Chapter 3

How Easy It Is to Ask the Wrong Question

> *The goal is to have differences in answers reflect differences in where people stand on the issues, rather than differences on their interpretations of the questions.*
>
> —Floyd Fowler, *Improving Survey Questionnaires: Design and Evaluation*

"Improving question design is one of the easiest, most cost-effective steps that can be taken to improve the quality of survey data" (Fowler 1995, vii), yet it is frequently one of the most disregarded. While many people focus a lot of attention on sampling where the discussion of errors often deals with few percentage points, "experiments suggest that the potential range of errors involved in sensitive or vague opinion questions may be twenty or thirty percentage points" (Warwick and Lininger 1975, 126).

Although there is no formal theory on the wording of a question, a general principle exists to substantially improve its design. That is, two basic rules make a good question: relevance and accuracy.

Relevance

Relevance is achieved when the questionnaire designer is intimately familiar with the questions, knows exactly the questions' objectives, and the type of information needed. To enhance accuracy, the wording, style, type, and sequence of questions must motivate the respondent and aid recall. "Cooperation will be highest [. . .] when the questionnaire is interesting and when it avoids items difficult to answer, time-consuming, [or] embarrassing" (Warwick and Lininger 1975, 127). A question is relevant if the information generated is appropriate for the purpose of the study. The objective of the question defines the information that is needed and models the words to be used. Sometimes this task is easy, for example, when asking the respondent's age. Other seemingly simple tasks, such as estimating the respondent's level of income is trickier. Hence, the questionnaire designer must force the analysts to be very specific about what they want to measure and why. "Until researchers decide specifically what their goals are it is impossible to write an ideal question" (Fowler 1995, 11).

Accuracy

A question is accurate if it collects the information sought in a reliable and valid manner. It serves no purpose to ask the respondent about something he or she does not understand clearly or that is too far in the past to remember correctly; doing so generates inaccurate information. As discussed later, respondents rarely admit ignorance. Rather, for a number of different reasons (the desire to be helpful or not appear ignorant), they tend to answer any question, even if they are not informed or barely understand the matter at hand. Because surveys query a variety of respondents, the questionnaire designer must always pose these questions only to people who are able to provide an accurate answer (Moser and Kalton 1971).

It is not always easy to determine whether the respondent has sufficient information to provide an accurate answer. The questionnaire designer should not fall into the trap of thinking easier questions give more accurate answers. This is especially true for opinion questions. By asking opinions on the budget deficit, for instance, we can not distinguish between whether the policy is wrong or the respondent is uninformed. Opinion questions require a validity check to screen "informed" respondents. This is accomplished by resorting to data on measurable behavior available from other sources (Moser and Kalton 1971) (that is, asking the current level of budget deficit) or by asking similar questions in different parts of the questionnaire to check the consistency of answers.[1]

Willingness

Finally, unless the respondent is *willing* to provide an answer, asking the right question to the right respondents still may not produce the desired outcome. In most surveys, respondents are not obliged to participate and are generally reluctant to do so.

> Many forces motivate people to participate in a survey: an interest in the topic, a desire to be helpful, a belief of the importance of the survey, a feeling of duty. . . . Other forces influence people to refuse: difficulty in understanding the questions, fear of strangers, the feeling of one's time being vested, difficulty in recalling information, and embarrassment at personal questions. (Plateck, Pierre-Pierre, and Stevens 1985, 17)

The way the survey is presented, how difficult the questionnaire is, and how sensitive questions are addressed influences the willingness of a prospective respondent to participate.[2]

[1] This second approach is the hardest to implement and not recommended.

[2] Issues of survey participation are addressed in the section on survey interview (chapter 5).

Practical Guidelines in Questionnaire Design

Constructing effective questions is an art in which field experience, along with a basic knowledge of linguistic and cognitive psychology plays a critical role (Peterson 2000). Although practitioners have developed techniques to help assess the level of readability and difficulty of questions, the ability to design a question cannot be learned from a book.[3] There is no substitute for experience of personally piloting and conducting interviews.

"A good rule to remember in designing questions [. . .] is that the respondent has probably not thought about these questions at the level of detail required by the survey" (Warwick and Lininger 1975, 158). When developing a question, the designer should first of all put himself "in the position of the typical, or rather the least educated, respondent" (Moser and Kalton 1971, 320). He or she must have a sense of the cognitive abilities of respondents and design the questions accordingly. Hence, while South Asia and East Asia are the regions with the highest share of businessmen with university training, in Sub-Saharan Africa less than half of the businessmen hold a university degree. Similarly Sub-Saharan Africa and Latin America are the only regions where approximately 10 percent of businessmen have not completed secondary education (figure 3.1).

Last but not least the mode of the interview[4] must also be taken into account when designing questions. The same word may generate confusion if spoken but be unambiguous if written. Homophonic words might elicit different interpretations in oral interviews, whereas in some languages different *intonations* of the same word will educe a completely different meaning (Peterson 2000).[5]

Question Wording

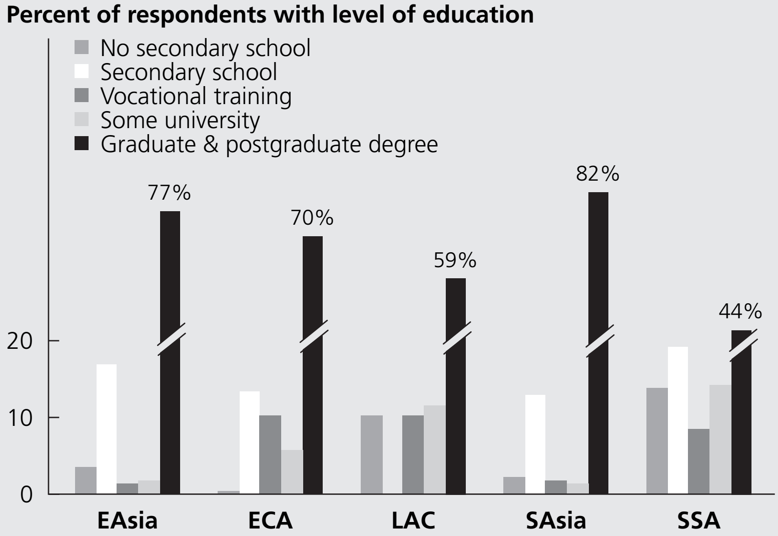
A number of studies have irrefutably shown that changing even a single word in a question can significantly alter the response distribution and accuracy. Three decades ago Loftus and Zanni (1975) reported the

[3] See Homan, Hewitt, and Linder 1994; Stevens, Stevens, and Stevens 1992; Gallagher and Thompson 1981; McConnell 1983.

[4] Possible modes include face-to-face, telephone, and mail interviews.

[5] The survey mode has a clear effect on a number of survey issues well beyond wording. Table 1 in Tourangeau and Smith (1996) shows survey mode effects on sensitive topics.

Figure 3.1.
Levels of Education of IC Survey Respondents

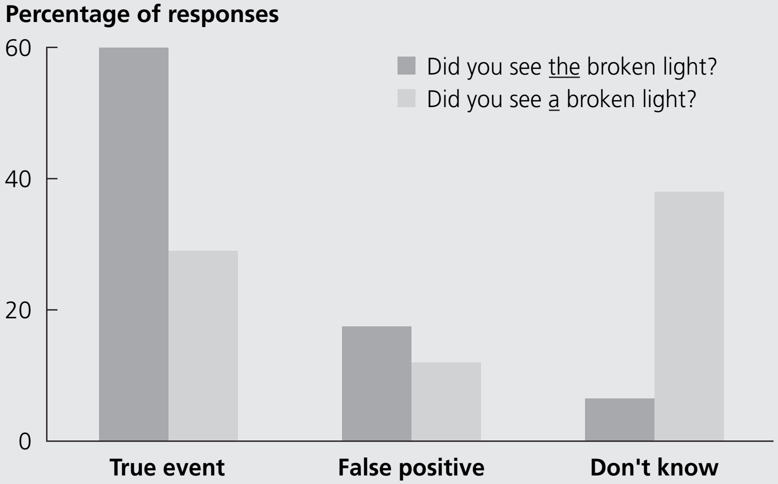


Source: Author's calculations.
Note: EAsia = East Asia; ECA = Europe and Central Asia; LAC = Latin America and the Caribbean; SAsia = South Asia; SSA = Sub-Saharan Africa.

results of two experiments in which a short movie is shown to two independent groups followed by a series of questions, some referring to events not even present in the movie. Figure 3.2 shows how changing one word for another—one group was asked, "Did you see *the* broken light?" the other group was asked, "Did you see *a* broken light?"—has a significant impact on the response distribution. In the mind of the respondent, "a" increases uncertainty about the existence of the event and consequently boosts (by more than half) the number of nonresponses. By contrast, "the" leads the respondent to infer the presence of an event, even if the event is nonexisting, hence encouraging false recognitions.

Because of the unique needs of each question, there is no universally accepted theory on question wording. There is, however, a general agreement on what constitutes good and bad questions. Four criteria should be followed when wording any question: it must be *brief, objective, simple,* and *specific* (or BOSS).

Figure 3.2.
A One-word Change Has a Significant Impact on Response



Source: Loftus and Zanni 1975.

Be Brief

All practitioners would agree that "unless a question is relevant to the research being conducted, it should not be included in a questionnaire. Likewise, unless a word is relevant to a question, it should not be included in the question" (Peterson 2000, 52). Questions should be short. Longer questions quickly become more complex and confusing for the respondent as well as the interviewer. Presser and Zhao (1992) show how a shorter question helps the interviewer do a better job by decreasing the tendency to misread it. Furthermore, the complexity of a long question is magnified by the intricacy of the subject matter covered.

As a rule of thumb, a question should not exceed 20 words (Payne 1951) and should not have more than three commas (Peterson 2000); however, brevity should not only be judged on physical appearance but also on contextual simplicity. In this sense, brevity means asking one question at a time. The designer must avoid the use of hidden questions, that is, questions that implicitly determine their relevance on another question. So asking "what interest rate are you paying on your loan?" implies the hidden question of "having a loan." More reliable data can be collected if we ask the questions separately: Do you have a loan? What interest rate are you paying? (Foddy 1993).

While brief questions are simpler, a question that is too short may also generate confusion. So the issue of brevity is not to reduce the length of a question by itself, but to choose the shortest way to pose the question without jeopardizing the intended meaning (see example 3.1). Likewise, a complex topic should not be phrased in one single question in the interest of brevity. This will only magnify its complexity and result in inaccurate answers.

The exception to the brevity requirement involves questions probing memory or sensitive topics. Experiments show that longer questions provide more accurate answers when memory or sensitive topics are covered (Peterson 2000).

Be Objective

"Nonobjective questions share a common characteristic: they tend to suggest an answer" (Peterson 2000, 57). The questionnaire designer should pay close attention to the neutrality of the words, because the question's objectivity can be subtly violated unintentionally. Hence he or she must be aware of the following:

Avoid leading questions. Leading questions are those questions that—by their content, structure, or wording—push the respondent in the direction of a certain answer by implication or suggestion (Warwick and Lininger 1975). So, for instance, a question that begins "Shouldn't something be done about . . . ?" leads to a positive answer. Similarly, when a question suggests only some of the alternatives, it leads in the direction of those alternatives, particularly if the respondent is not sure or does not understand the question properly (see example 3.2).

Response options

The set of response options have been proven to influence the answers given by the respondent in at least three different scenarios. First, failure to give equal weight to all options has the effect of suggesting what the usual or expected answers should be. Schwarz and others (1985) showed that compared with the true distribution on television viewing, respondents who were given a set of low-range categories to chose from were more likely to underreport. Similarly, respondents who were given a set of high-range categories did overreport television viewing (table 3.1).

Second, the actual set of options offered act as a source of information. This happens because respondents are reluctant to go beyond the list to avoid reporting behaviors that might appear unusual in the context of the range offered, or because respondents follow the easier path

Example 3.1

Does Brevity Mean Short?

Brevity in this case is achieved at the expense of clarity.

Original question: How frequently does your consignment arrive late at the gateway port and final destination in comparison with your planned schedule?

	Gateway port	**Final destination**
Average delay in the last year	(days)	(days)
Maximum delay in the last year	(days)	(days)

This question is extremely complex for a number of reasons:

a) it combines 4 different questions in one sentence. Generally, questions in table format are easy to write but extremely difficult to administer in a survey;
b) part of the question is not even included in the main text (average and longest delay);
c) it uses a general term, "frequently." Questions need to be specific; since we expect an answer in days we are to ask for "days";
d) There is no clear time reference. When? How long ago? Over what time period?
e) It assumes that the respondent experienced such an event. Filtering is missing.

A better way to ask this question(s) is:

Revised question:
In the last year, did you experience delays in delivering your goods from the factories to the gateway port? Yes / No
If yes, what is the average and maximum number of days that your export shipments arrived late at the gateway port in comparison with your planned schedule?

	Gateway port
Average delay in the last year	(days)
Maximum delay in the last year	(days)

In the last year, did you experience delays in delivering your goods from the gateway port to the final destination? Yes / No

If yes, what are the average and maximum number of days that your export shipments arrived late at the final destination in comparison with your planned schedule?

	Final destination
Average delay in the last year	(days)
Maximum delay in the last year	(days)

Example 3.2

Can I Ask You a (Leading) Question?

"How well is the Prime Minister managing monetary policy?"
1. Extremely well **2.** Very well **3.** Pretty well **4.** Well **5.** Not so well

This question has three elements that are designed to lead the respondent toward a favorable answer:

(1) The explicit inclusion of the word "well" in the main text of the question has the concealed intent of pulling the respondent toward a positive attitude. This is even more so if he or she is not aware of the event asked (see point 3).

(2) The range of options provided is not balanced, with all the choices referring to a different degree of positive attitude. Even the least favorable option rates the Prime Minister conduct of monetary policy as "not so well." A more balanced rating would include "Poor," "Very poor," and "Extremely poor."

(3) Finally there is no opt-out option (Don't Know). Because it is easier to influence the un-informed respondent, this is another way to lead him or her toward one of the "well" options reported.

Table 3.1

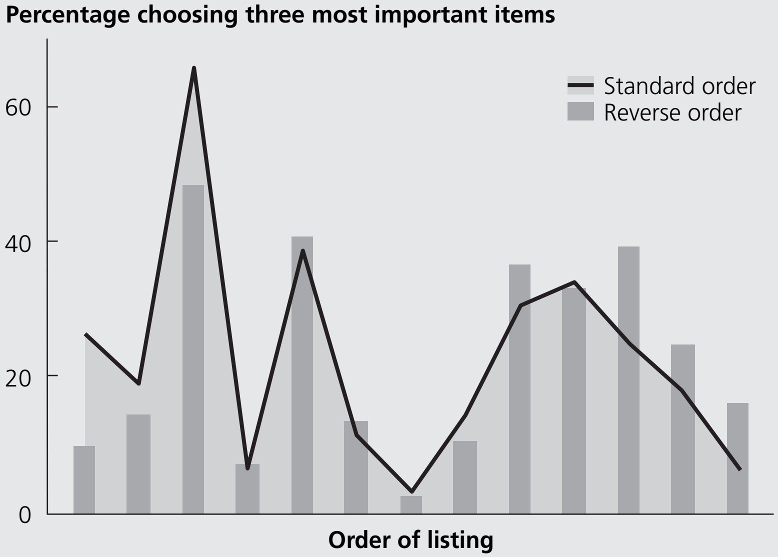
Reported Behavior Using Low and High Category Ranges

	Percentage of Estimated TV Usage		
Hours	**True Distribution**[a]	**Low Category**	**High Category**
up to 0.5	0	11.5	70.4
0.5 to 1	19.2	26.9	
1 to 1.5	15.4	26.9	
1.5 to 2	46.2	26.9	
2 to 2.5	0	7.7	
2.5 to 3	19.2		22.2
3 to 3.5	0	0	7.4
3.5 to 4	0		0
4.5 +	0		0
Mean	3.7	2.8	3.7

[a] Answers to an open-ended question.

Source: Schwarz and others 1985.

