Framing effects in surveys:

How respondents make sense of the questions we ask

Wändi Bruine de Bruin

Carnegie Mellon University

Department of Social and Decision Sciences

Department of Engineering and Public Policy

Chapter prepared for

Keren, G. (Ed.) *Perspectives on framing*. London, UK: Taylor & Francis.

Author note:

The author would like to thank Kirstin Appelt, Steve Atlas, Shahzeen Attari, Martine Baldassi, Bernd Figner, Baruch Fischhoff, Lauren Fleishman, David Hardisty, Eric Johnson, Gideon Keren, Maria Konnikova, Irwin Levin, Ye Li, Jenn Logg, Annie Ma, Juliana Smith, Katherine Thompson, Elke Weber, and Julie Zelmanova for their thoughtful comments on an earlier draft of this chapter, as well as Mandy Holbrook for research assistance.

Please address correspondence to Wändi Bruine de Bruin, Carnegie Mellon University, Department of Social and Decision Sciences, 5000 Forbes Avenue, Porter Hall 208, Pittsburgh PA 15213; wandi@cmu.edu (email); 412-268-3237 (phone); 412-268-6938 (fax)

Framing effects in surveys:

How respondents make sense of the questions we ask

In many countries, national surveys are conducted to learn about people's attitudes, expectations, and behaviors in a wide variety of domains. In the United States, well-known surveys include the Gallup Poll, the Consumer Expenditure Survey, the National Longitudinal Study of Youth, the Health and Retirement Study, and the Michigan Survey of Consumers. Other countries have their own national surveys, recruiting respondents from the general population.

Writing good survey questions may seem deceptively simple. A vast body of survey design research suggests that even slight variations in wording, response options, and question order can affect responses (e.g., Bradburn, 1982; Converse & Presser, 1986; Dillman, Smyth, & Christian, 2009; Schwarz, 1996, 1999; Sudman & Bradburn, 1982). In essence, these response inconsistencies are not all that different from the framing effects reported in the judgment and decision-making literature. Indeed, reviews of judgment and decision-making research sometimes include references to survey studies (e.g., Fischhoff, 1991; Hogarth, 1982).

Although researchers in judgment and decision making and researchers in survey design may have shared interests, their overall goals tend to be different. A main goal of behavioral decision making research is to understand the conditions under which people deviate from normative principles. One of these normative principles is *description invariance*, which posits that preferences should be unaffected by irrelevant variations in how a problem is presented. To examine when and why people may violate description

invariance, researchers in judgment and decision making have carefully crafted pairs of hypothetical decision problems that differ only slightly in their wording, yet can produce seemingly conflicting responses. Indeed, researchers have honed in on those decision problems that cause the largest violations of description invariance (Fischhoff, 1991; Kühberger, 1998).

Traditionally, researchers have studied *risky choice framing problems*, which present choice options with uncertain outcomes, and *attribute framing problems,* which ask participants to evaluate a single option (Levin, Schneider, & Gaeth, 1998). In each case, participants are randomly assigned to a positively worded or a negatively worded problem version. For example, a classic risky choice problem is the well known "Asian disease" (Tversky & Kahneman, 1981). When presented with the positive wording, most research participants prefer saving 200 people for sure over a gamble with a 1/3 probability of saving 600 and a 2/3 probability of saving no one. When presented with the equivalent negative wording, most prefer a gamble with a 1/3 probability that nobody will die and a 2/3 probability that 600 will die over losing 400 people for sure. A classic attribute framing problem shows that people rate ground beef as better when it is described as "80% lean" than when it is described as "20% fat" (Levin & Gaeth, 1988). Both these examples are examined in detail in several chapters of the present volume.

By contrast, most survey researchers aim to avoid response inconsistencies. Their main goal is to design clear questions that allow respondents to provide accurate reports of their beliefs, attitudes, and behaviors. Once a specific question has been used, survey researchers may even hesitate to change its wording, response options, its place in the survey, or other design elements – especially if they are following a panel of respondents

over a longer period of time (Dillman et al., 2009). Indeed, they recognize that such modifications may make it impossible to compare responses across different surveys, or interpret changes over time.

Survey researchers often discover response discrepancies by accident, when different versions of the same question appear on two different surveys. They may then conduct studies to understand these differences, and to identify the best way to measure the specific construct. In these studies, respondents are randomly assigned to one of the two question versions, and may be asked to explain their responses. To avoid affecting the continuity of longitudinal studies, these respondents may be drawn from a separate sample. If the new version is introduced to a longitudinal panel, it may first be given to a randomly selected subset of respondents, while the rest still answers the old question version. Both practices allow researchers to examine whether variations in question design may affect respondents' answers.

The theoretical framework for these survey studies typically interprets the interaction between researchers and respondents as a form of communication, which is subject to the tacit rules of everyday conversation (Grice, 1975; Schwarz, 1996, 1999). Effective communication occurs when respondents understand questions as intended by the researchers, and, in turn, give responses that the researchers understand. When designing questions, survey researchers therefore aim to be aware of the *lexical* meaning and the *pragmatic* meaning they are communicating (Schwarz, 1999). Understanding the lexical meaning requires that respondents recognize the words, and are able to retrieve their interpretation from semantic memory. Understanding the pragmatic meaning involves making accurate inferences about the questioners' intention, based on the logic

of conversation (Grice, 1975; Schwarz, 1996, 1999), assuming that survey designers adhere to (a) the *maxim of manner,* which holds that the most obvious interpretation of each question is also the correct one; (b) the *maxim of quantity,* which holds that each question asks for information that is relevant and interesting rather than assumed and already known; (c) the *maxim of relation,* which holds that each question should be interpreted in light of the previous questions that have been asked; and (d) the *maxim of quality,* which holds that each question presents correct information. It follows from these maxims that modifying a question's design may affect responses, if doing so systematically changes its interpretation.

This chapter aims to apply the logic of conversation to studies of framing effects due to variations in (a) question wording, (b) choice set, and (c) presentation order. Each section first presents survey design studies aiming to understand respondents' reasons for giving seemingly inconsistent responses. Each section then examines whether these reasons may also explain the framing effects reported in the judgment and decision-making literature. If two versions of a decision problem are interpreted as communicating different information, then seemingly conflicting response patterns may be defensible in terms of the logic of conversation. Moreover, if respondents perceive objective differences, then differential responses may not violate normative decision rules, as researchers claim (Frisch, 1993). The chapter ends with suggestions about how to improve question design, and a discussion of recent judgment and decision-making studies that have followed those suggestions.

Question wording

Studies in survey design. Communicating the lexical meaning of a question requires that researchers and respondents speak the same language. However, survey respondents are typically recruited from the general population, members of which vary widely in their reading ability. To reach people with lower reading ability, it is often recommended to write research materials at a 5th to 10th grade reading level (Paasche-Orlow, Taylor, & Brancati, 2003). Even people with higher literacy may experience less confusion when questions are easier to read. Formulas for reading ease consider the average number of words per sentence, and the average number of syllables per word (Kincaid, Fishburne, Rogers, & Chissom, 1975). Thus, "you can agree to be in the study now and change your mind later" is easier to read than "you have the right to choose not to participate or to withdraw your participation at any point in this study" (Paasche-Orlow et al., 1975). Similarly, survey questions are easier to understand when they use shorter everyday words (such as "bleeding") rather than official terminology (such as "hemorrhage") (Lerner, Jehle, Janicke, & Moscati, 2000).

