probably not be mentioned. But an interpretation that includes *creamer* makes sense in the context of questions about purchases, where the question *Coffee* may be construed as ''coffee-related purchases.'' Thus the frequency of different types of intrusions and justifications is bound to be a function of the interpretation that the context allows for. Second, the interpretation, and the type and number of items generated, will also depend on the respondents' level of knowledge or expertise about a particular topic. When asked about *Photographic Equipment*, a professional photographer would be more likely to list the different types of *cameras*, *lenses* and *film* than would an amateur photographer. Future research on the expertise topic should follow that of Alba and Hutchinson (1987) and Bickart (1992).

Third, the quality of the respondents' interpretations may also depend upon the mode of survey administration. Obviously, the more rushed respondents are in formulating a response, which is often the case in telephone surveys, the less likely it is that they will develop an appropriate criterion of inclusion. Future research on this topic should follow that of Schwarz et al. (1991) and Dillman et al. (1996). Taking these three factors together, a further understanding of how people interpret open-ended questions should improve data quality in the future.

8. References

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

Alba, J.W. and Hutchinson, J.W. (1987). Dimensions of Consumer Expertise. Journal of Consumer Research, 13, 411–454.

Baddeley, A. (1990). Human Memory: Theory and Practice. Hillsdale, N.J.: Erlbaum.

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

Barsalou, L.W. (1991). Deriving Categories to Achieve Goals. In The Psychology of Learning and Motivation: Advances in Research and Theory, ed. G.H. Bower, 27, 1–64, New York: Academic Press.

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

Bickart, B. (1992). Question-Order Effects and Brand Evaluations: Moderating Role of Consumer Knowledge. In Context Effects in Social and Psychological Research, eds. N. Schwarz and S. Sudman, 63–80, New York: Springer-Verlag.

Brooks, L.R., Norman, G.R., and Allen, S.W. (1991). Role of Specific Similarity in a Medical Diagnostic Task. Journal of Experimental Psychology: General, 120, 278–287.

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

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

Dillman, D. and Tarnai, J. (1991). Mode Effects of Cognitively Designed Recall Questions: A Comparison of Answers to Telephone and Mail Surveys. In Measurement Errors in Surveys, eds. P. Biemer, R. Groves, L.E. Lyberg, N.A. Mathiowetz, and S. Sudman, 73–94, New York: John Wiley & Sons, Inc.

Dillman, D., Sangster, R.L., Tarnai, J., and Rockwood, T.H. (1996). Understanding Differences in People's Answers to Telephone and Mail Surveys. In Advances in Survey Research. eds. M.T. Braverman and J.K. Slater, 70–110, San Francisco: Jossey-Bass, Inc.

Ericsson, K.A. and Simon, H.A. (1993). Protocol Analysis: Verbal Reports as Data. London: MIT Press.

Feldman, J.M. (1992). Constructive Processes in Survey Research: Explorations in Self-Generated Validity. In Context Effects in Social and Psychological Research, eds. N. Schwarz and S. Sudman, New York: Springer-Verlag.

Fleiss, J.L. (1981). Statistical Methods for Rates and Proportions, (2nd Edition). New York: John Wiley & Sons.

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

Groves, R.M., Fultz, N.H., and Martin, E. (1992). Direct Questioning About Comprehension in a Survey Setting. In Questions About Questions, ed. J.M. Tanur, 49–60, New York: Russel Sage Foundation.

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

Kalish, C.W. (1995). Essentialism and Graded Membership in Animal and Artifact Categories. Memory and Cognition, 23, 3, 335–353.

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

Lynch, E.B., Coley, J.D., and Medin, D.L. (2000). Tall Is Typical: Central Tendency, Ideal Dimensions and Graded Category Structure Among Tree Experts and Novices. Memory and Cognition, 28, 1, 41–50.

Malt, B.C. (1994). Water Is Not H_2O. Cognitive Psychology, 27, 41–70.

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

Medin, D.L. (1989). Concepts and Conceptual Structure. American Psychologists, 44, 12, 1469–1481.

Medin, D.L. and Ortony, A. (1989). Psychological Essentialism. In Similarity and Analogical Reasoning, eds. S. Vosniadou and A. Ortony, 179–195, New York: Cambridge University Press.

Murphy, G.L. (1993). A Rational Theory of Concepts. In The Psychology of Learning and Motivation, eds. G.W. Nakamura, R.M. Taraban, and D.L. Medin, 29, 327–359.

Nosofsky, R.M. (1991). Tests of an Exemplar Model for Relating Perceptual Classification and Recognition Memory. Journal of Experimental Psychology: Human Perception and Performance, 17, 3–27.

Putman, H. (1975). The Meaning of ''Meaning.'' In Mind, Language and Reality. Philosophical Papers, 2, 215–271, Cambridge: Cambridge University Press.

Rips, L.J. (1989). Similarity, Typicality, and Categorization. In Similiarity and Analogical Reasoning, eds. S. Vosniadou and A. Ortony, 21–59, Cambridge: Cambridge University Press.

Rockwood, T.H., Sangster, R.L., and Dillman, D.A. (1997). The Effect of Response Categories on Questionnaire Answers: Context and Mode Effects. Sociological Methods and Research, 26, 7, 118–140.

Ross, B.H. and Murphy, G.L. (1999). Food for Thought: Cross-Classification and Category Organization in a Complex Real-World Domain. Cognitive Psychology, 38, 495–553.

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

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

Schwarz, N. (1990). Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Question Construction. In Research Methods in Personality and Social Psychology, eds. C. Hendrick and M.S. Clark (pp. 98–119).

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

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

Schwarz, N., Strack, F., Hippler, H.J., and Bishop, G. (1991). The Impact of Administration Mode on Response Effects in Survey Measurement. Applied Cognitive Psychology, 5, 193–212.

Strack, F. (1992). Order Effects in Survey Research: Activative and Informative Functions of Preceding Questions. In Context Effects in Social and Psychological Research, eds. N. Schwarz and S. Sudman, 23–34, New York: Springer-Verlag.

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

Tourangeau, R. and Rasinski, K.A. (1988). Cognitive Processes Underlying Context Effects in Attitude Measurement. Psychological Bulletin, 103, 299–314.

Tversky, A. and Gati, I. (1978). Studies of Similarity. In Cognition and Categorization, eds. E. Rosch and B.B. Lloyd, 79–98, Hillsdale, NJ: Earlbaum.

Received October 1999
Revised March 2001