Figure 3.3.
Response rate distribution when the order of alternatives is reversed



Source: Krosnick and Alwin 1987.

of answering closed questions rather than recalling specific information (Foddy 1993). Third, the actual list of options provided will influence the respondent. Options that appear at the beginning of a long list seem to have a higher likelihood of being selected, which is known as the primacy effect (figure 3.3). Research on the primacy effect appears to show that this phenomenon is inversely correlated with the respondents' level of education (Krosnick and Alwin 1987). Furthermore, the interview's mode also plays a critical role. When the list of options are read to the respondents, there is evidence that that respondents tend to favor the ones they hear last (known as the recency effect). Conversely, when the respondent reads the list himself or herself (that is, when using show cards), the primacy effect seems to dominate (Foddy 1993).

Withholding information

Another case of a nonobjective leading question occurs when some information is withheld from the respondent. This would be case of asking "Are you in favor of a new road that would reduce rush hour traffic by 50 percent?" without mentioning that the road would be financed with a new tax (Peterson 2000). Finally, leading questions might

Politeness bias

generate the so-called "politeness or courtesy bias," when respondents, in their desire to be well-mannered toward the interviewer, might lean toward an answer that they think will please the interviewer (Plateck, Pierre-Pierre, and Stevens 1985).[6] This bias can be mitigated by using lead-in statements on both desirable and undesirable events, that is, "Many believe that . . . while others think that. . . . What is your opinion?" (Warwick and Lininger 1975).

Avoid loaded questions. Loaded questions bias answers through emotionally charged words, stereotypes, or prestige images such as "fair," "honest," "experienced," "colonialism," and so on. "Do you work?" is an example of a simple but emotionally charged question. Sometime this effect is more subtle. To describe the same phenomenon—for example, government help—the words "welfare" or "subsidy" are used if we refer to something we oppose, but the word "incentives" is used if we refer to something we favor (Browne and Keeley 2001). The "question designer must be continually on the alert for options which either flatter the respondent's self image or injure his pride" (Warwick and Lininger 1975, 144), because these options are a clever way to push the respondent in the desired direction.

Be wary of built-in assumptions. Generally speaking, questions should not take for granted that the respondent has familiarity with or carries out the activity asked in the question (Moser and Kalton 1971). The need for this awareness is even greater if the question refers to specific issues such as immigration laws or trade policy. Such questions could embarrass or annoy respondents who might claim knowledge they don't have so that they do not look ignorant or the respondents might refuse to continue with the interview (Plateck, Pierre-Pierre, and Stevens 1985). In these cases, filters should be used.

In fairness to some practitioners, it must be said that there are exceptional cases in which *not* using leading or loaded questions would bias the results. When you ask people whether they engage in certain disapproved practices (that is, paying bribes), they tend to lie and say no. However, if you provide more background information on the sensitive behavior[7] and then ask directly when, where, and how often

[6] This phenomenon is not limited to leading questions: it can very well occur with loaded questions.

[7] This to reduce the threatening nature of the question (see section on sensitive questions later in this chapter).

brides are paid, it is more likely that respondents will answer truthfully (Warwick and Lininger 1975).

Be Simple

The questionnaire designer should use language and terminology that exploits the simplest words and phrases. He or she should do the following:

Use words and expressions that are simple, direct, and familiar to all respondents. He or she must refrain from adopting "consider, initiate, purchase, or state," and use instead "think, start, buy, or say." He or she should not ask "Is it your opinion . . . ?" but simply "Do you think . . . ?" Similarly, the designer should refrain from employing slang expressions, because not everybody understands these expressions in the same way, if at all. It is not sufficient to ensure that all respondents understand the words used, it is necessary that they all understand the words in the same way (Moser and Kalton 1971). Take for instance the apparently simple and familiar expressions "majority" and "minority." What percentage value would you associate to these two commonly used expressions? Scipione (1995) asked this of a group of respondents and discovered that the average values associated with majority and minority were, respectively, 56.50 and 24.12 percent.[8]

Avoid technical jargons or concepts that are common only to those with specific and specialized training. The problem with technical terms, such as return on equity (ROE), discounted cash flow (DCF), and net present value (NPV), is determining whether the respondent understands the question or simply provides an answer in order not to appear ignorant. Furthermore, it is difficult to know whether the interpretation of the technical term is the same across respondents. Therefore, in these instances, if a technical term must be used and there is no simple correspondent concept, the technical term must be explained *before* the question is asked. Doing so may prevent the respondent from mentally framing the answer to the question based on his or her own interpretation of the technical term (Plateck, Pierre-Pierre, and Stevens 1985). After the respondent has framed the answer in his or her mind he or she will not listen to the definition provided afterward.

Adopt the same definitions throughout the form. If respondents are to answer accurately, the same definitions should be applied consistently

[8] With a wide standard deviation of 18.55 and 20.63, respectively. More examples are provided in table 3.7.

across all respondents. This is the only way their answers can be aggregated and compared not only within a country but also across regions. Although most practitioners agree with this predicament (that is, consistency in terminology), not all realize this also means avoiding the use of different terms with the same meaning. Unfortunately, we occasionally find that different definitions are used interchangeably to mean the same thing, such as when the same unit of investigation (the establishment) in a business survey is referred to as establishment, plant, factory, company, mother-company, firm, enterprise, or outlet (see example 3.3).[9]

If different terms are meant to indicate the same thing, then the questionnaire designer should use one term consistently throughout the whole document. Failure to do so will inevitably generate confusion in the respondents. If different terms do have different meanings, then all the definitions must be clearly explained in the questionnaire. This avoids possible confusion among respondents and ensures that the respondent answers questions on the basis of a consistent definition. Failing to take this into account in the questionnaire design is a major source of error (Fowler 1995).

When definitions are complex, it can become difficult to communicate a common definition to all respondents. In this case, it may be preferable to divide the single complex definition into a series of simple components. Hence, when asking for the geographic distribution of exports, it might be easier, and more accurate, for respondents to indicate the specific country of destination rather than the region (see example 3.4).

This approach has a number of benefits. First, it makes the question unambiguous because it is not necessary to communicate a common complex definition. Second, it makes the respondent's task easier because he or she does not have to add up or use the assigned definition to give an answer. Finally, this approach will provide the researcher with user-friendly data.

Avoid negative or double negative expressions. Double negatives not only generate cognitive complexity but also lead the respondent toward one answer. Suppose we ask a respondent whether he or she agrees or not with the following statement: "I am not satisfied with my job."

[9] Others examples of different terms used with the same meaning are (1) product, main product, main product line, most important activity, leading product, main line; (2) workforce, employees, workers; and (3) loan, term loan, line of credit, overdraft.

Example 3.3

To Whom Are We Talking?

Q 87 We have heard that ***establishments*** are sometimes required to make gifts or informal payments to public officials to "get things done" with regard to customs, taxes, licenses, regulations, services, etc . . .
Would you say this is true:

Always	1
Mostly	2
Frequently	3
Sometimes	4
Seldom	5
Never	6
Refuse	99
NA	100

Q 88 On average, what percent of annual sales value would such informal expenses cost to a typical ***firm*** like yours?
__________%

Q 89 Recognizing the difficulties many ***enterprises*** face in fully complying with taxes and regulations, what % of total sales would you estimate the typical establishment in your area of activity reports for tax purposes?
__________%

Q 90 Has your ***company*** been inspected or by or required to attend meetings with officials of national government, provincial, or municipal authority agencies during last 12 months?

Yes	1Go to Q90b
No	2Skip to Q91

This real case example shows how four different terms referring to the same unit of investigation (the establishment) are used interchangeably in four consecutive questions.

Note: Words in bold italic type emphasized here only; not on survey form.

Disagreeing with this statement of not been satisfied is a complex way of saying that he or she is satisfied (Fowler 1995). Similarly if we ask "You are going to do X, aren't you?" we imply the expectation of a yes answer. Conversely, if we ask "Aren't you going to do X?" we imply the expectation of a no answer (Foddy 1993). Experiments show that affirmative questions that are equivalent to superfluous negative questions take less time to answer (–7 percent) and prompt fewer requests for repetition or clarifications (–6 percent) (figure 3.4).

Simplicity is achieved when the level of cognitive effort the respondent is called on to perform in answering the question is minimized. This is not to say that difficult questions cannot be asked. Answers to

Example 3.4

Is Hong Kong Part of China?

X9. Please provide information on the percentage distribution of your plant's exports by destination regions:

	Percent of annual exports of *your plant* per year		Which year did *your plant* export to this region for the first time?
Regions of export	2001	2000	Year
a. West Europe	_____ %	_____ %	_____
b. East Europe	_____ %	_____ %	_____
c. North America (USA & Canada)	_____ %	_____ %	_____
d. Russia & Former Soviet Union countries	_____ %	_____ %	_____
e. China	_____ %	_____ %	_____
f. Rest of Asia (excluding China)	_____ %	_____ %	_____
g. Others specify	_____ %	_____ %	_____
TOTAL	100%	100%	

This example shows how difficult it is for the respondent to answer the question in the way presented. Exports to Hong Kong should be included in "China" or in "Rest of Asia"? Unclear definitions generate inaccurate answers.

complex questions can be obtained from surveys, but their accuracy will depend on the ability of the designer to match the level of cognitive complexity of the question with the respondent's level of cognitive ability.

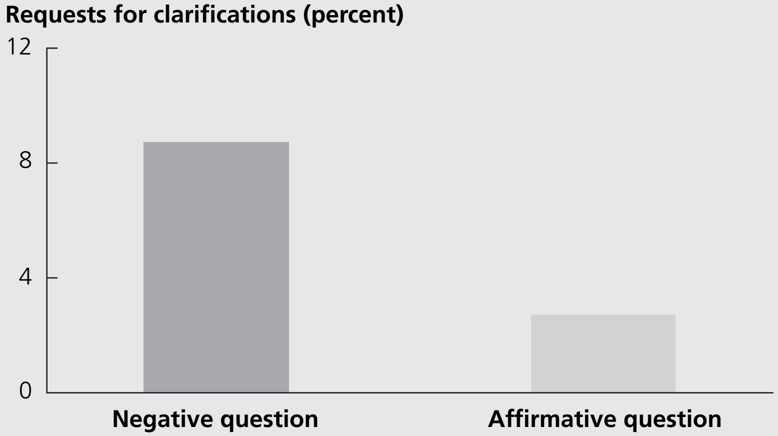
Be Specific

Being specific means asking precise questions. Vague queries will generate vague answers or, as is often the case in business surveys, will generate a sense of frustration in the respondent and lead to a perception that the study is not legitimate. Elite respondents do not like oversimplification of complex issues and when this happens they tend to ask detailed questions putting the interviewer in a difficult or embarrassing position (Zuckerman 1972).

Indefinite words

The questionnaire designer should avoid items that are too general, too complex, or too ambiguous. Indefinite words used in everyday con-

Figure 3.4.
Affirmative Questions Reduce Requests for Clarifications



Source: Bassili and Scott 1996.

versation such as *often, occasionally, usually, regularly, generally, rarely, normally* or *good, bad, approve, disapprove, agree, disagree, like, dislike* should also be avoided because they lack an appropriate objective dimension. For one person *often* may mean once a day, for another once a year (Warwick and Lininger 1975). The more general the question the wider the range of interpretations it invites (Converse and Presser 1986). If you ask "What kind of car do you have?" you should not be surprised to hear "a foreign car," or "a four-wheel-drive car," or "a sports car," or even "a very nice car."

Particular attention should be placed on the usage of words that imply great specificity, such as *ever, always,* and *never.* The meaning of these words extends the time horizon of the questions to the utmost limit and thus might render meaningless any answer because of its (almost) complete invariance. It might be legitimate to use these expressions when the phenomenon is rare or when respondents tend to answer randomly or untruthfully (Peterson 2000).

Abbreviations

Abbreviations should equally be avoided. Using MNC[10] can cause confusion for the respondents who might assume a different meaning

[10] Instead of Multi-National Corporations or Manila National Company.

Example 3.5

Are You from New Delhi or from India?

III.4 Do you expect to make a substantial increase in investment in order to increase capacity or improve quality?
code: Yes=1; No=2; Firm is closing=3

In 2003 □
In 2003–2005 □

In this example the question includes two options not mutually exclusive. Hence if one respondent intends to invest in 2003 and in 2004 while another intends to invest only in 2003, they will both answer YES to both questions and we will not be able to discriminate among them.

than intended by the interviewer. The use of abbreviations may also generate confusion for the interviewers, especially when interpersonal relations are tense. Abbreviations should be spelled out clearly unless they are common to all respondents or have already been defined in the questionnaire (Plateck, Pierre-Pierre, and Stevens 1985).

Answer alternatives

In closed-ended questions, the selection of answer alternatives, in itself, could become a source of confusion for the respondent unless they are mutually exclusive and collectively exhaustive (Peterson 2000). This apparently simple requirement is sometimes overlooked (example 3.5) and it is occasionally hard to fulfill. In some instances, in fact, the set of possible alternatives is too broad and their "neutral" classification in groups hard to determine.[11]

Double-barreled questions

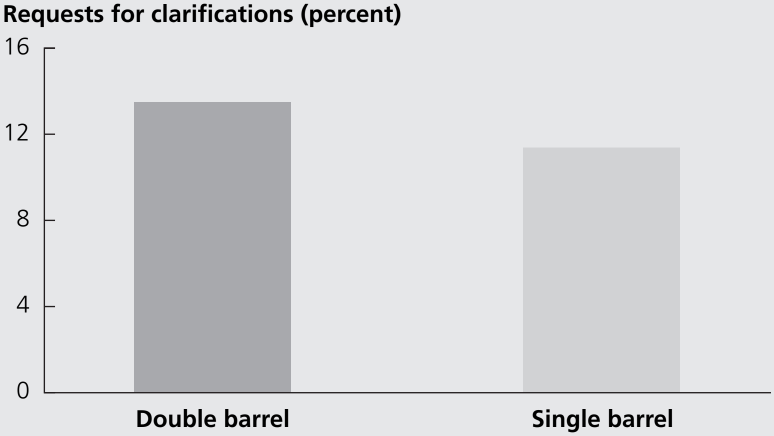
Another typical example of ambiguity is generated by double-barreled questions, that is, questions covering two or more issues at once. These questions cause uncertainty and confusion, particularly when both parts of the question apply to the respondent in different ways, and they usually require additional explanations (figure 3.5) As a consequence, they must be avoided. These type of questions should be divided into two questions or the choices provided for answering should cover all possible answer combinations.

Ability to answer

Finally, in evaluating the appropriate level of specificity to apply in a question, the designer should not ignore the ability of the respondent to answer. While it is important to ask specific questions, it is

[11] In this case, one possible solution is to adopt an open-ended question.

Figure 3.5.
Double-Barreled Questions Increase Requests for Clarifications



Source: Bassili and Scott 1996.
Note: For single barrel, the answer is the average of two simple questions (author's calculations).

equally essential to give the respondent an answer task that he or she can perform (Fowler 1995), as in example 3.6. It is very likely that the respondent knows the village where he or she sells his or her products, but at the same time, he or she might have no idea of its current population.

Example 3.6
Can You Tell Me How Many People Live in Pisa?

Sales and Supplies

13. a) What percent of your establishment's sales are:
    i) sold domestically
        a) to towns with 50,000 inhabitants or more ____%
        b) to towns with < 50,000 inhabitants ____%
    ii) exported directly ____%
    iii) exported indirectly (through a distributor) ____%
    TOTAL = 100%

Question Style

Unless respondents clearly understand a question, they will not be able to provide meaningful answers (Peterson 2000). Hence two fundamental concerns must be in the designer's mind when developing any question: will respondents be able to *understand* the question and will they be able to *answer* it? A well-understood question will not only increase the accuracy of the answers, but also their frequency. Two characteristics have a direct impact on these abilities: legibility and relevance.