Even if a survey question uses words that are easy to read, respondents do not necessarily use the interpretation that is common among experts. For example, some people may interpret the "greenhouse effect" as causing local weather changes rather than overall climate change (Bostrom, Morgan, Fischhoff, & Read, 1994), "abstinence" as including anal sex rather than meaning no sex at all (Schuster, Bell, & Kanouse, 1996), and a "pap smear" as testing for all possible sexually transmitted infections rather than "just" cervical cancer (Blake, Weber, & Fletcher, 2004). Moreover, in unpublished pilot interviews, heterosexual respondents living with a boy-friend or a girl-friend hesitated to

indicate that they are “married or living with a partner” because they interpreted that term as referring to a homosexual relationship (see also Hunter, 2005). Studies have also found varying interpretations of such seemingly simple terms as “ethnicity” (McKenney & Bennett, 1994), “family” and “unemployed” (Converse & Presser, 1986).

Furthermore, presumed antonyms may not actually communicate the exact opposite. In a classic poll, respondents answered either “Do you think that the United States should forbid public speeches against democracy?” or “Do you think that the United States should allow public speeches against democracy?” (Rugg, 1941). Of those who were asked about forbidding public speeches against democracy, 54% agreed that they should be forbidden. Of those who were asked about allowing those speeches, 75% indicated that, no, they should not be allowed. Thus, respondents were much more likely to be against forbidding anti-democratic speeches than they were to be in favor of allowing anti-democratic speeches. Subsequent studies have replicated this response pattern with questions about forbidding vs. allowing X-rated movies in cinemas (Hippler & Schwarz, 1986), military exercises near nature preserve areas (Holleman, 1999a), and many other topics (Holleman, 1999b).

Follow-up research found that the asymmetry occurs because respondents with a less extreme attitude as reported on a separate 1-7 scale will disagree with both “forbidding” and with “allowing” the same thing (Holleman, 2006). In Chapter 7 of this book, Schul provides similar examples of that response pattern, such that “not hot” is seen as less extreme than “cold” and “not including” as less extreme than “excluding.” The change in wording does indeed communicate something different, with respondents rating someone who supports “not forbidding” something as communicating less extreme

attitudes than someone who supports “allowing” it, and someone who supports “not allowing” something as communicating less extreme attitudes than someone who supports “forbidding” it (Hippler & Schwarz, 1986).

Ultimately, survey design researchers aim to determine the best way of asking the question. Their research on the forbid/allow symmetry contributes two suggestions. First, the asymmetry is less extreme when it is easier for respondents to form a determined opinion, as is the case when questions are concrete and linguistically simple (Hippler & Schwarz, 1986; Holleman, 1999b). Second, questions about forbidding may be more reliable than questions about allowing. That is, responses to questions about forbidding something tend to be relatively consistent with other questions about forbidding related things (Holleman, 2006). By comparison, respondents are more likely to switch between saying “yes” and “no” to questions about whether or not to “allow” related behaviors. These results suggest that the interpretation of “forbidding” is clearer than the definition of “allowing.” Below, the section on improving survey design describes research practices to further improve the reliability of survey questions.

Like presumed antonyms, presumed synonyms have also produced seemingly inconsistent responses. One recent example comes from surveys about inflation expectations, which tend to be followed by central banks. In the United States, the Michigan Survey of Consumers asks a monthly random sample “during the next 12 months, do you think that prices in general will go up, go down, or stay where they are now” (Curtin, 2006). Respondents who expect a change are then asked “by what percent do you think prices will go [up/down] over the next 12 months?” Similar questions

appear on economic surveys in other countries, typically asking about “prices” rather than directly asking about “inflation” (Ranyard et al. 2008)

Although the term “inflation” may be more complex than “prices in general” (Bruine de Bruin et al., 2009), people have a basic understanding of the concept (Leiser & Drori, 2004; Svenson & Nilsson, 1986). Some may even recognize that questions about “prices in general” are referring to inflation (Bruine de Bruin et al., 2009). Nonetheless, expectations for “inflation” tend to be much lower responses than expectations for “prices in general” (Bruine de Bruin et al., 2009), possibly because the two questions evoke different interpretations. The former elicits stronger thoughts about prices of things respondents spend money on, while the latter evokes stronger thoughts about the official inflation rate. Moreover, when respondents think about “prices” they spend money on, increasing ones seem more salient than decreasing ones (Kahneman & Tversky, 1979), creating an upward bias in responses. Compared to expectations for “inflation,” expectations for “prices in general” were more strongly correlated to expectations for gas and food prices, which had been increasing at the time of the survey. By contrast, no such difference was observed in correlations with expectations for housing prices, which had been decreasing at the time of the survey. Yet, researchers have not yet examined the external validity of these questions, in terms of their correlations with real-world financial decisions. It may be the case that people think about “prices in general” when making consumer decisions, and about “inflation” when making decisions about investing their money. If so, slight variations in question wording may lead respondents to make different assumptions about the intentions of the

researcher, possibly affecting responses to subsequent questions. Such effects of presentation order are discussed in a separate section below.

Studies in judgment and decision making. Researchers in judgment and decision making have also examined the effect of wording on responses to decision problems. As mentioned above, studies of the "Asian disease problem" have found that people respond differently when a choice is described in terms of lives saved or in terms of lives lost. The favored explanation is that these responses reflect loss aversion, such that the certain prospect of losing lives is more painful than the certain prospect of saving the equivalent number of lives is pleasurable (Tversky & Kahneman, 1981). However, conclusions about what drives responses to the Asian disease problem can not be drawn without fully understanding whether research participants comprehend the problem descriptions as intended (Fischhoff, 2005; Frisch, 1993; Keren & Willemsen, 2009). Although participants in judgment and decision making studies are typically recruited from undergraduate college students with good reading ability, they may still interpret the presented decision problems in ways that were not intended by the researchers.

Indeed, variations in the description of a decision problem may evoke systematically different interpretations (Frisch, 1993; Johnson, Häubl, & Keinan, 2007; Reyna & Brainerd, 1991, Shafir, Simonson, & Tversky, 1993). Participants may provide defensible justifications for interpreting two frames differently (Frisch, 1993; Mandel, 2001). For example, consider the certain prospect in the Asian disease problem, which states that 400 people will die (in the loss frame) or that 200 people will be saved (in the gain frame.) Because it is does not state what will happen to the rest of the 600 people who are exposed to the disease, some people interpret the sure option as referring to "at

least" 400 people lost or 200 saved. When the certain prospect is explicitly rephrased as "400 people will die and 200 people will not die" in the loss frame and "200 people will be saved and 400 people will not be saved" in the gain frame, response inconsistencies disappear (Kühberger, 1995).