Use Legible Questions

Ask questions that read well. This implies that conditional clauses, qualifications, and all other less important information must come ahead of the key content of the question. This placement prevents the respondent from jumping to an answer before the full question has been laid out. Likewise, punctuation should loosely follow proper grammatical rules and be more tailored to the flow of ideas stemming from the question. Thus, clarity is more important that grammatical correctness. This allows the interviewer to pause at the right time and place during the questioning. Similarly, words that need to be emphasized during the interview must be properly identified in the questionnaire and interviewers must be trained to recognize the identifiers. Finally, all words should be spelled out (Warwick and Lininger 1975).

Questions should not be formulated in a complex structure. Questions organized in a table format may appear well designed, and they give the impression of being easy to answer. However, they are extremely difficult to administer in a face-to-face interview and they put a big burden on the respondent's memory.

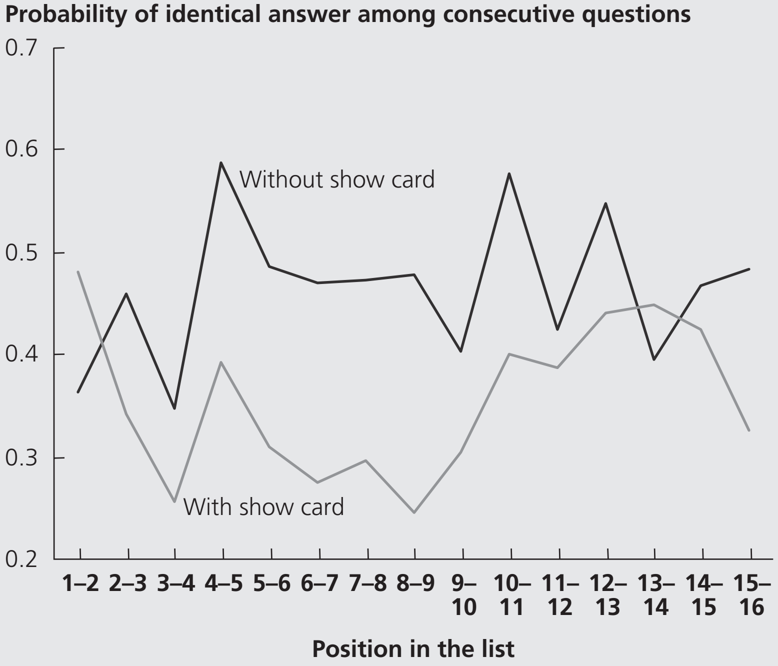
Similarly, the longer the list of questions the lower the quality of the data. "It is possible that respondents and/or interviewers recognize the 'production line' character of this survey strategy and that this promotes carelessness in the way questions are asked and answered" (Andrews 1984, 431).[12] Two types of errors can result from this behavior: acquiescence bias and position bias.

Acquiescence bias

Acquiescence bias is the tendency of respondents to choose a certain response category regardless of the item's content. For example, in questions with "agree-disagree," "yes-no," "0–5," and so on, respondents continually check "agree," "yes," or "2" even when the content of the

[12] Further experimental evidence has shown that these types of questions generate a higher tendency for the interviewer to misread them (Presser and Zhao 1992).

Figure 3.6.
Response Rate Effect of Long Repetitive Lists



Source: World Bank Investment Climate Surveys 2003 (author's calculations).

question is reversed. This is particularly true when a long sequence of questions in exactly the same order[13] is asked. This approach infringes conventional conversational norms and can easily become boring and irritating, pushing the respondent to answer mechanically without thinking carefully about the individual alternatives (Plateck, Pierre-Pierre, and Stevens 1985). When repetitive questions need to be asked, data accuracy can be improved by using show cards. Figure 3.6 illustrates how the adoption of show cards reduces the probability of obtaining the same answer (answering mechanically) among consecutive questions. When show cards are not used, evidence shows that consecutive questions are answered more mechanically (the probability of obtaining the same answer is higher). This effect appears to pick up when the list contains

[13] For example, questions like (1) How big of a problem is Telecommunication? (2) How big of a problem is Electricity? (3) How big of a problem is Transportation?

more than four repetitive questions. Using show cards instead helps eliminate this “contagion” effect, although only up to a point: when the list reaches 10 items the same “mechanical” effect reappears.[14]

One proposed solution to this bias is to give specific content to each response option. So instead of asking, “Do you agree or disagree that your current company is efficient in delivering packages?” you might want to ask, “Do you think that your current company is efficient or inefficient in delivering packages?” (Moser and Kalton 1971; Warwick and Lininger 1975). Yet a different possible strategy is to have two forms of the questionnaire in which the order of the alternatives is reversed or to have as many forms as the possible combinations of alternatives (Warwick and Lininger 1975).[15]

Position bias

In other cases, when the respondent is asked to select from a list of alternative answers, their choice may be affected by the order in which the alternatives are presented rather than true relevance to the respondent. When a set of alternatives is ordered, such as a set of numerical variables, respondents may consistently lean toward the middle, right, or left irrespective of the meaning of the order. Experiments have shown that the alternatives presented at the beginning or at the end are favored (Moser and Kalton 1971; see also figure 3.3). This phenomenon is more likely to occur when the list of alternatives is long, so the best solution is to use a short list of alternatives (no more than eight) or to elicit a response from each individual alternative. If it is not possible to reduce the list of alternatives, another useful approach is the filter-unfolding method. With this technique, major categories are first presented to the respondent. Then on the basis of his or her choice, a set of more specific alternatives are shown or the interviewer moves on to the next major category (example 3.7). This method optimizes the use of time by focusing only on the choices perceived by the respondent as most relevant. A less efficient solution is to use separate versions of the questionnaire, allowing each alternative to appear in a given position with equal frequency or use different show cards in which the order of the alternatives is different to make its position neutral.[16]

[14] This figure is based on question 18 of the Investment Climate Surveys’ core questionnaire (see appendix 1 for exact wording of the question). Data refer to pooled answers from Bangladesh2002, Brazil2003, Cambodia2003, China2002, Ethiopia2002, Honduras2003, India2002, Kenya2003, Nicaragua2003, Nigeria2001, Pakistan2002, Peru2002, Philippines2002, Tanzania2003, and Uganda2003.

[15] Avoid, however, using different orders within the same form. This could be misleading for the interviewer and the respondent.

[16] The last two alternatives carry a higher risk of error during data entering and coding.

Example 3.7

Filter-Unfolding Method

In this example, instead of asking questions d(1)–d(7), the interviewer first asks question d and only if the response is a 3 or 4 are options d(1)–(7) asked. Question d works as a filter for the more detailed questions d(1)–(7). This approach saves time and maintains focus during the interview.

V.2. Please tell us if any of the following issues are a problem for the operation and growth of your business. If an issue poses a problem, please judge its severity as an obstacle on a four-point scale where:

0 = No obstacle 1 = Minor obstacle 2 = Moderate obstacle 3 = Major obstacle 4 = Very severe obstacle

	No Problem	Degree of Obstacle			
a. Telecommunications	0	1	2	3	4
b. Electricity	0	1	2	3	4
c. Transportation	0	1	2	3	4
d. Access to land for expansion/relocation	0	1	2	3	4
1) Procurement process	0	1	2	3	4
2) Cost of land	0	1	2	3	4
3) Availability of infrastructure	0	1	2	3	4
4) Disputed ownership	0	1	2	3	4
5) Small size of land ownership	0	1	2	3	4
6) Government ownership of land	0	1	2	3	4
7) Others (please specify ______________________)	0	1	2	3	4
e. Tax rates	0	1	2	3	4
f. Tax administration	0	1	2	3	4
g. Customs and trade regulations	0	1	2	3	4
h. Labor regulations	0	1	2	3	4
1) Minimum wages	0	1	2	3	4
2) Mandatory non-salary benefits	0	1	2	3	4
3) Restrictions on employment of local staff	0	1	2	3	4
4) Visa/work permit for foreign staff	0	1	2	3	4
5) Hiring and firing regulations	0	1	2	3	4
6) Labor dispute settlement	0	1	2	3	4
7) Others (please specify ______________________)	0	1	2	3	4
i. Skills and education of available workers	0	1	2	3	4
j. Business licensing and operating permits	0	1	2	3	4
1) Constructing operational facilities	0	1	2	3	4
2) Fire department	0	1	2	3	4
3) Environmental clearance	0	1	2	3	4
4) Intellectual property, trademark registration	0	1	2	3	4
5) Company registration	0	1	2	3	4
6) Others (please specify ______________________)	0	1	2	3	4

Example 3.7 (continued)

k. Access to financing (e.g., collateral)	0	1	2	3	4
l. Cost of financing (e.g., interest rates)	0	1	2	3	4
m. Economic policy uncertainty	0	1	2	3	4
n. Macroeconomic instability (inflation, exchange rate)	0	1	2	3	4
o. Corruption	0	1	2	3	4
p. Crime, theft and disorder	0	1	2	3	4
q. Anticompetitive practices	0	1	2	3	4

Source: Author's creation.

Use Relevant Questions

Ask questions applicable to all respondents. Few things are more irritating than to be asked a question that is not applicable like "where did you complete your doctorate?" or "how many children do you have?" This is even more frustrating when elites are interviewed. As a matter of fact, while ordinary respondents are more willing to discuss issues about which they have little information, elites are quickly irritated if the topic of the questions is not of interest to them (Zuckerman 1972). Inapplicable questions are not only irritating but also potentially misleading. The individual who is not a parent may still give a positive answer to save embarrassment or simply to oblige the interviewer (known as false positives) (Warwick and Lininger 1975). One solution is to add proper lead-in questions and devise various skip patterns, filters, or conditional questions.

Hypothetical questions

Hypothetical questions, especially, should be avoided. People cannot reliably forecast their future behavior in a hypothetical scenario. Thus, the questionnaire designer should make careful use of this style of questioning. First, it is advisable to ask a question related to a hypothetical situation *only* of those who have already experienced the phenomenon in the past. For example, ask "Would you like to live in an apartment or in a house?" only of those who have lived in both. Otherwise you run the risk of picking up the answers of those who would like to try new things. Second, the designer should be wary of asking hypothetical questions in which the answer is obvious, such as "Would you like a reduction of metro fares?" In this case, answers are biased because the respondent is asked to get something for nothing (Moser and Kalton 1971).[17]

[17] Or, put differently, the question is not objective.

Use Painless Questions

Finally, the question asked should require the least possible effort to be answered. As the level of cognitive complexity of the question increases, the respondent is more likely to reply "I don't know" or, worse, inaccurately. If the researcher suspects that the respondent might not be candid in his or her answer, the researcher should not ask the question in the traditional way but rather adopt alternative strategies to elicit a truthful answer (Peterson 2000).

Question Type

Avoid Sensitive Questions

A sensitive question refers to a behavior that, when answered truthfully, is judged by society as undesirable or illegal, or when the question itself is perceived by the respondent as an invasion of privacy. Sensitive questions should be avoided. Two types of respondent's behavior can threaten the accuracy of answers to sensitive questions: nonresponse and response error. Respondents might refuse to answer sensitive questions, thus biasing the results because "the very persons with the most sensitive information to report may be the least likely to report it" (Tourangeau and Smith 1996, 276). Likewise, research on response accuracy has shown that respondents are prone to distort answers in ways that will make them look better (known as *social desirability bias*). Responses on illegal or immoral behavior, such as corruption, are consistently underreported not because respondents have forgotten them but rather because the behavior does not conform to social norms (Fowler 1995).

Survey firm bias

When compared internationally, sensitive questions are subject to another often overlooked source of potential measurement error associated with the type of survey firm conducting the interviews. The survey literature clearly identifies the interviewer's sponsoring agency as a potential source of measurement bias. Moreover, research has demonstrated that sensitive questions are answered more or less candidly depending on the person conducting the interviews. Evidence from the Investment Climate Surveys not only confirms this, but also it allows the estimation of such bias. The type of agency conducting the fieldwork—government agency, a private local survey company, or a private international survey firm—has a different effect on data accuracy depending on the sensitivity of question asked. Sensitive questions on corruption and taxation show a different pattern than questions on red tape. Not surprisingly, when the interviewer is a government employee, sensitive questions on bribes and sales are consistently underreported.[18] Although the magnitude of the bias varies depending on the specific question, the impact of underreporting appears to be up to 60 percent of one standard deviation (table 3.2). This phenomenon is present also with questions on the perception of corruption.[19] Hence, respondents interviewed by a government employee are 13 percent less likely to rate corruption as a major concern (figure 3.7). On the contrary, when the same objective sensitive questions are asked by an international survey firm, no underreporting effect appears in the data. Again only the perception question on corruption shows the same magnitude of underreporting (11%).

When questions on red tape are asked, however, whether the survey firm is a government agency or an international company, a similar pattern of under- or overreporting appears in the answers. Such a bias seems even higher for the latter than for the former (table 3.2).

When designing and analyzing survey data, it is important to keep in mind that people may vary in what they consider sensitive. Questions on apparently nonsensitive issues, such as questions on infrastructure, are subject to the same measurement bias discussed above. When the interview is conducted by a government employee, respondents tend to underreport such constraints, although the impact is most of the time relatively small (table 3.2).

Strategies to minimize bias

When sensitive questions are asked, two major forces operate to produce a distortion in the reported answer: the desire to avoid responses that could pose a threat and the tendency of respondents not too look bad (Fowler 1995). A number of different steps can be taken to minimize these forces. First, the level of detail of the question can be tailored to address sensitivity concerns. It might be easier for respondents to provide answers in categories or percentages rather than in absolute values. When following this approach, however, the questionnaire designer should be aware of the fact that changing the format of the question has an impact not only on response rate but also on response accuracy. Peterson and Kerin (1980) have shown that while the refusal rate for an open-ended question on income was higher (8%) than that

[18] Compared with a private local survey company.

[19] See appendix 2 for exact wording of objective questions and appendix 3 presents the parametric results. Appendix 1 reports the perception questions. All questions are from the "core" questionnaire of the World Bank Investment Climate Surveys.

Table 3.2

Interviews Conducted by Government Agencies and International Private Firms Affect Survey Responses

	Government Agency		International Private Firm	
	Coefficient	Effect on Standard Deviation	Coefficient	Effect on Standard Deviation
CORRUPTION				
Unofficial payments to get things done (% sales)	–1.854**	–0.33	ns	
Gifts expected as % value government contracts	–1.974**	–0.41	ns	
Estimated % of total sales declared for tax purposes	–17.854**	–0.60	ns	
RED TAPE				
% of management's time dealing with gov't officials	–0.709*	–0.05	–6.256**	–0.43
Total days spent with officials from tax inspectorate	2.307**	0.13	4.803**	0.26
Days on average to claim imports from customs	ns		–4.150**	–0.34
Days on average to clear customs for exports	1.815**	0.22	ns	
Optimal level of employment (% of current level)	2.769**	0.06	30.147**	0.64
INFRASTRUCTURE				
Days of power outages from public grid	–31.998**	–0.49	–16.366**	–0.25
Days of insufficient water supply	–2.673**	–0.05	ns	
Days of unavailable mainline telephone service	–3.69**	–0.12	ns	
% of sales lost because of power outages	–0.229*	–0.03	ns	
% of sales lost because of insufficient water supply	0.653**	0.08	ns	
% of sales lost because of unavailable telephone service	–0.844**	–0.15	ns	
% of average cargo value lost in transit	–0.176*	–0.04	–1.629**	–0.33

Source: World Bank Investment Climate Surveys 2003. (Author's calculations.)

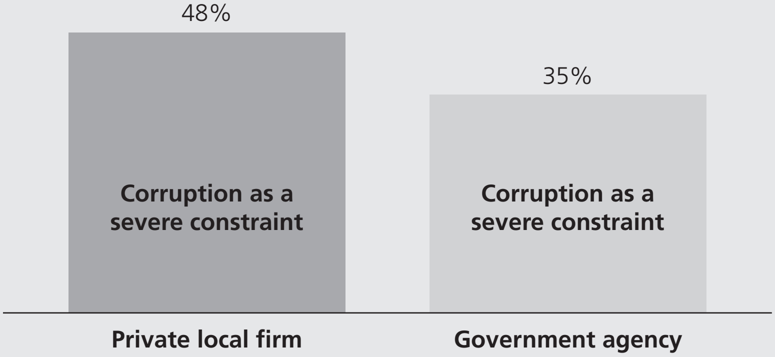
Note: See appendix 2 and 3 for description of questions and a compete set of regression results. Results are in comparison with private local survey company.

ns = Not significant.

* Significant at 5%.

** Significant at 1%.

Figure 3.7.
Interviews Conducted by Government Agencies Decrease Reports of Corruption



Source: Author's calculations.

on a closed-ended question (2%) the quality of the answers to narrative questions was higher (table 3.3).

Second, as mentioned earlier, the length of the question itself can also mitigate the threatening nature of the topic. Longer questions seem to have a positive impact on the accuracy of sensitive questions on behavior, while the opposite appears true for attitudinal questions (Sudnam

Table 3.3
Accuracy is Higher for Open-Ended Questions

	Percent of Respondents	
	Open-Ended Question	Closed-Ended Question
Overreported	26.7	29.7
Accurate	47.4	43.8
Underreported	25.9	26.5
Actual-reported correlation coefficient[a]	0.93	0.70

Source: Peterson and Kerin 1980.
a. Using midpoint of range category.

and Bradburn 1974). A similar approach is to moderate the extent to which respondents feel that their answers will be used to put them in a negative light. With this strategy, the question is asked in a way to explain to the respondent that there are various reasons why people behave in one way or the other so that he or she feels more relaxed in providing an unbiased answer to the sensitive topic. In other cases, it might be appropriate to ensure the confidentiality of responses and communicate effectively that protection measures are in place. This implies that no association between respondents and answers should be apparent during the interview, that sensitive questions should be asked only when the respondent is alone with the interviewer, and that, if they exist, specific laws protecting the confidentiality of answers should be mentioned and clearly stated in the questionnaire. Explaining the appropriateness of the question to the research objectives of the survey is yet another way to reduce resistance in respondents (Fowler 1995). Sometime respondents consider a question sensitive because they don't see the link between the goal of the survey and the question itself, or they don't see the usefulness of their answer. Likewise there are instances in which it is advisable to use words that imply the same sensitive behavior by others. Finally, another possible way of dealing with sensitive questions is to put the threatening topic in a list of less threatening topics or to use a randomized response technique (Plateck, Pierre-Pierre, and Stevens 1985).[20]

In business surveys, the inclination not to disclose information considered critical to the business activity should be taken into account when developing a survey instrument. Questions on taxes, profits, and names of suppliers or clients could be the subject of distorted answers or outright refusal. Conversely, questions on bribes are generally answered, unless the admission of this behavior is in itself condition for criminal prosecution. This was the case in Ethiopia where, although it was not possible to ask the amount of bribes paid by entrepreneurs, because this would have guaranteed jail time, 60 percent of the respondents were still willing to discuss how big of a problem corruption was. The pre-test is critical in detecting the respondent's reaction to a delicate question.

[20] With this technique, the respondent chooses to answer either the sensitive question or a nonsensitive question. The process of choosing uses a random mechanism. The interviewer is not aware whether the respondent is answering the sensitive question or the nonsensitive statement. In this way, we expect the respondent to be more truthful. At the same time, the probability of selection of each statement must be known so that it is possible to calculate the aggregate value of the sensitive question as a weighted average of the probability of selection of each statement (See Moser and Kalton 1971, 328).

Memory Questions

Retrospective questions are those questions drawing on long-term memory. Although past events are not really forgotten, their recall might be very memory intensive and therefore incorrectly reported. How many of us can remember the exact number of days we were sick two years ago? Or how many classmates we had the first year of high school? Recall questions should hence be avoided.

Because "little is known about how time-related information is mentally encoded, stored, and retrieved [and] the effect of mood or motives on memory" (Peterson 2000, 93) retrospective questions are subject to "recall bias." Survey research has identified three types of memory errors: respondents might forget the recalled events (omission), might recall events that did not occur (commission), or might correctly report events but place them at the wrong time (telescoping) (Gaskell, Wright, and O'Muircheartaigh 2000).[21]

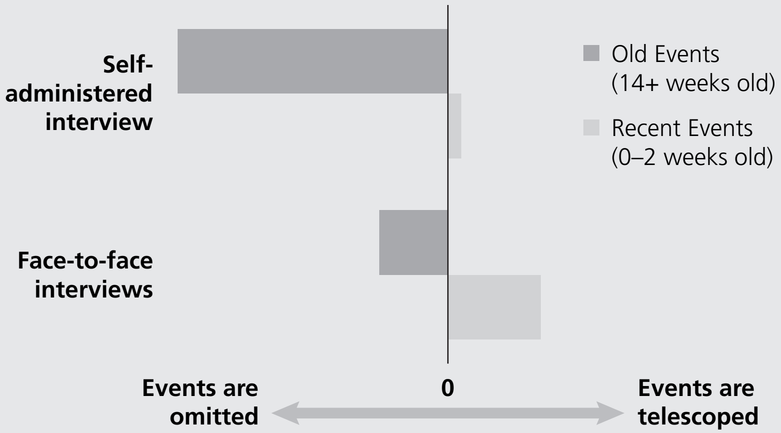
Recurring event

Behavioral frequency questions such as "How many times last year did you visit a doctor?" are a common occurrence in many surveys. Because frequency questions rely heavily on the respondent's mnemonic ability, they require a deep understanding of the cognitive process behind it. Questions that ask respondents to recall events in the past are an increasing source of error the farther back in time the event is located and the less important the event was in the respondent's life. Research shows that recall errors are associated with the interview mode and that there is always a trade-off between the accuracy of the event reported and the length of time of the event recalled (Fowler 1995). As one would expect, self-administered questionnaires appear more vulnerable to these errors than face-to-face interviews (figure 3.8) (Sudnam and Bradburn 1974).

How far back in time can a question go without seriously compromising the accuracy of the data collected? This depends on three factors: the *saliency of the event, its frequency, and how the question is designed*. More relevant events will be recalled more accurately. Unfortunately, the reality is that "contrary to what many researchers [. . .] might think (or desire!) much of what is investigated is not significant to study participants. Thus, the container size of toothpaste used in

[21] Telescoping can be "backward" if the event is reported to happen before it actually did, or "forward" if the event is reported to happen after it actually did. Telescoping occurs when the memory of the past event is so detailed that the respondent mistakenly blends recency with clarity (Bradburn, Rips, and Shevell 1987).

Figure 3.8.
Index of Memory Error by Mode of Interview



Source: Sudnam and Bradburn 1974.

1991 is probably long forgotten" (Peterson 2000, 20). Three factors influence the saliency of the event: the emotion generated by the event, the marking of a turning point, and its financial impact on the respondent's life (Auriat 1993).

In addition to saliency, another factor influencing the cognitive abilities of the respondent is the frequency of the event to be recalled. Respondents use different protocols to answer frequency questions: episode enumeration, rule-based estimation, availability estimation, automatic estimation, and various combinations of these protocols.[22] In deciding which protocol to use, respondents balance the level of accuracy they feel must be achieved with the level of effort required by the cognitive process itself. Empirical evidence shows that the accuracy of recall can be improved if episode enumeration is adopted. However, respondents will use episode enumeration only if there are not too many events to recall (Burton and Blair 1991). Because more distant events are harder

[22] Episode enumeration implies recalling and counting the occurrences of the event; rule-based estimation involves the use of some sort of rule, such as decomposing the time period in shorter time periods or in subdomains; availability estimation involves estimation on the basis of the ease of recalling; and automatic estimation involves resorting to some sort of innate or normative sense of frequency (for example, once a week).

to locate and to retrieve, the longer the time frame, the fewer respondents will adopt episode enumeration (Blair and Burton 1987).

While the questionnaire designer cannot change the frequency of the event, he or she can adjust the wording of the question to facilitate the use of episode enumeration. The National Crime Survey and the National Health Survey improved the accuracy of data by asking about six-month reporting instead of one year (Fowler 1995). But what time period would most likely promote the adoption of episode enumeration? Blair and Burton (1987) show that respondents are less inclined to use episode enumeration if the event happens more than 10 times during the reference period (table 3.4). Consistent with this result, Burton and Blair (1991) show that when holding the time reference constant an increase in the number of events within that period appears to be associated with a decrease in the accuracy of responses (table 3.5). This demonstrates that it is not time reference

Table 3.4

As Frequency of Event Decreases, Use of Episode Enumeration Increases

	Percentage of Respondents Using	
	Episode Enumeration	Other Enumeration
Frequency of event		
1	100	0
2	68	32
3	93	7
4–5	63	37
6–10	15	85
11–25	0	100
26–100	0	100
100+	0	100
Time frame		
2 weeks	56	44
2 months	25	75
6 months	4	96

Source: Blair and Burton 1991.

Table 3.5
Higher Event Frequency Has a Negative Effect on Accuracy

	Number of Events Reported[c]	Correlation Between Reported and Recorded Data
Checks[a]	16.3	34%
ATM[b]	4.2	67%

Source: Burton and Blair 1991.
a. Number of checks written.
b. Number of ATM withdrawals.
c. Average value over a 3-week period.

that increases accuracy but rather the number of events to be recalled. Thus, the time referenced in the question should cover approximately 10 episodes of the event to be recalled. For example, if there are five power outages every week, the question should ask how many power outages have there been in two weeks.

The wording of the question has also an impact on the ease of recall. One way to stimulate recall is to ask a long, rather than short question. This means adding an introduction that helps the respondent to put his or her state of mind in the time period of the event recalled. A second approach is to ask multiple questions or to ask questions that trigger associations with the event recalled (called a landmark). These methodologies have been proven to facilitate recall because the respondent is asked to dig into his or her memory (Fowler 1995). Furthermore, it has been shown that communicating to respondents the importance of the accuracy of their answer has a positive effect. Thus, using specific phrases like "please take your time to answer this question," or "the accuracy of this question is particularly important," or "please take at least 30 seconds to think about this question before answering" has a positive impact on the accuracy of responses (table 3.6) (Burton and Blair 1991). Similarly, asking "how often" as opposed to "how many times" might discourage episode enumeration in favor of rule-based estimation (Blair and Burton 1987). Finally, if the recall question asks the respondent to provide a list, it is desirable to provide him or her with a comprehensive and mutually exclusive list of events (Moser and Kalton 1971).

Table 3.6
Response Time and Episode Enumeration Have a Positive Effect on Accuracy

Time Given	Correlation Between Reported and Recorded Data	No. of Respondents Using Episode Enumeration
10–20 seconds	0.58	60%
35–50 seconds	0.81	80%
70 seconds	0.86	84%
unspecified	0.46	

Source: Burton and Blair 1991.

Unique event

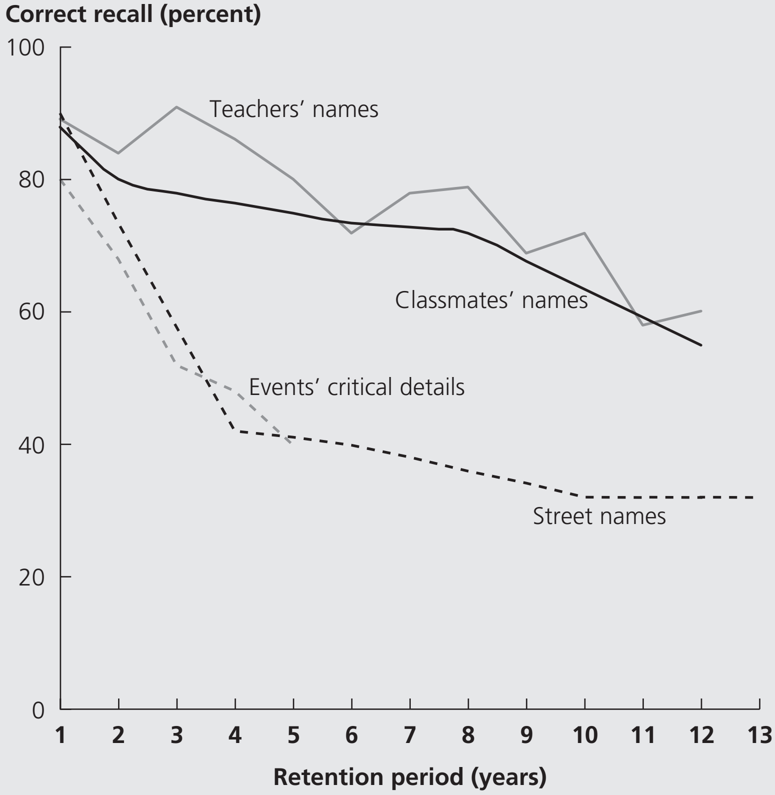
While the questionnaire designer can, to some extent, improve the accuracy of the recall through a careful question design, there are cases in which his or her ability is severely limited. This happens when the past event to be recalled is unique, such as "How many employees did you have when you started your business?" In this case, the questionnaire designer cannot adjust the time reference of the question to facilitate recall nor can episode enumeration be encouraged. In this case, accuracy rests solely on the recall ability of the respondent. Saliency remains the only critical factor to which the designer must appeal. A number of experiments have attempted to determine the accuracy of such a recall. A first experiment shows that between 10 and 20 percent of details are irretrievable after only one year, and as much as 60 percent can be lost after four years. In any case, even salient events are very hard to access and retrieve after 10 years (figure 3.9) (Sudnam, Bradburn, and Schwarz 1996). Even higher nonresponse rates were reported in a study carried out by the U.S. National Center for Health Statistics. That study shows the percentage of underreporting errors is more than 40 percent after just one year (figure 3.10)

Subjective (or Objective) Questions

"Subjective phenomena are those that, in principle, can be directly known, if at all, only by the persons themselves. [. . .] Objective phenomena are those that can be known by evidence that is, in principle, directly accessible to an external observer" (Duncan, Fishhoff, and

Figure 3.9.

Accuracy of Recall Decreases Over Time



Source: Sudnam, Bradburn, and Schwarz 1996.

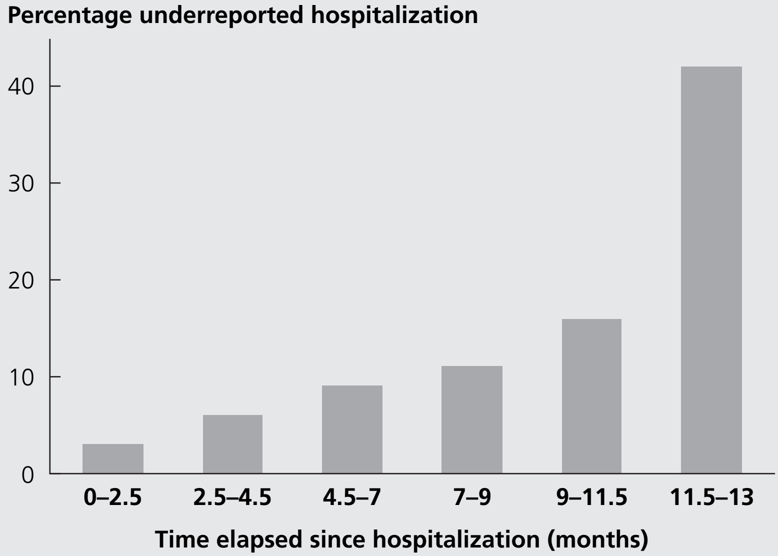
Turner 1984, vol. 1, 8). Subjective questions are questions tailored to measure people's subjective states (that is, their opinions, knowledge, feelings, and perceptions).[23]

Rating scales

One of the most popular ways of asking a subjective question is to use rating scales, that is, a single, well-defined continuum in which the answer is expected to be placed (see example 3.8) When employing this type of questions two issues must be addressed: how many scale

[23] That is why they are often referred to as perception or opinion questions.

Figure 3.10.
Effect of Time on Underreporting Hospitalization



Source: Foddy 1993.

categories should be used, and what words or numbers should be associated with each scale category.

There is no consensus in the literature on the optimal number of categories to use. Using too few categories gives less-refined information while using too many categories makes the question very hard to administer (Fowler 1995). The choice of scale frequency must be guided

Example 3.8
Rating Scales in Subjective (or Perception) Questions

Q.42 To what extent do you agree with this statement? "I am confident that the legal system will uphold my contract and property rights in business disputes."