Similar explanations may apply to attribute framing effects. For example, evaluations tend to be more positive for ground beef that is described as 80% lean rather than 20% fat (Levin & Gaeth, 1988), for condoms that are described as having a 95% success rate rather than a 5% failure rate (Linville, Fisher, & Fischhoff, 1993), and for a medical treatment with a 75% survival rate rather than a 25% mortality rate (Levin, Schnittjer, & Thee, 1988). The positive and negative frames of these problems may convey systematically different information about the researcher's frame of reference. When participants are asked to take the role of the communicator, they tend to use the positive frame to communicate a positive change from the previous reference point, and the negative frame to indicate a negative change from the previous reference point. For example, they prefer to describe a new medical treatment in terms of its survival rate if it has more survivors than the previously available alternative, and in terms of its mortality rate if it has fewer survivors than the previous alternative (McKenzie & Nelson, 2003). If communicators choose positive frames to reflect positive attitudes, and negative frames to reflect negative attitudes then the two versions are not informationally equivalent – suggesting that the reported response variations do not violate description invariance (McKenzie, 2004).

Researchers in judgment and decision making have identified many other framing effects on responses, due to alterations in wording. For example, people prefer

"insurance premiums" over "sure losses" of the same amount (Fischhoff, Slovic, & Lichtenstein, 1980; Hershey & Shoemaker, 1980), "rebates" over "deductables" with the same effect on their overall expenses (Johnson, Hershey, Meszaros, & Kunreuther, 1993), donating $1 a day over $350 a year (Gourville, 1998). Whether these seemingly inconsistent responses violate normative principles of decision making depends on whether the variations in the descriptions are interpreted as conveying objectively different meanings (Frisch, 1993). The section on improving survey design provides more information about research methods for examining these interpretations.

Choice set.

Studies in survey design. Questions can be asked with open-ended or closed-ended response modes. Open-ended questions require that researchers interpret respondents' answers. For example, consider respondents who are asked to list the things they usually spend money on. Responses such as "groceries" and "dining out" may be categorized as "food" while coding "movies" and "computer games" as referring to "entertainment." However, some responses may not perfectly map onto researchers' categories, introducing error into the coding process. For example, a respondent who mentioned "going out on dates" may have referred to going out for dinner and/or a movie, thus making it unclear whether it should be categorized as referring to "food," "entertainment," or both.

In addition to being subject to error, the coding of open-ended responses can also be very time consuming, especially when surveys have many respondents. To save time and money, survey researchers often use closed-ended questions, which ask respondents

to select the answer they would like to have given from a set of response options. However, doing so requires respondents to map their answers onto the categories – still leaving room for interpretation errors.

When response options are presented, they become part of the communication between researchers and respondents. Unintentionally, these response options may provide information about how to best answer the question. Hence, people may artificially increase their score on a multiple-choice knowledge test (Bruine de Bruin & Fischhoff, 2000). Teachers may take advantage this testing effect, because it improves students' learning of course material (McDaniel, Anderson, Derbish, & Morrisette, 2007; Roediger & Marsh, 2005). However, students may also incorrectly deduce that some of the false statements are true, as a result of acquiescence (Roediger & Marsh, 2005; Toppino & Brochin, 1989). If so, then test results are not measuring what they are intended to measure. In the classroom, testing effects will lead to a misrepresentation of what students have learned from the course. In survey research, testing effects will lead to a misrepresentation of the actual knowledge that is held in the general population.

Learning from tests may be less likely with true/false statements (Bruine de Bruin & Fischhoff, 2000) than with multiple-choice items, or with multiple-choice items that provide two response options rather than more (Haladyna, Downing, & Rodriguez, 2002). As a result, it may be preferable to present fewer response options when using closed-ended knowledge questions. However, acquiescence may be more likely with true-false statements than with multiple-choice items (Cronbach, 1941), possibly because alternative answers are not evoked (Koriat, Lichtenstein, & Fischhoff, 1980).

The presentation of response options can also affect answers to questions about beliefs and attitudes. For example, respondents who are asked about "the most important thing for children to prepare them for life" are more likely to select "to think for themselves" from a closed-ended response list, than to generate it as an open-ended response (Schuman & Presser, 1981). Moreover, respondents may report stronger dislike of a technology like carbon capture and sequestration when they rate its attractiveness alone, compared to when they rank it in a set of other options for low-carbon electricity-generation, such as nuclear power, natural gas, and renewables (Fleishman, Bruine de Bruin, & Morgan, 2009). As argued in the section on improving survey design, presented below, the chosen response mode should probably depend on the real-world decision researchers aim to understand. When making real-world decisions about child rearing strategies, parents may typically have to generate their own – making the open-ended response mode more comparable. When making real-world decisions about low-carbon technologies, however, people may choose between different ones – akin the closed-ended response mode.

Response modes also affect the seemingly inappropriate use of 50% when assessing small risks, such as getting lung cancer from smoking, or being the victim of a burglary. Although some 50% responses may reflect actual probabilistic beliefs, the majority of 50% responses may reflect not knowing what number to use (Hurd, Manski, & Willis, 2007). Use of 50% is more likely with open-ended questions than with a visual response scale ranging from 0% to 100% (Fischhoff & Bruine de Bruin, 1999), which may provide respondents with answers they would not otherwise consider. The scale may also change the logic of the conversation, by affecting respondents' interpretation of

what the question itself is asking. That is, an open-ended question may evoke verbal phrases like “fifty-fifty,” whereas the numbers on the scale encourage more numerical thinking.

Even when self-reporting on their own behaviors, which should be known to respondents, answers can be affected by the response mode. An example provided by Schwarz (1999) suggests that respondents who are asked the open-ended question “what have you done today?” will omit events that are known to the researcher (e.g., took survey) or that may be otherwise assumed (e.g., showered). Doing so follows conversational norms, according to which there is no need to report known, irrelevant or uninteresting activities. However, if those activities were presented on the response list of a closed-ended question, respondents would likely have reported them.

The survey designer’s choice of response options can also change the meaning of a question. For example, one survey asked respondents to indicate how often they watch TV on a typical day (Schwarz, Hippler, Deutsch, & Strack, 1985). Half received a high-frequency scale (on which 2.5 hours a day constituted a low response), while the other half received low-frequency scale (on which 2.5 hours a day constituted a high response). With the high-frequency scale, 37.5% of respondents indicated that they watched TV for more than 2.5 hours per day. With the low-frequency scale, 16.2% did so. Similar response patterns have been observed for self-reported sexual activity (Schwarz & Scheuring, 1988; Tourangeau & Smith, 1996), consumer choices (Menon, Raghubir, & Schwarz, 1995), and health behaviors (Schwarz, 1996). Compared to a low-frequency scale, a high-frequency scale may communicate that the researcher is asking respondents to self-report a behavior that typically occurs more often. Low-frequency behaviors may

evoke more strict definitions than do high-frequency behaviors (Schwarz, Strack, Müller, & Chassein, 1988). For example, the low-frequency scale may lead respondents' to narrowly define "watching TV" as actually sitting down to watch a program. To make sense of the high-frequency scale, they may expand that definition to include having the TV on while doing something else.

Respondents are less likely to glean information from the range on the frequency response scale, when the behavior they are asked to self report is better described. For example, the effect of the scale is smaller with questions about clearly defined experiences, such as "excessive perspiration" than with questions about ambiguous experiences such as "responsiveness to the weather" (Schwarz & Scheuring, 1992). Thus, researchers should aim to clearly define the behaviors they are asking respondents to report, and use frequency scales that are interpreted as the appropriate range for that definition of the behavior.