Strongly Disagree	Disagree in Most Cases	Tend to Disagree	Tend to Agree	Agree in Most Cases	Strongly Agree
1	2	3	4	5	6

Example 3.9

Do You Agree or Disagree with "The Government is Providing Good Services"?

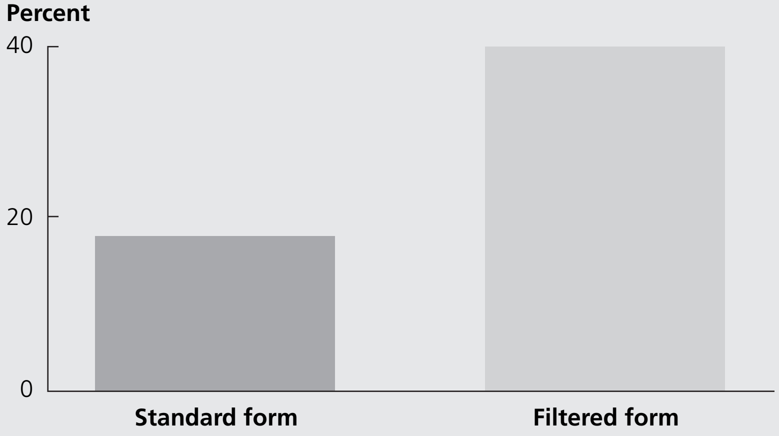
A.	________	________	________	________	________	
	Strongly Agree		Neither Agree Nor Disagree		Strongly Disagree	
B.	________	________	________	________	________	________
	Strongly Agree		Neither Agree Nor Disagree		Strongly Disagree	No Knowledge or No opinion

by the mode of the interview, the respondent's ability to interpret the categories, and the research goal.[24] Although some scales adopt up to 12 categories, experiments show that it is preferable to use between 5 and 9 categories (Cox 1980; Finn 1972; Leigh and Martin 1987; Miller 1956). Related to this is the decision whether to adopt a middle category. Here again no one choice fits all. It is not clear whether the presence or exclusion of a midpoint improves data quality (Andrews 1984). The content and the analytical purpose of the question will determine the optimal choice. If a neutral answer is a possibility, a midpoint should be included. If the researcher wants the respondent to take one side or the other, an even number of categories can force the respondent away from the middle alternative (Peterson 2000). The researcher, however, should be extremely careful in the latter case. Forcing respondents to choose an alternative with which they are not familiar has little analytical value and introduces bias into the data.

More important, the questionnaire designer must ensure that the scale categories are sufficient to discriminate between and "indifferent" response and "no opinion" answers. As example 3.9 shows, allowing respondents to "opt out" if they lack the required information (option B) improves the quality of the data because it avoids the risk that respondents with no opinion on the subject might otherwise select the middle alternative.

[24] In telephone interviews and in self-administered surveys, it is advisable to adopt a lower number of categories. The same is true if we interview a child rather than an adult. If, on the other hand, there is reason to believe that respondents are homogeneous, a higher frequency of categories should be adopted (Peterson 2000).

Figure 3.11.
Respondents Answering "Don't Know" Increases with Option to Opt Out



Source: Peterson 2000.

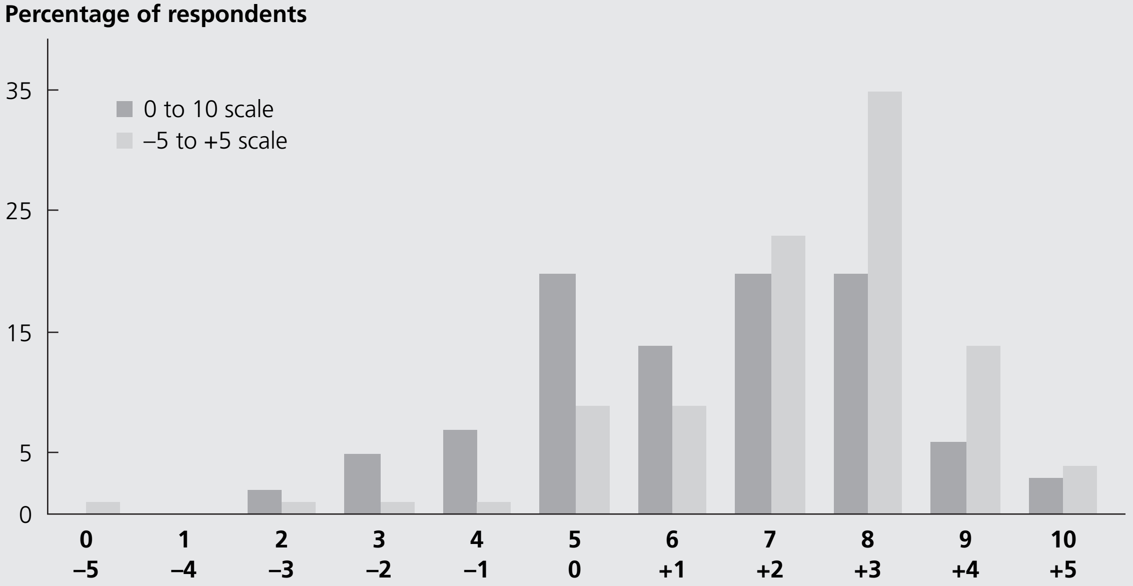
An experiment conducted by Peterson (2000) demonstrated that more than 20 percent of respondents venture an answer although they would have chosen "don't know" if given the option (figure 3.11).

Once the optimal number of categories has been determined, there are numerous ways respondents can be asked to assign answers to a position on a continuum. Generally adjectives (that is, good, fair, poor) or numbers[25] (that is, 1, 2, 3) are used. The ongoing debate is whether to use one approach or the other, and whether to label all categories in the continuum or only some of them. What seems an apparently trivial task can have a substantial effect on the data collected. As a matter of fact, far from being "neutral measurement devices," the response categories offered are used by respondents to interpret the question and therefore can influence the answers. Thus, asking the same question but using an 11-point rating scale from 0 to 10 or from –5 to +5 can generate different results (figure 3.12).

[25] Peterson (2000) reports a number of different verbal stimuli most commonly used: comparison stimuli (much more . . . much less), endorsement stimuli (definitely true . . . definitely not true), frequency stimuli (always . . . never), influence stimuli (major problems . . . minor problems), and intensity stimuli (strongly agree . . . strongly disagree).

Figure 3.12.

Impact of Numeric Scales on Response Patterns



Source: Schwarz and others 1991.

Accordingly, Schwarz and others (1991) conclude that scales intended to measure bipolar concepts should use negative to positive values, while scales intended to measure the intensity (or absence of intensity) of a single attribute should use zero to positive values. In general, evidence shows that adjectives are better because they improve consistency of interpretation (Wildt and Mazis 1978), and because it might generate cognitive complexity in the respondent, it is not preferable to label all categories. It is preferable to label only the extreme values and the middle point (Andrews 1984).

Rank ordering

A second approach used by researchers in designing qualitative questions is the use of rank ordering. Respondents are asked to compare objects on some dimension. They are provided with a list of items and asked (1) to rank all of them from most to least important; (2) to identify the most (second, third) important; or (3) to rate each of the items using some scale. The first task is the easiest only if the list is short (four to five items). The second one is preferable if the list is long, but it becomes difficult if the list grows too long (more than 10 items). The third alternative is the easiest to perform and provides more information, but it is the most time-consuming (Fowler 1995).

Agree-disagree format

If the goal of a question is to measure respondents' feelings about policies or ideas, then the agree-disagree format is probably the best option. Respondents will agree or disagree if their opinion falls within a reasonable distance from the point where they see the statement's opinion is located. Sometimes this type of question is employed to spread the responses along a continuum, differing from the previous category in that the continuum is simply bipolar (Fowler 1995).[26] Attention should nonetheless be paid when developing this type of question. First, the designer must ensure that the contextual environment is appropriate for this format. Cognitive complexity is increased when we ask someone whether they agree or disagree with the statement "My health is good" instead of asking directly "How do you rate your health? Good, Fair, or Poor." Second, the designer must ensure that the wording of the question is unequivocal so that disagreeing can be interpreted unambiguously. If a person disagrees with "I am sometimes depressed," we don't know whether he or she is never depressed or always depressed. Finally, the designer must not overlook the tendency among less-educated respondents toward acquiescence, that is, to consistently answer "agree" even when they don't know whether they do or don't (Fowler 1995).

The questionnaire designer should pay particular attention to this type of question because it is very easy to be misled by its apparent simplicity. Suppose we ask a respondent whether he or she agrees with an increase in taxes to improve parks. The respondent's role, implied by this type of question, is to figure out whether the policy alternative is close enough to his or her views to "agree" with the statement. This, however, is based on a subtle but critical assumption: that the respondent has some general opinion about this specific issue. The risk of this question is that if the respondent has no opinion about this issue, but he or she is generally opposed to raising taxes *for any reason,* then he or she will "disagree," in effect answering a different question (Fowler

[26] Sometime four categories are used, such as strongly agree, agree, disagree, strongly disagree. In many cases, however, the answers are usually analyzed in two response categories, so such questions do not yield much more information. Some researchers also consider a question so worded as emotionally charged, violating the objectivity rule of good question design (Fowler 1995).

1995). This violates the fourth criteria of good question design, specificity: respondents should all answer the same question. Put differently,

> Respondents read meaning into the question and answer in terms of some general predisposition toward the [issue. . . .] When people rely on such a predisposition they are showing ideological or constrained attitudes, since they are using general attitudes to supply responses to specific questions. (Smith 1984, 223)

Limitations of subjective questions

Consequently extraordinary care must be taken when designing and interpreting subjective questions. There is no doubt that questions of this type are easy to administer and easy to answer. Their response rate in Investment Climate Surveys tops 99 percent. However, what is easy to get is not necessarily easy to use. Researchers should be wary that subjective questions have the following three analytical limitations:

Plausibility of the answers. A key feature of these questions is that there are no right or wrong answers independent of what respondents tell. While we can, to some extent, measure errors in reporting the percentages of sales that went to bribes, there is no way to assess the accuracy of answers on a six-point scale on the severity of corruption. There is, in fact, no objective standard against which to measure the rightness of that answer (Fowler 1995).

Comparability of the responses. No matter whether we use adjectives or numbers to define the continuum, different respondents may interpret the same categories differently. Hence, when Scipione (1995) conducted an experiment in which a group of respondents were asked what value each associated to "majority," the answers ranged from 38 percent to 76 percent with a mean value of 57 percent (table 3.7).

Furthermore, the researcher cannot be sure that the same answer from different respondents has the same weight, or that the same reply to different questions by the same respondent has the same meaning. In other words, there is no guarantee that "agree" from one respondent is different from "strongly agree" from another respondent. If we ask respondents to rate their social standing on a 1 to 10 scale, we cannot concluded that those rated 9 are three times as high as those rated 3 (Fowler 1995). Finally, the analyst should not "confuse the 'extremity' of judgments with the 'importance' of topics for respondents and with the 'certainty' or 'sureness' of their responses" (Foddy 1993, 160). In answering these types of questions, people put their perceptions up

Table 3.7
Perceived Percentage Values Associated with Descriptive Words

Expression	% Value	Standard Deviation
An overwhelming majority	74	19
A substantial majority	67	20
A large majority	62	21
A majority	57	19
A large minority	41	21
A substantial minority	32	16
A minority	24	21
Most	69	21
Hardly anyone	12	20
Much more than	33	22
Somewhat more than	31	25
Somewhat less than	30	25
Much less than	26	13
A slight change	20	28

Source: Scipione 1995.

against a self-imposed standard unknown to the analyst rather than an objective standard. What one respondent perceives as "good" may be considered only "fair" or even "poor" by another (Fowler 1995). This is because opinions on virtually any issue are often many sided, with cultural, moral, religious, legal, medical, professional, or even geographic dimensions. So an inhabitant of Canada will have different perceptions on what constitutes a cold winter from an inhabitant of Indonesia. During a pilot test in Ethiopia in December 2001, I interviewed the manager of the St. George's Beer factory. He was extremely interested in our project and he answered many questions, often looking at his laptop to provide the most accurate information. When I asked him about the severity of corruption in Ethiopia, he told me that corruption was a minor problem. Given the previous discussions I had had with other managers and experts in that country, his answer surprised me. Later on during the interview I found out why he rated corruption as minor. He had just moved from Nigeria where corruption was rampant. To him the level of corruption in Ethiopia was low because he used a different standard of reference: Nigeria's corruption level. This practically meant that in his mind the question I asked him was "How do you rate corruption in Ethiopia *compared to Nigeria?*"[27] "If respondents adopt different perspectives when answering a question, it can be argued that they are not answering the same question. If this happens, the answers that they give cannot be meaningfully compared" (Foddy 1993, 79). All this implies that,

> answers to questions about subjective states are always relative [and] never absolute. The kinds of statements that are justified, based on answers to these kinds of questions, are comparative. It is appropriate to say that Group A reports more positive feelings than Group B. It is appropriate to say that the population reports more positive feelings today than it did a year ago. It is not appropriate [. . .] to say that people gave the president a positive rating [or] that they are satisfied with their schools. (Fowler 1995, 72–73)

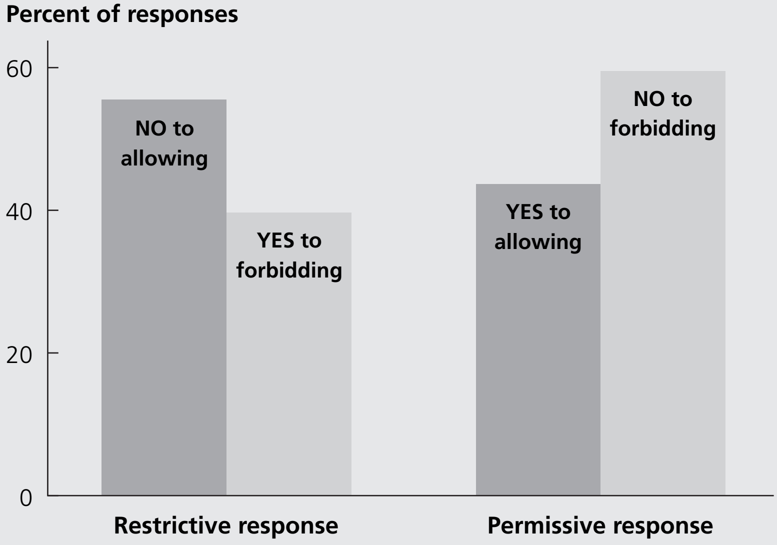
Reliability of the respondent. Questions based on the subjective assessment of the event are subject to idiosyncratic factors, such as the respondent's mood at the time of the interview, the wording of the question, or even external events that might cause the respondent to present only one aspect of his or her reaction to the object of the question (Dexter 1970; Narayan and Krosnick 1996). Words that appear to be the same to researchers often are not so from the point of view of the respondent. When respondents were asked, "Do you think that one should generally *forbid* the use of salt?" or "Do you think that one should generally *allow* the use of salt?" 62 percent of respondents sided in favor of forbidding it and 79 percent in favor of not allowing it (Hippler and Schwarz 1986). See figure 3.13 for an analogous example.

Similarly, using words like "dealing with drug addiction" rather than "drug rehabilitation" elicits a more active stance on the issue (Rasinski 1989).[28] Different words might stimulate different feelings and generate different reactions. In 1940 Cantril and Wilks showed that the percent-

[27] This example is additional proof that using the same rating scale to compare answers across countries creates serious methodological problems.

[28] This is one of the reasons why enumerators should read the question exactly as it is stated in the questionnaire.

Figure 3.13.
Negative or Positive Words Influence Respondents Differently



Source: Schuman and Presser 1981.

age of people supporting U.S. involvement in World War II almost doubled if the word "Hitler" appeared in the question.[29] Half a century later, following a controversial opinion poll commissioned by the European Union, the European Commission stated that "it would *change this unfortunate perception by asking the question differently* in future"[30] (*The Economist* 2003, 8).

An external factor that can "influence" answers to subjective questions is, once again, the affiliation of the interviewer. Perception questions appear to be more subject to this bias than objective questions. Furthermore and contrary to what happens with objective questions, this measurement error is more evident if the interview is conducted by an international survey firm than by a government agency. Evidence from the Investment Climate Surveys shows that, on average, compared to a private local survey agency respondents interviewed by an international

[29] Thirteen percent in favor without "Hitler" and 22 percent in favor with the word "Hitler."
[30] Italics added by the author.

survey firm are 10 percent less likely to rate any bottlenecks as a major constraint. A similar bias, although lower in magnitude (4%), is present when a government agency is conducting the interviews (table 3.8). The presence of this different fixed effect makes the international comparison of perception questions much harder.

Table 3.8

Interviews Conducted by Government Agencies and International Private Firms Reduce Probability of Rating Major Constraints

	Government Agency	International Private Firm
A. Telecommunications	2%	–5%
B. Electricity	0%	–14%
C. Transportation	0%	–12%
D. Access to Land	6%	–3%
E. Tax Rates	1%	–9%
F. Tax Administration	3%	0%
G. Customs and Trade Regulations	–4%	–8%
H. Labor Regulations	0%	–10%
I. Skills and Education of Available Workers	2%	–8%
J. Business Licensing and Operating Permits	3%	–3%
K. Access to Financing (e.g., collateral)	–9%	–13%
L. Cost of Financing (e.g., interest rates)	–16%	–18%
M. Economic and Regulatory Policy Uncertainty	–12%	–12%
N. Macroeconomic Instability (inflation, exchange rate)	–1%	0%
O. Corruption	–13%	–11%
P. Crime, theft, and disorder	–18%	–17%
Q. Anticompetitive or informal practices	–12%	–27%
R. Legal system/conflict resolution	–2%	–17%
Average effect	–4%	–10%

Source: Calculations based on World Bank Investment Climate Surveys 2003.
Note: 0% change means no significant difference; negative sign means less likely; positive sign means more likely.
All results presented are significant at 1% or 5%.
Results are in comparison with the results of the interview being conducted by a private local survey agency.
Results for international private firms refer only to East Europe and Central Asian countries.
The exact wording of the questions is presented in appendix 1.

One approach researchers have come up with to address the fundamental problem of subjective questions, and their inability to provide quantitative measures that can be compared across individuals, is to use magnitude estimation. With this technique, the respondent is asked to rate a phenomenon by comparing it with another phenomenon for which a rating has already been assigned. For example, one such question for a physician would be as follows:

> Suppose we want to compare the amount of work involved in a splenectomy with the amount of work involved in a tonsillectomy. If we assume that the amount of work involved in a splenectomy is 10, what number would you assign to the work involved in a tonsillectomy?

This approach produces more reliable responses because it introduces an objective point of reference for respondents to rate the phenomenon. However, this technique has its limitations. First, it cannot be used for many of the subjective states that researchers want to measure. Second, the respondents must be able to understand the technique itself, a task that requires a certain level of cognitive ability. And, finally, it takes a fair amount of training by respondents, which is time-consuming. As a result, this approach is not commonly used in surveys (Fowler 1995).

The best solution is to move away from subjective questions. In some cases, this is easier than it may seem. Consider, for example, the subjective question in example 3.10.

The question is better addressed by the corresponding objective question in example 3.11.

Example 3.10

Subjective Question

Business-Government Relations

34. How would you generally rate the efficiency of government in delivering services (e.g., public utilities, public transportation, security, education, and health). Would you rate it as (*read 1–6*)?

1. Very inefficient
2. Inefficient
3. Somewhat inefficient
4. Somewhat efficient
5. Efficient
6. Very efficient

Source: Investment Climate Surveys.

Example 3.11

Objective Question

H3. What is the share of government officials that deliver efficient services (e.g., public utilities, public transportation, security, education, and health)?

_______%

Source: Investment Climate Surveys.

When a subjective question is the only way to ask a question, the researcher should refrain from rushing into the analysis without first looking at the possible factors that might influence the respondent. Thus, it is important for the same subjective issue to be addressed from different angles. The presence of inconsistencies among answers to subjective questions by the same respondent remains the only critical source of information on their "quality."

Narrative Questions

What characterizes a narrative or open-ended question is the freedom enjoyed by respondents to answer with their own words. Because open-ended questions do not force respondents into a set of predetermined answers, this is the only type of inquiry that allows them maximum spontaneity of expression. Furthermore, not being influenced by predetermined alternatives allows the researcher to identify the respondent's level of knowledge and information, the salience of the event, the strength of his or her feelings, and his or her motivational influences while avoiding format effects (Foddy 1993).

Open-ended questions have their own set of drawbacks. They take more time and effort than closed questions. As a consequence, they have a higher refusal rate and a higher cost per completed questionnaire. Secondly, the freedom they give to respondents generates a higher variability of answers. "Because of different word choices, verbal skills, and the like, study participants seldom give identical answers, even though they may be saying essentially the same thing" (Peterson 2000, 33). The associated diversity of answers results in a great variety of interpretations making the analysis extremely labor intensive (Peterson 2000). Finally, narrative questions rely more heavily on the interviewer's ability and experience. The more expansive and complex the respondent's answer (verbosity effect) the more important the interviewer's ability to probe and the greater the risk that only those aspects of the response that the interviewer considers interesting and relevant will be reported (Warwick and Lininger 1975).

There is nevertheless a general agreement among all practitioners—both those for and against open-ended questions—that the open format of a question generates a different distribution of results from a closed version (table 3.9), although there seems to be no evidence that one form is preferable to the other (Foddy 1993).

Table 3.9

Open- and Closed-Question Formats Generate Different Responses

Most important thing for children to learn to prepare for life (percentage of respondents).

Answer	Closed Format	Open Format
1 To obey	19.0	2.4
2 To be well liked or popular	0.2	0.0
3 To think for themselves	61.5	4.6
4 To work hard	4.8	1.3
5 To help others in need	12.6	0.9
6 To be self-reliant		6.1
7 To be responsible		5.2
8 To have self-respect		4.1
9 To have respect for others		6.7
10 To have self-discipline		3.5
11 To be honest		7.4
12 To have other moral qualities		3.0
13 To be religious		5.4
14 To love others		2.0
15 To get an education		12.8
16 To learn a trade of job skill		0.9
17 To get along with others		5.0
18 Multiple answers not classifiable		16.1
19 Other	0.0	9.3
20 DK	0.0	1.3
21 NA	1.8	2.0

Source: Schuman and Presser 1979.
Note: DK = don’t know; NA = not available.

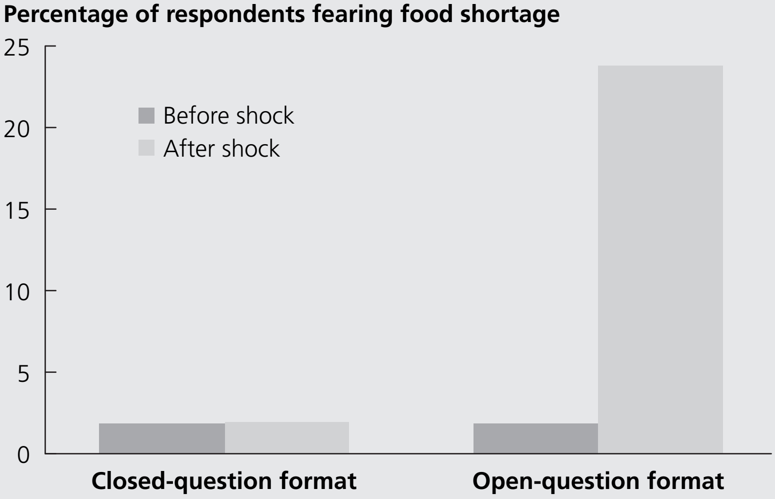
In an open-ended format, the respondent puts forward a set of information (the answer to the question) and an "expert" filters out the relevant answer on the basis of some detailed instructions (stimuli). In the closed format, the expert provides the "stimuli" and the respondent filters the information to extract the relevant answer. The decision to adopt a closed or open question is fundamentally a decision about who will interpret the information and extract the relevant answer. Given the same set of "stimuli" and the same set of information, using either of the two formats leads to the same relevant answer. Should the two formats lead to different answers the difference is attributable only to the agent who performs the interpretation of the information set. It is reasonable to assume that for a given set of stimuli the best agent to interpret the set of information provided by the respondent is the respondent himself or herself. This consideration, along with the benefits highlighted earlier and the realization that respondents might voluntarily or involuntarily not reveal all relevant information for the expert to extract the right answer, provides justification for a well-designed closed format over the open format.

Because it is not always possible or feasible to construct a well-designed closed question, narrative questions retain an important role in survey research. When information about a potentially complicated phenomenon is sought, when the range of possible alternatives is so extensive that it is practically impossible to list them,[31] or when the possible answers cannot be reduced to few words, the use of open-ended questions is recommended. If knowledge is being measured, then a narrative question is better than a multiple choice question, because in the latter case some correct answers may occur by chance. Similarly, when the reasoning behind a behavior or preference is of interest, the best way too learn it is through the respondent's own words (Fowler 1995). In other cases, when there is reason to suspect that external events might affect the respondent's answers, open questions should be used. So when asked about the most important problem facing the country following the 1977 winter storm, which generated a worry about food shortages, the closed question was unable to detect this shift in public opinion (figure 3.14) (Schuman and Presser 1979).

Finally, it is good practice to use narrative questions during the pilot test to ascertain how the respondent reacts to the question, to identify the optimal set of response categories, and to determine whether the

[31] As mentioned earlier, good closed question design requires that the list of alternatives be exhaustive and mutually exclusive (Peterson 2000).

Figure 3.14.
Event Contamination: Closed-Question Format Is Unable to Detect External Shocks



Source: Schuman and Presser 1979.

closed question is itself appropriate. Notwithstanding their limitation and apart form the above-mentioned limited circumstances, research has shown that providing people with a list of answers to chose from gives more reliable responses than using the open-ended format (Fowler 1995; Schuman and Presser 1979).

Question Sequence

Question sequence should not only facilitate the administration of the interview but also "arouse the respondent's interest, overcome his suspicions and doubts, [. . .] facilitate recall, and motivate [him] to collaborate" (Warwick and Lininger 1975, 148). Therefore, questions should flow in an orderly sequence, with exact instructions on how to move ahead without having to look back and forth throughout the form (Warwick and Lininger 1975).

Three main aspects of the question sequence need attention: opening questions, flow of questions, and location of sensitive questions (Warwick and Lininger 1975).

Opening Questions

At the start of an interview, respondents are a bit suspicious about the study and unsure about their role as informants. Thus, the first questions should be easy, pleasant, and interesting. It should allow them to build up confidence in the survey's objective, stimulate their interest and participation, and eliminate any doubt that they may have about being able to answer questions (so-called motivation and confidence building). Good opening questions are conversational and encourage the respondent to express himself or herself in positive terms, while remaining within the purpose of the study (Warwick and Lininger 1975). Questions such as "What do you do in your free time?" or "Which school do your kids attend?" are irrelevant to the purpose of a business survey, might immediately generate suspicion, and should be avoided. A much better opening is "Can you tell me a little bit about your company?" In any question originating from the interviewer, even the most conversational one, the respondent must be able to see the relationship between the question asked and the purpose of the study. This is the only way to build and keep trust.

Question Flow

The sequence of questions in the body of the form should be tuned to a good flow of ideas and to the logical reasoning of the respondent. Once a general topic has been addressed, all related questions should come up before a second topic is raised. Similarly, if a long list of events is asked, each with dependent questions, it might be confusing for the respondents to go back and forth to each event to provide additional details each time.[32]

It is a good practice, especially in business surveys, not only to start with a narrative question but also to add open questions at regular intervals throughout the form. Elites "resent being encased in the straightjacket of standardized questions" (Zuckerman 1972, 167) and feel particularly frustrated if they perceive that the response alternatives do not accurately address their key concern (Dexter 1970).

[32] In this case, however, the designer must keep in mind that the respondent might understand the structure of the question itself and modify his or her answer to avoid the dependent questions to speed up the interview. For this reason, it is suggested to ask the main questions in their entirety first and then proceed with the dependent questions (Atkinson 1971; Moser and Kalton 1971; Plateck, Pierre-Pierre, Stevens 1985).

Illogical jumps or an abrupt change of topic should also be avoided because it will create confusion, and possibly frustrate the respondent and compromise the accuracy of the data. It is good practice to divide the survey subject in different topics linked in the interview by transitional explanations, such as "Okay, let's now move to . . ." or "All right, I am now going to ask you a few questions on. . . ." These transitions play a critical role in the sequencing of questions in both introducing a new topic and showing how a new topic relates to the purpose of the study. More important, transitional phrases help respondents foresee what type of questions they are going to be asked. This will focus their minds and help them relax. Furthermore, transitional explanations exert a positive psychological effect on respondent by giving a sense that we are moving toward the end of the interview (Atkinson 1971). Finally, the use of bridging remarks, such as "You mentioned earlier that . . . ," should be encouraged because it shows that the interviewer is attentive and interested in what the respondent has to say.

The order of questions can also be used to aid individual's memory or to gradually introduce respondents to unpleasant or embarrassing topics.[33] For example, questions on awareness of a program should precede questions on their use. Questions with a common reference period should be grouped together, with the most recent period coming first (Warwick and Lininger 1975). Easier questions should be asked at the beginning or the end of the interview.

Filter questions and conditional questions should be used to guide the interviewer and exclude respondents from a question sequence that does not apply to them. It is time-consuming and annoying for a respondent to be asked "How many days did it take you to export your goods?" if he or she does not export at all. Filter questions are more efficient, and therefore should be preferred over conditional questions. They allow us to discriminate between "not applicable" and "nonresponse." As a matter of fact a "not applicable" response to the question "If you have a loan, what is the interest rate?" could mean that the respondent does not have a loan or that he or she simply does not pay interest on an existing loan.

[33] To some extent, the actual order of questions might affect respondents. Although the literature is split on this issue (Benton and Daly 1991; McAllister and Wattenberg 1995; McClendon and O'Brien 1988; Sigelman 1981), it seems that "question order effects occur only when a prior question establishes a response set to a subsequent question; that is, that the order effects are selective or conditional on the substance of specific questions" (Crespi and Morris 1984, 580).

The importance of filter questions cannot be stressed enough, especially for attitude or knowledge questions. Respondents are reluctant to admit ignorance. They are open to offering opinions not only on subjects they know little about, but also on fictitious information presented as fact.

> Gallup [. . .] finds that while 96 percent [of respondents] had an opinion on the importance of a balanced budget, 25 percent did not know whether the budget was currently balanced, 8 percent wrongly thought that it was balanced, 40 percent knew it was unbalanced but didn't know by how much, 25 percent knew it was unbalanced but overestimated or underestimated the amount by 15 percent or more, and 3 percent knew it was unbalanced and knew the approximate level. (Smith 1984, 221)

When asked about a fictitious "Public Affairs Act" one-third of respondents volunteered an answer in a form without a filter (Bishop, Oldendick, Tuchfarber, and Bennett 1980; Bishop, Tuchfarber, and Oldendick 1986) Filters are extremely important because they can screen out from 5 to 45 percent of responses, depending on the wording of the filter and how familiar or emotive the issue covered is (Bishop, Oldendick, and Tuchfarber 1983).[34]

Location of Sensitive Questions

While all agree that sensitive questions should not come at the beginning of the interview, when the main goal of the interviewer is to gain trust from the respondent, the belief that these questions should not be placed at the end is not unanimous either. Some practitioners in fact favor the "hit-and-run" method, in which as many sensitive questions as possible are asked toward the end of the interview until the respondent becomes unwilling to continue.[35] This method shows a poor understanding of the psychology of the interview process and leaves the respondent with negative attitudes toward the study (Warwick and Lininger 1975). In the worst case, the respondent may reject the whole interview and ask the interviewer to hand over the questionnaire.