Studies in judgment and decision making. In the field of judgment and decision making, researchers have also studied the effects of presenting options. Decisions may be presented with one explicit option, also referred to as the default. For example, consider decisions about organ donation. Some countries, including France and Austria, have policies that automatically enter people into organ donation programs, unless they opt out. Other countries, including the United States and the Netherlands, require people to actively opt in, if they want to be an organ donor. Countries in which opting in is the policy default tend to have many more organ donors than countries in which opting out is the policy default (Johnson & Goldstein, 2003). Similar default effects have also been observed in decisions about pension saving plans (Madrian & Shea, 2001), insurance

(Johnson et al., 1993) and internet privacy (Johnson, Bellman, & Lohse, 2002). These default effects seem to violate normative rules of decision making, which posit that people's choices should be solely made on the basis of their preferences, and be unaffected by seemingly meaningless variations in the problem presentation.

Several explanations have been offered for people's sensitivity to default settings. First, it may be psychologically painful to give up the positive characteristics of the suggested default, which has become the decision-maker's reference point (Johnson et al., 1993). Such loss aversion would violate decisional norms, because it means that decision makers are affected by the way the problem is presented. Second, switching to the alternative may require too much time, money, and energy to be perceived as worthwhile. Although that reason would be normatively defensible, it turns out that large default effects still occur when no costs are associated to switching (Johnson & Goldstein, 2003). Finally, another alternative explanation is that, as with variations in question wording, different default settings may convey variations in the communicators' preferred choice of action (McKenzie, Liersch, & Finkelstein, 2006). Therefore, default settings may defensibly affect people's choices, especially if they are uncertain about their preferences.

Researchers in judgment and decision making have also found that changing the number of options can affect evaluations. Judging one option in the context of alternatives may be different from judging it on its own, if the alternatives obtain information about how to evaluate specific attributes. Indeed, controlled laboratory experiments have suggested that some attributes, such as the number of entries in a dictionary, are difficult to evaluate without a comparison to other dictionaries, while

other attributes, such as whether or not the cover of the dictionary is torn, are easy to evaluate on their own. People tend to give more weight to the attributes that they are able to evaluate. As a result, they are willing to pay more for a dictionary described as having 20,000 entries and a torn cover than for a new dictionary with 10,000 entries when the two descriptions are presented *separately*. When the two descriptions are presented *jointly*, however, the latter is assessed as more valuable (Hsee, 1996).

The presentation of a third option may also provide information about how to choose between two other options (Huber, Payne, & Puto, 1982; Simonson, 1989; Simonson & Tversky, 1992). For example, when people are choosing between a computer with 960k memory that costs $1200 and another computer with 640K memory that costs $1000, a third option can provide tacit information about whether it is worthwhile to pay $200 more for 320K worth of additional memory (Simonson & Tversky, 1992). A third option may even affect the relative preference between two other options if it is inferior to one of those two other options, and not even under serious consideration. For example, research participants are more likely to prefer an elegant pen over $6 when they are presented with the third option of a low-quality pen than when they are not (Simonson & Tversky, 1992).

Researchers in judgment and decision making may deem such choices defensible, if the third option truly provides information about how to evaluate the features of the other two. However, research participants are affected by the presentation of a third option even when they are evaluating familiar products, whose quality should be relatively easy to assess (Simonson & Tversky, 1992), and after receiving extensive information about how to interpret attribute values (Huber et al., 1982).

Yet, supporters of the logic of conversation may argue that research participants who assume that they are in a collaborative conversation should pay attention to a third option, if it is presented. The maxim of quantity holds that the information that is provided is relevant to the conversation. Thus, when researchers add a low-quality pen to a choice set consisting of $6 and a luxury pen, they may be inadvertently signaling to participants that they find it important for their participants to have a pen. If participants are sensitive to this implicit message, then they may feel that they should choose one of the pens, with the luxury pen being obviously better than the low-quality pen. Indeed, after the presentation of the inferior option, research participants find it easier to justify a choice for its luxury superior – and are even more likely to choose it if they expect that they will have to explain their choice (Simonson, 1989). Thus, variations in the choice set may defensibly affect responses because they convey different meanings to the decision maker.

Presentation order.

Studies in survey design. After questions have been designed, they need to be combined into one survey. Among other things, researchers may add instructions about how to answer the questions, design an informed consent form, and post an ad inviting people to participate. Each may become part of the conversation between the researchers and the respondents, potentially affecting how the survey questions are interpreted (see Schwarz, 1996, 1999). Here, we focus on order effects, with questions that are presented earlier influencing the interpretation of, and responses to, questions that are presented later (Schuman & Presser, 1981; Sudman et al., 1996; Schwarz, 1999).

The previous section discussed the effect of wording on a question's responses. These wording effects may also influence responses to subsequent questions. For example, studies of eyewitness testimony have shown participants a video of a car accident (Loftus & Palmer, 1974). In one condition, participants were then asked "about how fast were the cars going when they smashed into each other?" That question led to significantly higher estimates of the observed speed than did similar questions in which "smashed into" was replaced with synonyms such as "hit," "collided with," or "bumped into." These wordings also affected responses to a subsequent question, which asked participants whether they had seen broken glass. There had been none in the video. Participants were more likely to report having seen broken glass if they had originally been asked about the cars smashing into each other rather than hitting each other. Thus, participants may have interpreted the experimenter's choice of words as conveying accurate information about the video-taped accident, especially if they did not remember the exact details.

Overall, general questions are more prone to order effects than specific ones, (Schwarz, Strack, & Mai, 1991; Tourangeau, Rasinski, & Bradburn, 1991). For example, respondents report being less satisfied with their lives in general when they are first asked about how happy they are with their romantic relationships, compared to when they receive the opposite presentation order (Schwarz et al., 1991; Tourangeau et al., 1991). Conversational norms require that respondents should avoid redundancy in their answers, across consecutive questions. When the general question about life satisfaction is asked first, it may be interpreted as covering various types of satisfaction (e.g., regarding one's career, family relations, and romantic relationship). However, when the general question

about life satisfaction is asked after the specific question about relationship satisfaction, respondents will interpret the question as asking "Aside from your marriage, which we have already asked you about, how satisfied are you with other aspects of your life?" (Schwarz, 1999). Indeed, reports of general life satisfaction are more strongly correlated to reports of relationship satisfaction when the general question appears first (vs. second). Moreover, when the question about general life satisfaction is re-worded to explicitly include relationship satisfaction, it is not affected by presentation order (Schwarz et al., 1991). Researchers in survey design therefore recommend writing specific questions, or presenting general questions before specific ones (Converse & Presser, 1986; Sudman & Bradburn, 1982).

Response options may also be presented in sequence, potentially causing order effects. On written surveys, options that appear near the beginning of the list are more likely to be selected (Krosnick, 1991). Such primacy effects are even more pronounced on web surveys using drop-down menus (Couper, Tourangeau, Conrad, & Crawford, 2004). It appears that earlier options receive more cognitive processing, with respondents presumably thinking of more reasons for selecting them (Couper et al., 2004). Possibly, respondents assume that earlier options are more important to the researchers. If so, the logic of conversation suggests that these options should receive more consideration.

Studies in judgment and decision making. Studies in judgment and decision making tend to use a between-subjects design, randomly assigning participants to one of two decision scenarios, followed by one simple question. Thus, all information is presented at the same time. If participants are asked to evaluate more than one option, the choice set is often presented simultaneously. As a result, order effects should not

play a role in the observed responses. One exception involves studies of belief updating, in which participants are asked to repeatedly judge one option after receiving pieces of evidence that are presented one at a time (Hogarth & Einhorn, 1992). Another exception, discussed here, involves studies in which participants are asked to judge different options that appear in sequence.

Outside of the psychological laboratory, people regularly face real-world decisions in which options appear sequentially. Consider for example, buying a house, tasting wines, searching for a job, or judging contestants in a music competition. If presentation order is not random, it is difficult to determine whether order effects on judgments are due to the presentation order per se, or due to other factors that are confounded with presentation order. For example, realtors may present houses in a certain sequence, if they believe that it will make it more likely that they will make a sale. If so, the presentation order may present decision makers with information about how to judge the different options.

In some real-world situations, order of presentation is truly random. For example, organizers of international sports and music competitions tend to randomly determine the sequence in which contestants compete. When performance order is determined by a random draw, candidates' serial position does not contain information about their quality, and should not affect how they are judged. Yet, contestants who perform later tend to be evaluated as better, and, therefore, are more likely to win. Such order effects have been found in judged competitions of figure skaters, gymnasts, synchronized swimmers, classical musicians, and pop groups (Bruine de Bruin, 2005; Glejser & Heyndels, 2001; Haan, Dijkstra, & Dijkstra, 2003; Scheer, 1973; Wilson, 1977).

At least three different explanations have been offered for these order effects. First, because each performance is judged in light of previous ones, each additional performance can change how new appearing options are evaluated. As argued in the previous section, adding a new option to a choice set can teach decision makers how to evaluate specific attributes, and change their perceived importance. Competing contestants may also have attributes that become easier to appreciate after seeing multiple performances. Perhaps as a result, order effects in judgments of contestants are more pronounced in evaluations made by audience members rather than by experts (Haan et al., 2003).

A second explanation for order effects focuses on the comparison process. When two options appear in sequence, the first becomes the referent to which the second is compared (see chapter 9). Thus, it is more natural for judges to examine how the second option is different from the first, rather than vice versa. The unique features of the second option therefore tend to receive more weight than the unique features of the first option, or the features that are shared. If the second option has positive unique features, it tends to be judged as better than the first. If the second option has negative unique features, it tends to be judged as worse than the first. This direction-of-comparison process also creates order effects when more than two options are presented, and occurs when options are evaluated one at a time, or after all have been seen (Bruine de Bruin & Keren, 2003). Because competitions generally tend to invite contestants because of their good performance abilities, this direction-of-comparison effect would cause later performers to be judged as better than earlier ones.

A third explanation for order effects focuses on the role of memory. As time passes, judges may forget the details of the earlier alternatives they have seen. Such imperfect recall may introduce uncertainty, and lead judges to regress their earlier judgments towards the mean (Li & Epley, 2009). When options are relatively good, early ones will therefore appear less desirable over time, compared to later ones that are remembered better. When options are relatively bad, ones that are presented early will similarly be remembered as more desirable with the passing of time. Thus, in competitions between good performers, memory-induced regression towards the mean would lead later performers to be judged as better than earlier ones.

Whether the observed order effects in judgments are defensible, in terms of the logic of conversation, depends on whether judges interpret presentation order as relevant to their evaluations. That may be the case if judges assume that contestants appear in a specific order, determined by the organizers, for example, to increase their viewing pleasure. Indeed, people generally prefer their experiences to improve rather than deteriorate over time – as seen in preferences for salary profiles (Hsee, Abelson, & Salovey, 1991), as well as sequences of pleasurable experiences (Loewenstein & Prelec, 1993), and painful experiences (Varey & Kahneman, 1992). If people do indeed expect that the order in which they are asked to judge options conveys information about how to evaluate the options, then the logic of conversation posits that presentation order *should* dictate the process by which sequentially presented options are judged. To examine whether this is indeed the case, studies would need to examine how judges interpret the presentation order of options, and whether these interpretations may explain the reported order effects in research of judgment and decision making.

Improving survey design.

Several research practices have been suggested to avoid effects of wording, choice set, and presentation order on responses (e.g., Bradburn, 1982; Converse & Presser, 1986; Dillman et al., 2009; Presser et al., 2004; Schwarz, 1996, 1999; Sudman & Bradburn, 1982). First, as noted above, questions that are difficult to read or otherwise ambiguous, may lead respondents to use other information – contained in the question wording, the choice set, or the presentation order – to interpret what is being asked. It is therefore recommended to write research materials in specific and simple wording, leaving little room for interpretation (Converse & Presser, 1986; Dillman et al., 2009; Sudman & Bradburn, 1982). At the same time, if the goal is to understand real-world decisions, questions should reflect the tasks that respondents would face in the real-world. For example, response options should reflect information that respondents would typically consider (Bruine de Bruin & Fischhoff, 2000; Sudman & Bradburn, 1982; Schwarz, 1999).

Second, qualitative pilot interviews can reveal whether questions are understood as intended, and whether any of the presented information affects respondents' interpretations. In these interviews, participants are asked to think out loud while answering drafts of questions, criticize their wording, and repeat their meaning back in their own words (Blair, Ackermann, Piccinino, & Levenstein, 2007; Converse & Presser, 1986; Dillman et al., 2009; Presser et al., 2004). Because these interviews are labor-intensive, they are typically conducted with small samples, rendering insufficient statistical power to test whether different question interpretations affected responses.

However, they may inspire hypotheses about what is communicated by variations in wording, choice sets, and presentation order, and which version is most likely to reflect participants' real-world decisions. These hypotheses can then be tested on a subsequent survey with a larger sample.

Third, when conducting surveys and experiments, researchers should ask participants how they interpreted those questions that are crucial to the researchers' hypotheses. Such comprehension checks can confirm whether questions were interpreted as intended, and allow researchers to separately analyze responses by participants who had the intended interpretations, rather than alternative ones (Converse & Presser, 1986; Fischhoff, 2005; Frisch, 1993; Keren & Willemsen, 2009). The interviews mentioned above should inform the design of questions to perform these comprehension checks. Indeed, without these interviews, researchers may not always know what to ask in these comprehension checks.

Fourth, assessing reliability and validity will help researchers to decide on the best way to measure constructs of interest. Indeed, if questions are interpreted in the same way by most respondents, they will have (a) good reliability, in terms of measuring a construct consistently, and (b) good validity, in terms of showing correlations with external criteria. Both are discussed in detail below.

Reliability is gauged across multiple measurements. The same person should provide the same response at each measurement – as long as the conditions that determine the underlying construct have not changed between the two times of measurement. For example, two self-ratings of personal health assessed a week apart should be similar for people whose health has not changed. Therefore, the correlation

between the two measurements will be higher when most of their variation reflects meaningful individual differences in the construct rather than random noise due to respondents' confusion.