Sensitive questions should be introduced only at a point of the interview at which the respondent is likely to have developed confidence in

[34] Examples of filters include the following: "Do you have an opinion on this or not?" "Have you been interested enough in this to favor one side over the other?" "Have you thought much about this?" and "Where do you stand on this issue, or haven't you thought much about it?" (Bishop, Oldendick, and Tuchfarber 1983).

[35] See Moser and Kalton (1971, 346).

the purpose of the study and trust in the interviewer. They should be placed where they are least likely to be sensitive, such as where the topic being discussed is the most appropriate. Finally, to mitigate the perceived threatening nature of sensitive questions it is good practice to introduce them gradually by a series of warm-up questions (Warwick and Lininger 1975).

Questionnaire Length

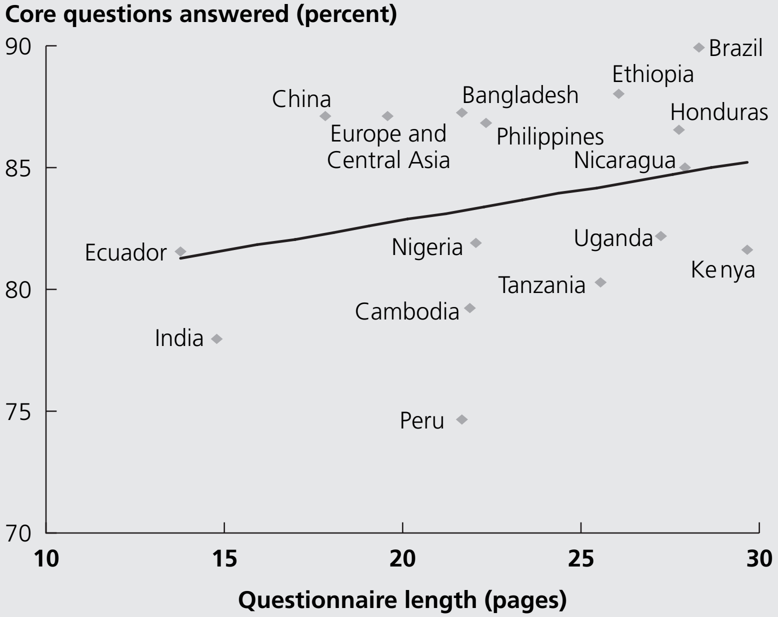
While the literature investigates extensively the relationship between question length and data accuracy, few authors have analyzed the effect of questionnaire length on data quality. A review of the literature by Bogen (1996) finds no clear association between questionnaire length and survey participation. While she blames a lack of experimental research on this issue, she points out that the existing evidence does not lead to the assertion of a negative relationship between questionnaire length and response rate. Most of the papers reviewed in this article refer to mail and telephone interviews and, in the few papers on face-to-face interviews, the average interview length is one hour. For longer interviews, like Investment Climate Surveys interviews, which can run between one and two hours, very little experimental research can be found.

As discussed in chapter five, the existing literature suggests that survey participation is only marginally associated with questionnaire length. Evidence form the Investment Climate Surveys confirms that questionnaire length[36] has no impact on response rate (figure 3.15).[37]

[36] The number of pages is only an imperfect measure of questionnaire length. The real length is a function of three components: the length of individual questions, the number of questions, and the format of questions included in the form. A questionnaire of only 10 questions, but with each of them in a table format, is more difficult and time-consuming than a questionnaire with the same amount of information asked in a sequence of questions with appropriate skip patterns. To account for this, we "standardized" the definition of questionnaire length by counting the words of each questionnaire and calculating the number of pages, assuming 422 words per page. All figures and analyses in this book assume this standardized definition.

[37] The response rate is calculated as the percentage of "core" questions answered to control for question effect. A copy of the core questionnaire used in the investment climate surveys is available at: http://www.ifc.org/ifcext/economics.nsf/Content/IC-SurveyMethodology. Countries included in the figure are: Bangladesh2002, Brazil2003, Cambodia2003, China2002, Ecuador, Ethiopia2002, Honduras2003, India2000, Kenya2003, Kyrgyzstan 2003, Moldova2003, Nicaragua2003, Nigeria2001, Peru2002, Philippines2002, Poland2003, Tajikistan2003, Tanzania2003, Uganda2003, Uzbekistan2003. The data on Investment Climate Surveys used throughout this book are available online at the following URL: http://iresearch.worldbank.org/ics/jsp/index.jsp.

Figure 3.15.
Relationship Between Questionnaire Length and Response Rate



Source: World Bank Investment Climate Survey 2003.

Questionnaire length has a significant impact on data accuracy. Longer questionnaires put an unfair burden on the time and memory of the respondent and will inevitably result in higher response errors. Although respondents may not appear to refuse to answer a question, they may provide any answer to complete the interview more quickly, hence providing biased information. The problem is that we cannot control for this bias.

Sudnam and Bradburn (1974) suspected that interviews over two hours might endanger response accuracy. They affirm that fatigue does not jeopardize data quality in interviews up to one-and-a-half hours long, while pointing out that "a fatigue factor [. . .] could become a serious problem in interviews lasting more that two hours" (Sudnam and Bradburn 1974, 90). Andrews (1984) also found a similar association by demonstrating that on a questionnaire with up to 348 items "better data quality comes from items that fell in the 26th to 100th positions" (432).

Notwithstanding this evidence it is hard to determine the optimal length of a questionnaire. In face-to-face interviews there is a general agreement that the interview should take no longer than 45 to 60 minutes.[38] After 75 minutes, the managers manifest clear signs of fatigue (uncomfortable on the chair, watching the clock, looking around the room, asking more often for the question to be repeated). Data from the Investment Climate Surveys also seems to confirm that data accuracy starts to suffer when the questionnaire reaches 20 pages. Allowing for the fact that only part of the whole Investment Climate Surveys questionnaire is administered through a formal face-to-face interview,[39] evidence seems to show that in general face-to-face interviews should not exceed 14 pages (or approximately one-and-a-half hours).[40]

Questionnaire Layout

Often not enough attention is paid to the physical layout of the questionnaire, which results in a greater likelihood of errors by interviewers, editors, coders, key operators, and ultimately respondents.

> A common reason for poor layout is the desire to fit all of the questions in a single page, even if the type has been reduced to miniscule proportions and the items crammed together. [. . .] Though such forms mean savings on paper and printing expenses, they are ultimately wasteful if they reduce the quality of the information obtained. (Warwick and Lininger 1975, 151)

A number of principles should be followed to ensure that the questionnaire is convenient for the interviewer and respondent, as well as easy to identify, code, and store.

[38] See Rea and Parker 1997.

[39] The Investment Climate Surveys questionnaire is usually divided into two parts. The first is administered through a face-to-face interview with the chief executive officer or manager of the selected establishment. The second part is to be filled in by the accountant and human resources manager under the supervision of the interviewer, and requires referencing books and records. Although the questionnaire length is measured for the whole questionnaire, the time to complete the face-to-face interview refers only to the first part.

[40] In determining the actual length of the interview, the survey manager should keep in mind that a host of other factors beyond the questionnaire length play a role, not the least being the interviewer's ability.

Identification

Each form should contain one or more unique identifying numbers assigned in advance and marked on each questionnaire. It is good practice to have the same number assigned to each sample unit be the identifier of its paper questionnaire as well (Warwick and Lininger 1975).

Numbering

Questions should be numbered sequentially throughout the instrument without omissions or repetitions. Even if the questionnaire is divided in sections or parts it is preferable to use progressive numbers throughout the form.

Space

Sufficient space between questions should be left to facilitate questionnaire administration. Saving space will ultimately be uneconomical if it compromises data accuracy. It is advisable to print only on one side of the page and to leave enough space for notes from interviewers, editors, or coders on both sides of the questions. To facilitate data entering, it is good practice to align the answer boxes to the margin of the pages. If it is not practical to do so (that is, when questions are of varying length) then it might be appropriate to use a table format in which questions are spread across the page in an orderly fashion allowing the justification of answer categories (example 3.12). Finally, if it is not feasible to place the answer boxes next to the question it is advisable to use dotted lines to connect them (Plateck, Pierre-Pierre, and Stevens 1985).

Open-ended questions should have sufficient space for the expected average answer length. For questions that need to be coded at a later stage, "For Official Use Only" space should be clearly allowed in the questionnaire.

Instructions

Instructions are critical both for the administration of the form as well as for the collection of accurate data. Two types of information must be readily distinguishable in the questionnaire: questions to be read and instructions to be followed.

One effective way to eliminate confusion between instructions and questions is to use different formats or to put one of them in a box. So, for instance, all verbatim questions can be typed in regular font while instructions are typed in capital letters. Instructions should be placed

Example 3.12

Questionnaire Layout

Language

11. WHAT IS THE LANGUAGE . . . FIRST LEARNED IN CHILDHOOD AND STILL UNDERSTANDS? (Mark all that apply)
    English..○
    French ....○
    Other ....○

12. CAN . . . SPEAK ENGLISH OR FRENCH WELL ENOUGH TO CONDUCT A CONVERSATION?
    NO—Neither English nor French ...○
    YES—English only.........................○ Go to 14
    YES—French only..........................○
    YES—Both English and French.......○ Go to 13

13. IN GENERAL, WHICH OF THESE TWO LANGUAGES DOES . . . PREFER TO SPEAK?
    English ....○
    French...○
    Neither..○
    Don't know..................................○
    No preference..............................○

14. WHAT LANGUAGE DOES . . . SPEAK MOST OFTEN AT HOME? (Mark all that apply)
    English ....○
    French...○
    Other ....○

Education

15. HAS . . . (EVER) ATTENDED A UNIVERSITY, COMMUNITY COLLEGE, OR OTHER POSTSECONDARY INSTITUTION AS A FULL-TIME STUDENT?
    Yes ○ No ○ Go to 20

16. WHAT IS THE HIGHEST LEVEL OF EDUCATION . . . COMPLETED?
    □□ Enter code
    If code 99 Go to 20

17. IN WHAT YEAR WAS . . . 'S LAST DEGREE, DIPLOMA, OR CERTIFICATE GRANTED?
    1 9 □□ Year

18. IN WHICH PROVINCE, TERRITORY, OR OTHER COUNTRY WAS THIS DEGREE, DIPLOMA, OR CERTIFICATE GRANTED?
    □□ Enter code

19. WHAT WAS . . . 'S MAJOR FIELD OF STUDY?
    □□ Enter code

20. HAS . . . LIVED IN ANY OTHER PROVINCE, TERRITORY, OR OTHER COUNTRY SINCE JUNE 1, 1976?
    Yes ○ No ○ Go to 48

Example 3.12 (continued)

Migration History

21. IN WHICH PROVINCE, TERRITORY, OR OTHER COUNTRY DID . . . LIVE ON JUNE 1, 1976?
    □□ Enter code

22. TO WHICH PROVINCE, TERRITORY, OR OTHER COUNTRY DID . . . FIRST MOVE AFTER JUNE 1, 1976?
    □□ Enter code

23. WHEN DID . . . MAKE THIS MOVE?
    Mo. □□□□ Yr.

24. TO WHICH PROVINCE, TERRITORY, OR OTHER COUNTRY DID . . . MOVE NEXT?
    □□ Enter code
    If code 99 (No other moves) go to 34

25. WHEN DID . . . MAKE THIS MOVE?
    Mo. □□□□ Yr.

26. TO WHICH PROVINCE, TERRITORY, OR OTHER COUNTRY DID . . . MOVE NEXT?
    □□ Enter code
    If code 99 (No other moves) go to 34

27. WHEN DID . . . MAKE THIS MOVE?
    Mo. □□□□ Yr.

28. TO WHICH PROVINCE, TERRITORY, OR OTHER COUNTRY DID . . . MOVE NEXT?
    □□ Enter code
    If code 99 (No other moves) go to 34

29. WHEN DID . . . MAKE THIS MOVE?
    Mo. □□□□ Yr.

30. TO WHICH PROVINCE, TERRITORY, OR OTHER COUNTRY DID . . . MOVE NEXT?
    □□ Enter code
    If code 99 (No other moves) go to 34

31. WHEN DID . . . MAKE THIS MOVE?
    Mo. □□□□ Yr.

32. TO WHICH PROVINCE, TERRITORY, OR OTHER COUNTRY DID . . . MOVE NEXT?
    □□ Enter code
    If code 99 (No other moves) go to 34

33. WHEN DID . . . MAKE THIS MOVE?
    Mo. □□□□ Yr.

Source: Plateck, Pierre-Pierre, and Stevens 1985.

Example 3.13

Different Emphasis Implies Different Answers to the Same Question

Overt Emphasis on		Implied Emphasis on	Possible Replies
Why did you buy this book?	→	motivation	(gift, self interest)
Why did ***you*** buy this book?	→	person	(self, other)
Why did you ***buy*** this book?	→	action	(rent, borrow)
Why did you buy ***this*** book?	→	object	(other book, magazine)

Source: Peterson, 2000.

directly above the question concerned or the section of the questionnaire to which they apply. It is not advisable to put instructions at the beginning of the questionnaire or in the manuals (Plateck, Pierre-Pierre, and Stevens 1985).

Fonts and Formats

Given that the same question can educe different meanings if different words are emphasized (see example 3.13), critical words should be underlined or printed in bold to ensure uniform emphasis by interviewers and uniform interpretations by respondents.

Similarly, when the time reference changes from one question to the next, it is particularly important to ensure uniformity of interpretation. Using bold when asking "**In the last two weeks,** how many . . ." focuses the respondent's attention on the new time reference (Plateck, Pierre-Pierre, and Stevens 1985).

In a multicultural survey in which each question is printed first in one language and then in the second language, it is good practice to use two different fonts for each language throughout the questionnaire (Plateck, Pierre-Pierre, and Stevens 1985).[41]

Symbols

Symbols such as circles, arrows, boxes, triangles, and asterisks, as well as different colors or shades are excellent visual tools and should be used

[41] Alternatively, the questionnaire could be printed double-sided with each side having the same questions but in different languages. When more than two languages are used, it is advisable to use separate forms for each language.

not only to guide the interviewer and respondent throughout the form, but also to facilitate the work of supervisors and key operators. The questionnaire can be designed in such a way that each type of answer has an associated type of symbol. So, for instance, arrows might help identify the skipping pattern, a circle might be used whenever a check mark is called for, and a box whenever numbers are called for (see previous example 3.12) (Plateck, Pierre-Pierre, and Stevens 1985).[42]

Similarly, if the questionnaire is divided into two parts, it has been proven useful to have the two parts printed on two different paper colors.

Translation

Surveys are often conducted in multiethnic and multilinguistic societies. Thus, asking the same question in countries (or regions) with different cultures, traditions, beliefs, and languages becomes even harder and the solutions are more complex.

Three major concerns are raised by comparative research. The first is whether one concept has the same meaning in different cultures. So, for instance, the concept of illegal party contribution might have different meanings in the United States and the Philippines. Second, even if the same concept has the same meaning in different cultures, it does not imply that the same indicator of that concept would apply to all of the meanings. The notion of political activity is the same in the United States and Europe, but the pattern of activities is different for each country. The third concern refers to the analytical value of the information collected. A well-translated identical indicator can still generate a nonequivalent response pattern across cultures. This happens when different response styles occur in different cultures. So, for instance, if acquiescence is more common among small companies, then appropriate controls should be introduced to avoid the fact that apparent differences across countries actually reflect differences in the strata's composition (Warwick and Lininger 1975).

The implication is that translation in the local language should not be seen as simple "transliteration" of the words. Rather it should be a transformation of the instrument to "conceptual equivalence" (Hunt, Crane, and Wahlke 1964). In other words, it is essential that the trans-

[42] This helps not only in the data-collection phase but also in the data-checking stage of the survey.

lation convey a consistent message in different cultures. So if we want to measure "unlawful party contribution to political parties," we cannot use the same measurement for the United States and the Philippines, because in the latter no laws forbid such contributions. A technique to ensure good translation quality is the so-called back translation as outlined in the following four steps:

Step 1. The questionnaire is first translated from language A to language B.

Step 2. The translated version is then translated back from B into A by *a different* translator.

Step 3. The two versions of the questionnaire, original and back translated, are then compared and discrepancies clarified and corrected.