Of course, it may not always be possible to ask respondents the same questions again at a later time. Another way of examining reliability is to give respondents different questions about the same construct, given on the same survey. Their responses should be correlated, if variations in responses reflect meaningful individual differences in the construct, rather than random measurement error. For example, questions about forbidding vs. allowing anti-democratic speeches, which were discussed above, may be measuring similar attitudes, as their responses are actually correlated (Holleman, 1999a).

Reliability is a prerequisite for validity, which means that responses are related to external standards. Concurrent validity refers to correlations with measures of related concepts that are taken at the same time. Predictive validity refers to correlations with measures that are taken some time in the future. For example, female adolescents' judged probabilities of getting pregnant in the next year are correlated with (a) their concurrent self-reports of their sexual activity, suggesting concurrent validity (Fischhoff et al., 2000), as well as with (b) subsequent-year reports of having gotten pregnant, suggesting predictive validity (Bruine de Bruin, Parker, & Fischhoff, 2007a).

Questions that are not well-understood beget responses with low reliability and low validity. If respondents are confused about what a question means, they will have to guess at how to answer it. Doing so will introduce noise to their responses, reducing the reliability with which the intended construct is measured. Such reduced reliability also limits the potential for finding significant correlations with the external criterion it aims

to predict, thus threatening validity. However, questions may show reliability without validity, if they are systematically misinterpreted. For example, questions about "prices in general" may elicit reliable responses that are not valid predictors of inflation, if respondents systematically think of gas prices when answering it.

Conclusions and final comments.

Like studies in survey design, studies in judgment and decision making have found that responses can be systematically affected by slight variations in question wording, choice sets, and presentation order, among other things. Researchers in judgment and decision making tend to interpret such response patterns as violating normative decision-making principles. By contrast, researchers in survey design tend to interpret them as resulting from their own violations of conversational norms. Ultimately, either conclusion can not be drawn without knowing whether respondents interpreted the questions in the way the researchers intended.

Fortunately, studies in judgment and decision making have started to examine how research participants interpret hypothetical decision problems, sometimes revealing interpretations that are different from those previously espoused by researchers (Fischhoff, 2005; Frisch, 1993; Keren & Willemsen, 2009; Johnson et al., 2007; McKenzie, 2004; McKenzie & Nelson, 2003; McKenzie et al., 2006; Reyna & Brainerd, 1991, Shafir et al., 1993). Such process-oriented approaches will likely contribute to new insights into how people actually make decisions, and improve measures of how people make decisions.

Traditionally, researchers in judgment and decision making have paid little attention to reliability, largely ignoring individual differences (Levin, 1999). They typically recruit homogenous samples of undergraduate research participants, assuming that the results generalize to a broader population – which, fortunately, has been shown to be the case for framing effects (Kühberger, 1998). However, because the presented problems tend to be hypothetical, they may have limited external validity, and be unrelated to real-world decisions (Gigerenzer et al., 2000; Klein, 1999).

Following the lead of Stanovich & West (1998, 2000), recent research has started to examine the reliability and validity of commonly studied decision problems. An individual difference measure of decision-making competence (DMC) has been developed for adolescents (Parker & Fischhoff, 2005) and for adults (Bruine de Bruin et al., 2007b). Among other things, components measure people's ability to apply decision rules, to be appropriately confident in their knowledge, and to consistently assess risks. The ability to avoid framing effects is seen in giving the same response to a pair of items, with one presenting a decision problem in positive terms, and the other presenting it in negative terms. The overall framing measure includes 14 such item pairs, which were adapted from the literature, and include the Asian disease problem (Tversky and Kahneman, 1981) and the ground beef problem (Levin & Gaeth, 1988). Participants received the 14 positively framed items at the beginning of the survey session, and the 14 negatively framed items near the end. Across these 14 item pairs, performance appears reliable, such that respondents who avoid framing effects in one item pair also tend to avoid framing effects in other item pairs (Bruine de Bruin, Parker, & Fischhoff, 2007b; Levin, Gaeth, Schreiber, & Lauriola, 2002; Parker & Fischhoff, 2005). Moreover, the

ability to avoid framing effects is reliably correlated to performance on other decision-making tasks measuring under/overconfidence, applying decision rules, consistency in risk perception, and to existing measures of general cognitive ability (Bruine de Bruin et al., 2007b; Stanovich & West, 1998).

Unlike the other components of Adult Decision-Making Competence, the ability to avoid framing effects was uncorrelated to the number of negative life decision outcomes (such as having been diagnosed with type 2 diabetes, having declared bankruptcy, and having spent a night in jail for any reason) reported on the Decision Outcome Inventory (Bruine de Bruin et al., 2007b). There may be several related explanations for that seeming lack of external validity. First, the ability to avoid framing effects may be irrelevant to real-world choices, because real-world frames sometimes convey meaningful information to which people should pay attention (McKenzie, 2004; McKenzie & Nelson, 2003; McKenzie et al., 2006). Second, the study presented people with both frames, whereas most real-world decisions may only be based on one frame. Third, the study may have measured some other ability than the ability to avoid framing errors. Because measures were presented in the same survey session, giving consistent responses may be a reflection of the ability to remember the first frame when the second frame is being presented, and to recognize the second frame as similar to the first (LeBeouff & Shafir, 2003). To reduce the influence of such memory skills, it may have been better to present the two sets of frames in separate sessions spaced a week or more apart (Levin et al., 2002).

Research on individual differences may ultimately provide insights on how to teach people to make better decisions. Having a reliable and valid measure of decision-

making competence is the first step towards identifying decision-making processes that are in need of improvement, and that are relevant to the real world. Although more studies are needed to warrant serious conclusions, the available research suggests that teaching decision-making skills may help people to improve their real-world decisions – although avoiding framing effects may not be one of these skills.

Without understanding how people interpret a question, it is impossible to draw conclusions about their decision-making skills, and what they need to know to improve their decisions. Researchers in survey design have developed the methodological tools needed to understand how respondents interpret questions, and to identify the best way of asking questions when slight variations in design cause differences in interpretations. Researchers in judgment and decision making have started to apply these methodological tools to understand potential decision-making errors. Overall, these efforts may lead researchers to ask better questions, allowing respondents to give better answers.

References.

Blair, J., Ackermann, A., Piccinino, L., Levenstein, R. (2007). Using behavior coding to validate cognitive interview findings. *Proceedings of the American Statistical Association: Survey Research Methods Section*, pp. 3896-3900.

Blake, D.R., Weber, B.M., & Fletcher, K.E. (2004). Adolescent and young adult women's misunderstanding of the term pap smear. *Archives of Pediatrics and Adolescent Medicine, 158,* 966-970.

Bostrom, A., Morgan, G.M., Fischhoff, B., & Read, D. (1994). What do people know about global climate change? 1. Mental models. *Risk Analysis, 14,* 959-970.

Bradburn, N. (1982). Question-wording effects in surveys. Hogarth, R.M. (Ed.). *Question framing and response consistency.* San Francisco, CA: Jossey-Bass (pp. 65-76).