Step 4. The revised translated version is translated back into A again for comparison, and this process keeps going until there are no more inconsistencies between the two versions (Warwick and Lininger 1975).

During this translation process, it is useful to treat the pre-test as an additional tool for checking the translated version of the questionnaire. A lot of the problems of designing a cross-cultural study can be addressed

> if the investigators take to the field early in the study and allow ample time and resources for pre-testing the research instrument. The day when a single questionnaire is designed in the United States or Europe and sent to 'hired hands' in other countries is hopefully over. (Warwick and Lininger 1975, 167)

This brings us to the last step of questionnaire design: pre-testing.

Pre-Test

Armchair discussions cannot replace direct contact with the population being analyzed (Warwick and Lininger 1975).

> It is all too easy to think that one can draft a perfectly worded questionnaire while sitting in an office. In fact, it is very difficult to imagine all the possible interpretations and the variety of answers respondents may give, or the different circumstances or conditions which may alter the sense of the questions. (Plateck, Pierre-Pierre, and Stevens 1985, 21)

By the time the form reaches the pre-test stage all issues of wording, style, content, layout, and language should be resolved. The pilot represents the first "live" test of the instrument, as well as the last step in the finalization of the questions. No matter how experienced the questionnaire designer is, any attempt to shortcut this step will seriously jeopardize the accuracy of the data about to be collected. Time constraints should not come at the expense of this essential last step in the design of the questionnaire (Moser and Kalton 1971).

As discussed previously in this chapter, good question design requires clarity in the terminology adopted. "Survey questions [. . .] should mean the same thing to all respondents, and they should mean the same thing to respondent as well as to the researcher" (Fowler 1992, 218). The pilot test represents the only opportunity to verify this and the data collected will ultimately reflect any poorly defined question or concept. When asked "How many *servings* of eggs do you eat in a typical day?" some 33 percent of respondents interpreted one serving as one egg and 47 percent of them interpreted one serving as two eggs. Therefore, when the meaning of "servings" was clarified, the number of respondents reporting two eggs went from 15 to 62 percent (figure 3.16).

Bassili and Scott (1996) and Fowler (1992) show that clearer questions reduce both the time to answer as well as the requests for clarifications, thus reducing the time to complete the interview (figure 3.17).

Pre-test methods

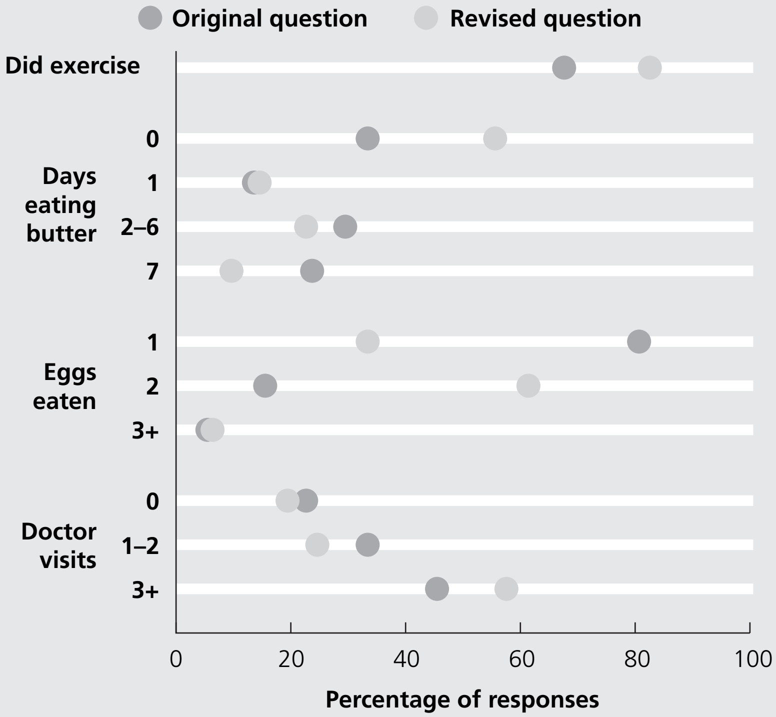
The pre-testing of a questionnaire can be conducted following three different methods: conventional, behavioral, and cognitive.

The conventional method involves a small number of interviews followed by a debriefing in which experiences are shared and problems identified.[43] The behavioral pre-test involves structured interviews monitored by an expert whose role is to identify and code problems. In the cognitive pilot, the respondent is asked to report everything that comes to his or her mind while or after answering the questions. Preliminary experimental results show that each method serves a different purpose. The behavioral and conventional methods are more appropriate for detecting problems with both the respondent and the interviewer, whereas the cognitive method assesses the analytical accuracy of the answers by evaluating the questions from the point of view of the effort required to answer. Conventional and cognitive pre-tests also perform well in identifying semantic problems.

[43] This methodology is efficient if the investigator designs effective debriefing questions for both the interviewer and the respondent and if he or she is able to determine which information is not relevant (Campanelli, Martin, and Rothgeb 1991).

Figure 3.16.

Effect of Unclear Terms on Survey Results

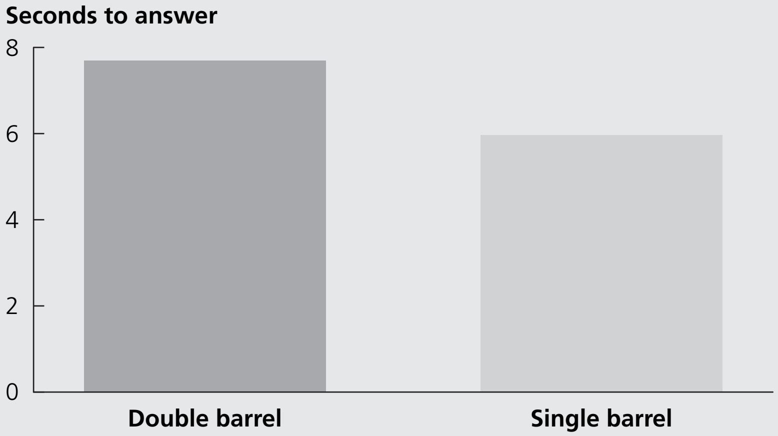


Original question[a]	Unclear term in original question	Clarification in revised question
Have you done any exercise?	Exercise	Walking included
Days per week you have butter?	Butter	Margarine excluded
How many servings of eggs?	Serving	Number of eggs
Times you have visited a doctor?	Visit	See a doctor, or see a nurse, or see a medical assistant, or receiving medical advice, prescription or test results

Source: Fowler 1992.

a. Meaning of question. For exact question wording see reference.

Figure 3.17.
Unclear Terms Take Longer to Answer



Source: Bassili and Scott 1996.
Note: For single barrel, the answer is the average of two simple questions (author's calculations).

In addition to these three field methods, there is a fourth method in which expert designers review the questionnaire in the office. This has proven beneficial, particularly for the identification of problems related to the analytical value of the questions (figure 3.18) (Presser and Blair 1994). Biemer and Lyberg (2003) present a useful list of coding categories that experts can use in their assessment of the questionnaire (box 3.1).

Pre-test goals

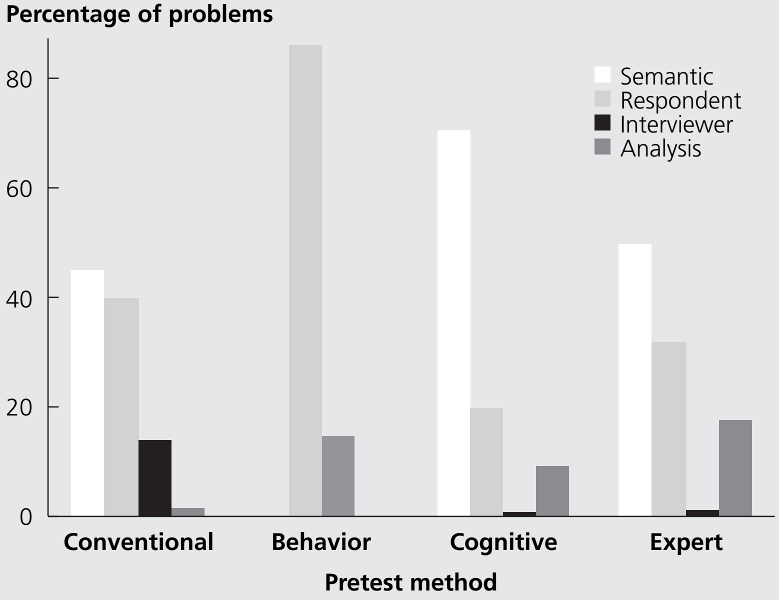
The purpose of the pre-test is threefold:

- To evaluate the adequacy of the questionnaire,[44]
- To estimate the length of the interview, and
- To determine the quality of the interviewers.

It is not easy to determine a priori the warning signs of design defects. Many different situations can occur during an interview leading to a great variety of answers to any single question. In general, too many answers at one extreme may indicate a leading question. Too many "don't know" or requests for clarifications definitely indicate vague questions, questions using uncommon words, or questions going out-

[44] Occasionally, the pilot can also be used to evaluate different wording of the same question. In this case, however, the survey manager must ensure that the alternative versions of the question are tested by the same interviewer on two equivalent random sample of respondents (Moser and Kalton 1971).

Figure 3.18.
Percentage of Problems Identified by Different Pre-Test Methods



Source: Presser and Blair 1994.
Note: Weighted averages of all trials, author's calculations.

side the respondent's experience. If respondents add qualifications to their answer, the question needs to be clarified. If many refuse to answer or answer in the same way, the question must be reworded, repositioned, or cut out altogether (Moser and Kalton 1971).

Following is a checklist of concerns regarding the questionnaire that the designer should address during the pilot:[45]

- Do respondents understand what the survey is all about?
- Do they feel comfortable answering questions?
- Is the wording clear?
- Is the time reference clear to respondents?
- Are the response categories compatible with the respondent's experience?

[45] A systematic approach to rate questions during the pilot has been suggested by Fowler (1995, 116–124).

Box 3.1

List of Questionnaire Problems for Pre-Test Expert Review

1. PROBLEMS WITH READING: Determine whether it is difficult for the interviewers to read the question uniformly to all respondents.
   1a – WHAT TO READ: Interviewers may have difficulty determining what parts of the question are to be read.
   1b – MISSING INFORMATION: Information the interviewer needs to administer the question is not contained in the question.
   1c – HOW TO READ: Question is not fully scripted and therefore difficult to read.

2. PROBLEMS WITH INSTRUCTIONS: Look for problems with any introductions, instructions, or explanations from the respondent's point of view.
   2a – CONFLICTING OR INACCURATE INSTRUCTIONS, introductions, or explanations.
   2b – COMPLICATED INSTRUCTIONS, introductions, or explanations.

3. PROBLEMS WITH ITEM CLARITY: Identify problems related to communicating the intent or meaning of the question to the respondent.
   3a – WORDING: The question is lengthy, awkward, ungrammatical, or contains complicated syntax.
   3b – TECHNICAL TERMS are undefined, unclear, or complex.
   3c – VAGUE: The question is vague because there are multiple ways in which to interpret it or to determine what is to be included and excluded.
   3d – REFERENCE PERIODS are missing, not well specified, or are in conflict.

4. PROBLEMS WITH ASSUMPTIONS: Determine whether there are problems with assumptions made or the underlying logic.
   4a – INAPPROPRIATE ASSUMPTIONS are made about the respondent or his/her living situation.
   4b – ASSUMES CONSTANT behavior: The question inappropriately assumes a constant pattern of behavior or experience for situations that in fact vary.
   4c – DOUBLE-BARRELED question that contains multiple implicit questions.

5. PROBLEMS WITH KNOWLEDGE/MEMORY: Check whether respondents are likely to not know or have trouble remembering information.
   5a – KNOWLEDGE: The respondent is unlikely to know the answer.
   5b – An ATTITUDE that is asked about may not exist.
   5c – RECALL failure.
   5d – COMPUTATION or calculation problem.

Box 3.1 (continued)

6. PROBLEMS WITH SENSITIVITY/BIAS: Assess questions for sensitive nature or wording, and for bias.
   6a – SENSITIVE CONTENT: The question is on a topic that people will generally be uncomfortable talking about.
   6b – A SOCIALLY ACCEPTABLE response is implied.

7. PROBLEMS WITH RESPONSE CATEGORIES: Assess the adequacy of the range of responses to be recorded.
   7a – OPEN-ENDED QUESTIONS that are inappropriate or difficult.
   7b – MISMATCH between question and answer categories.
   7c – TECHNICAL TERMS are undefined, unclear, or complex
   7d – VAGUE response categories.
   7e – OVERLAPPING response categories.
   7f – MISSING response categories.
   7g – ILLOGICAL ORDER of response categories.

Source: Biemer and Lyberg 2003.

- Which items require respondents to think hard before they answer?
- What cognitive processes do they adopt to answer difficult questions?
- Which items seem to produce irritation, embarrassment, or confusion?
- Are there any items that respondents consider comical?
- Does the style of the question generate bias?
- Are the answers we get what we really want for the purpose of the study?
- Is there enough variability in the answers received?
- Are there local expressions that should be incorporated into the items to avoid ambiguity?
- Is the questionnaire too long?
- In the eye of the respondent, have any other important issues been overlooked in the questionnaire?

Many of those issues are hard to judge so it is important for experienced staff with a profound understanding of the analytical purpose of each question to participate in the pre-test. Take for instance the following question:

Example 3.14

Pre-Testing Helps Determine Whether Question Goals Are Met

What share of your plant machinery and equipment is:

a. <5 years old _____%
b. 5–10 years old _____%
c. 10–20 years old _____%
d. >20 years old _____%

The goal of this question was to get an estimate of capital vintage in the establishment and relate it to both technological progress and productivity improvement. For this question to be of analytical value, the answer should be based on the market value of the machinery, not on its book value. Unfortunately a cognitive evaluation of this question during a pre-test has shown that the answers provided were based on the physical number of the machinery. Even when the question was reworded to clearly indicate that the market value of machinery should have been referenced, the respondents had so much difficulty answering it that they would purely guess an answer.

The pilot is also an excellent opportunity to test both the convenience of the form and the ability of interviewers.[46] Some of the questions that the survey manger should consider addressing during the pilot to assess the convenience of the instrument are as follows:

- Is the questionnaire easy to administer?
- Do filters and skip patterns work properly?
- Are instructions clear?[47]
- Are transitions from question to question smooth?

Similarly, some of the issues relevant to assess the interviewer's ability are as follows:

- How does he or she read the questions?
- How does he or she behave in difficult situations?

[46] Although it is not feasible to have all potential interviewers participate in the pre-test, the survey manager's experience with the pre-test can serve as a first step in establishing an expectation on the average level of interviewers. This will help him or her better tailor the training.

[47] A useful technique to test the usefulness of instructions, paraphrasing, is suggested by Gower (1993). "Respondents are asked to repeat the question in their own words, or to explain the meaning of terms and concepts [. . .] used."

- How does he or she handle questionnaire instructions?
- How does he or she explain concepts not clear to the respondent?
- How does he or she probe and record answers?

The size of the pre-test is more a matter of convenience and availability than the result of a random selection process. Generally it should be carried out in 15 to 25 establishments. Although firms of different sizes should be included in the test, the selection of respondents is more purposive than random. It is not necessary to visit establishments in different locations or industries unless there are reasons to believe that regional or sectoral differences in the interpretation of questions might exist.

Because it is difficult for one individual to carry out a good interview and observe and take notes on how to improve the instrument, it is good practice for the survey manager or an expert to accompany the interviewer during the pre-test. The expert will then observe the interviewer's behavior and the respondent's reactions, and will perform a cognitive assessment of the most challenging questions. To better perform these tasks, it is advisable to focus only on a subset of the questions on each pre-test and to apply what has been learned in previous pilots to subsequent interviews. With all the information gathered during the pre-test the questionnaire is then finalized.[48]

[48] At this stage, the form is almost ready to be administered. A few useful comments and modifications could come up during the training, especially if the instrument is translated into the local language.