Bruine de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects on jury evaluations. *Acta Psychologica, 118*, 245-260.

Bruine de Bruin, W. & Fischhoff, B. (2000). The effect of question format on measured HIV/AIDS knowledge in detention center teens, high school students, and adults. *AIDS Education and Prevention, 12,* 187-198.

Bruine de Bruin, W. & Keren, G. (2003). Order effects in sequentially judged options due to the direction of comparison. *Organizational Behavior and Human Decision Processes, 92,* 91-101.

Bruine de Bruin, W., Parker, A.M., & Fischhoff, B. (2007a). Can teens predict significant life events? *Journal of Adolescent Health, 41*, 208-210.

Bruine de Bruin, W., Parker, A.M., & Fischhoff, B. (2007b). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology, 92*, 938-956.

Bruine de Bruin, W., VanderKlaauw, W., Downs, J.S., Topa, G., & Armantier, O. (2009). The effect of question wording on reported inflation expectations. Paper presented at the conference on Subjective Probability and Utility in Decision Making (SPUDM), Rovereto, Italy.

Converse, J.M., & Presser, S. (1986). *Survey questions. Handcrafting the standardized questionnaire.* Thousand Oaks, CA: Sage Publications.

Couper, M.P., Tourangeau, R., Conrad, F.G., & Crawford, S.D. (2004). What they see is what we get: Response options for web surveys. *Social Science Computer Review, 22*, 111-127.

Cronbach, L.J. (1941). An experimental comparison of the multiple true-false and multiple multiple-choice tests. *Journal of Educational Psychology, 32,* 533-543.

Curtin, R. (2006). *Inflation expectations: Theoretical Models and Empirical Tests.* Paper presented at the National Bank of Poland Workshop on The Role of Inflation Expectations in Modelling and Monitary Policy Making, Warsaw.

Dillman, D.A., Smyth, J.D., & Christian, L.M. (2009). *Internet, mail, and mixed-mode surveys. The tailored design method.* Hoboken, NJ: Wiley.

Fischhoff, B. (2005). Cognitive processes in stated preference methods. In K-G. Mäler & J. Vincent (Eds.), *Handbook of Environmental Economics* (pp. 937-968). Amsterdam: Elsevier.

Fischhoff, B. (1993). Transaction analysis: A framework and an application to insurance decisions. *Journal of Risk and Uncertainty, 7*, 53-69.

Fischhoff, B. (1991). Value elicitation: Is there anything in there? *American Psychologist, 46*, 835-847.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1980). Knowing what you want: Measuring labile values. In: T. Wallsten (Ed.), *Cognitive processes in choice and decision behavior*. (pp. 17-141). Hillsdale, NJ: Erlbaum.

Fischhoff, B. & Bruine de Bruin, W. (1999). Fifty-fifty=50%? *Journal of Behavioral Decision Making, 12*, 149-163.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1980). Knowing what you want: Measuring labile values. In T Wallsten (Ed.), *Cognitive processes in choice and decision behavior* (pp. 117-141). Hillsdale, NJ: Erlbaum.

Fischhoff, B., Parker, A.M., Bruine de Bruin, W., Downs, J.S., Palmgren, C., Dawes, R. & Manski, C. (2000). Teen expectations for significant life events. *Public Opinion Quarterly, 64*, 189-205.

Fleishman, L., Bruine de Bruin, W., & Morgan, M.G. (2009). Public Perceptions of Carbon Capture and Sequestration and other Carbon-Reducing Technologies. Manuscript under review.

Frisch, D. (1993). Reasons for framing effects. *Organizational Behavior and Human Decision Processes, 54*, 399-429.

Gigerenzer, G., Todd, P. M., & the ABC Group (2000). *Simple heuristics that make us smart*. Cary, NC: Oxford University Press.

Glejser, H., & Heyndels, B. (2001). Efficiency and inefficiency in the ranking in competitions: the case of the Queen Elisabeth Music Contest. *Journal of Cultural Economics, 25*, 109–129.

Gourville, J.T. (1998). Pennies-a-day: The effect of temporal reframing on transaction evaluation. *Journal of Consumer Research, 24*, 395-403.

Grice, H.P. (1975). Logic and conversation. In: P. Cole & J.L. Morgan (Eds.). Cole & J.L. Morgan (Eds.). *Syntax and semantics. Volume 3. Speech acts.* (pp. 41-58). New York: Academic Press.

Haan, M., Dijkstra, S., & Dijkstra, P. (2005). Expert judgment versus public opinion–evidence from the Eurovision Song Contest. *Journal of Cultural Economics, 29*, 59–78.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*, 309-334.

Hershey, J.R., & Schoemaker, P.J.H. (1980). Risk taking and problem context in the domain of losses: An expected utility analysis. *Journal of Risk and Insurance, 47*, 111-132.

Hippler, H.J., & Schwarz, N. (1986). Not forbidding isn't allowing: The cognitive basis of the forbid-allow asymmetry. *Public Opinion Quarterly, 50*, 87-96.

Hogarth, R.M. (1982). *Question framing and response consistency*. San Francisco, CA: Jossey-Bass.

Hogarth, R.M., & Einhorn, H.J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology, 24*, 1-55.

Holleman, B. (2006). The meanings of "yes" and "no." An explanation of the forbid/allow asymmetry. *Quality and Quantity, 40,* 1-38.

Holleman, B. (1999a). The nature of the forbid/allow asymmetry: Two correlational studies. *Sociological Methods Research, 28,* 209-244.

Holleman, B. (1999b). Wording effects in survey research: Using meta-analysis to explain the forbid/allow asymmetry. *Journal of Quantitative Linguistics, 6,* 29-40.

Hsee, C.K. (1996). The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational Behavior and Human Decision Processes, 67,* 247-257.

Hsee, C.K., Abelson, R.P., Salovey, P. (1991). The relative weighting of position and velocity in satisfaction. *Psychological Science, 2,* 263-266.

Huber, J., Payne, J.W., & Puto, C. (1982). Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research, 9,* 90-98.

Hunter, J. (2005). *Report on cognitive testing of cohabitation questions.* Washington, DC: Statistical Research Division. U.S. Bureau of the Census. Survey Methodology 2005-06.

Hurd, M., Manski, C., & Willis, R. (2007). *Fifty-fifty responses: Equally likely or don't know the probability.* Paper presented at the Cognitive Economics Conference, Jackson Hole, WY.

Johnson, E.J., Bellman, S., & Lohse, G.L. (2002). Defaults, framing, and privacy: Why opting in-opting out. *Marketing letters, 13,* 5-15.

Johnson, E.J., & Goldstein, D. (2003). Do defaults save lives? *Science, 302,* 1338-1339.

Johnson, E.J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology, 33,* 461-474.

Johnson, E.J., Hershey, J., Meszaros, J., & Kunreuther, H. (1993). Framing, probability distortions, and insurance decisions. *Journal of Risk and Uncertainty, 7,* 35-53.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263-291.

Keren, G., & Willemsen, M.C. (2009). Decision anomalies, experimenter assumptions, and participants' comprehension: Reevaluating the uncertainty effect. *Journal of Behavioral Decision Making, 22,* 301-317.

Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S. (1975). Derivation of new readability formulas (automated readability index, fog count, and Flesch reading ease formula) for Navy enlisted personnel. Research branch report 8-75. Memphis: Naval Air Station.

Klein, G. (1999). *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.

Koriat, A., Lichtenstein, S. & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 107-118.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5,* 213-236.

Kühberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes, 75,* 23-55.

Kühberger, A. (1995). Framing effects: A new look at old problems. *Organizational Behavior and Human Decision Processes, 62,* 230-240.

Leiser, D., & Drori, S. (2004). Naïve understanding of inflation. *Journal of Socio-Economics, 34,* 179-198.

Lerner, E.B., Jehle, D.V.K., & Janicke, R.M. (2000). Medical communication: Do our patients understand? *American Journal of Emergency Medicine, 18,* 764-766.

Levin, I. P. (1999). *Why do you and I make different decisions? Tracking individual differences in decision making.* Presidential address to the Society for Judgment and Decision Making, Los Angeles, CA.

Levin, I. P., & Gaeth, G. J. (1988). Framing of attribute information before and after consuming the product. *Journal of Consumer Research, 15,* 374–378.

Levin, I.P., Gaeth, G.J., Schreiber, J., & Lauriola, M. (2002). A new look at framing effects: Distribution of effect sizes, individual differences, and independence of types of effects. *Organizational Behavior and Human Decision Processes, 88,* 411-429.

Levin, I.P., Schneider, S.L., & Gaeth, G.J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes, 76,* 149-188.

Levin, I. P., Schnittjer, S. K., & Thee, S. L. (1988). Information framing effects in social and personal decisions. *Journal of Experimental Social Psychology, 24,* 520–529.

Li, Y., & Epley, N. (2009). When the best appears to be saved for last: Serial position effects on choice. *Journal of Behavioral Decision Making, 22,* 378-389.

Linville, P. W., Fischer, G. W., & Fischhoff, B. (1993). AIDS risk perceptions and decision biases. In J. B. Pryor & G. D. Reeder (Eds.), *The social psychology of HIV infection* (pp. 5–38). Hillsdale, NJ: Erlbaum.

Loewenstein, G., & Prelec, D. (1993). Preferences for sequences of outcomes. *Psychological Review, 100,* 91-108.

Madrian, B.C., & Shea, D.F. (2001). The power of suggestion: Intertia in 401 (k) participation and savings behavior. *Quarterly Journal of Economics, 116,* 1149-1187.

Mandel, D.R. (2001). Gain-loss framing and choice: Separating outcome formulations from descriptor formulations. *Organizational Behavior and Human Decision Processes, 85,* 56-76.

McDaniel, M.A., Anderson, J.L., Derbish, M.H., & Morisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19,* 494-513.

McKenney, N.R., & Bennett, C.E. (1994). Issues regarding data on race and ethnicity: The Census Bureau experience. *Public Health Reports, 109,* 16-25.

McKenzie, C.R.M. (2004). Framing effects in inference tasks – and why they are normatively defensible. *Memory & Cognition, 32,* 874-885.

McKenzie, C.R.M., Liersch, M.J., & Finkelstein, S.R. (2006). Recommendations implicit in policy defaults. *Psychological Science, 17,* 414-420.

McKenzie, C.R.M., & Nelson, J.D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychological Bulletin & Review, 10,* 596-602.

Menon, G., Raghubir, P., & Schwarz, N. (1995). Behavioral frequency judgments: An accessibility-diagnosticity framework. *Journal of Consumer Research, 22,* 212-228.

Paasche-Orlow, M.K., Taylor, H.A., & Brancati, F.L. (2003). Readability standards for informed-consent forms as compared with actual readability. *The New England Journal of Medicine, 348,* 721-726.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making, 18,* 1–27.

Presser, S., Couper, M.P., Lessler, J.T., Martin, E., Martin, J., Rothgeb, J.M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68,* 109-130.

Ranyard, R., Del Missier, F., Bonini, N., Duxbury, D., & Summers, B. (2008). Perceptions and expectations of price changes and inflation: A review and conceptual framework. *Journal of Economic Psychology, 29*, 378-400.

Reyna, V.F., & Brainerd, C.J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making, 4,* 249-262.

Roediger, H.L., & Marsh, E.J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155-1159.

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly, 5,* 91-92.

Scheer, J. K. (1973). Effect of placement in the order of competition on scores of Nebraska high school students. *Research Quarterly, 44*, 79–85.

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys.* New York: Wiley.

Schuster, M.A., Bell, R.M., & Kanouse, D.E. (1996). The sexual practices of adolescent virgins: Genital sexual activities of low-risk adolescents who have never had vaginal intercourse. *American Journal of Public Health, 86,* 1570-1576.

Schwarz, N. (1999). Self reports. How the questions shape the answers. *American Psychologist, 54,* 93-105.

Schwarz, N. (1996). *Cognition and communication: Judgmental biases, research methods, and the logic of conversation.* Hillsdale, NJ: Erlbaum.

Schwarz, N., Hippler, H.J., Deutsch, B., & Strack, F. (1985). Response categories: Effects on behavioral reports and comparative judgments. *Public Opinion Quarterly, 49,* 388-395.

Schwarz, N., & Scheuring, B. (1988). Judgments of relationship satisfaction: Inter- and intra-individual comparison strategies as a function of questionnaire structure. *European Journal of Social Psychology, 18,* 485-496.

Schwarz, N., Strack, F., & Mai, H.P. (1991). Assimilation and contrast effects in part-whole question sequences: A conversational logic analysis. *Public Opinion Quarterly, 55,* 3-23.

Schwarz, N., Strack, F., Müller, G., & Chassein, B. (1988). The range of response alternatives may determine the meaning of the question: Further advice on informative functions of response alternatives. *Social Cognition, 6,* 107-117.

Shafir, E., Simonson, I., & Tversky, A. (1993). Reason-based choice. *Cognition, 49,* 11-36.

Simonson, I. (1989). Choice based on reasons: The case of attraction and compromise effects. *Journal of Consumer Research, 16,* 158-174.

Simonson, I., & Tversky, A. (1992). Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research, 29,* 281-295.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645–726.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161–188.

Sudman, S., & Bradburn, N. (1982). *Asking questions. A practical guide to questionnaire design.* San Francisco, CA: Jossey-Bass.

Svenson, O., & Nilsson, G. (1986). Mental economics: Subjective representations of factors related to expected inflation. *Journal of Economic Psychology, 7,* 327-349.

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true–false examinations. *Journal of Educational Research, 83,* 119–124.

Tourangeau, R., Rasinski, K.A., & Brandburn, N. (1991). Measuring happiness in surveys: A test of the subtraction hypothesis. *Public Opinion Quarterly, 55,* 255-266.

Tourangeau, R., & Smith, T.W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly, 60,* 275-304.

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference-dependent model. Quarterly Journal of Economics, 106, 1039–1061.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211,* 453-458.

Varey, C.A., & Kahneman, D. (1992). Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making, 5,* 169-185.

Wilson, V. E. (1977). Objectivity and effect of order of appearance in judging of synchronized swimming meets. *Perceptual and Motor Skills*, *44*, 295–298.