6

Pretesting and Pilot Testing

INTRODUCTION

In this chapter, we detail the possibilities and pitfalls presented by **pretesting**, the methods of validating the survey instrument and its measurements, and **pilot testing**, the "dress rehearsal" of survey administration and procedures (Rothgeb 2008, 584). Pretesting and pilot testing are invaluable components of survey research, affording researchers a valuable opportunity for reflection and revision of their project before the costs of errors begin to multiply later on. We begin this chapter with a discussion of the goals of and guidelines for pretesting followed by a summary checklist to help you make the most of this procedure. Then we provide an elaboration of the broader process of pilot testing the entire project from start to finish. If you pay close attention to the issues outlined in this chapter, pretesting and pilot testing could lead you to that early "stitch in time" that saves countless dollars and hours later on down the road.

PRETESTING

Once you have a complete draft of your survey, a pretest is a necessary next component of the research process. Pretesting your survey is an important way to pinpoint problem areas, reduce measurement error, reduce respondent burden, determine whether or not respondents are interpreting questions correctly, and ensure that the order of questions is not influencing the way a respondent answers. In other words, a pretest is a critical examination of your survey instrument that will help determine if your survey will function properly as a valid and reliable social science research tool (Converse and Presser 1986).

Using a pretest of the survey, researchers are able to ensure that the questions are clearly articulated and that the response options are relevant, comprehensive, and mutually exclusive—and not just in their own estimation, but from the point of view of the respondents as well. Making sure that researchers and respondents interpret the survey in the same way is of the very highest concern in survey design, and pretesting is one of the best ways to do this (Converse and Presser 1986). Pretesting can bring to light those inevitable instances of obscure terminology, unfamiliar references, and ambiguous words and phrases that the developer did not initially see as problematic, but that could confound and frustrate the respondent and hurt data quality and response rates. Furthermore, the pretest also allows the researcher to assess **response latency**, the amount of time it takes to complete individual items in the survey as well as the full survey, which can then be reported in the introduction of the full-scale survey (for reasons discussed in Chapter 3; Bassili and Scott 1996; Draisma and Dijkstra 2004).

Another important feature of pretesting a survey is the technical report (paper trail) left for future research endeavors. When a survey researcher has conducted similar research in the past, and has pretested and sufficiently documented survey materials, these tried and true measures help design a valid survey instrument. Thus, a meticulous record of the pretest process helps avoid future problems encountered at the various stages of study design.

When considering research funding sources, pretesting your survey instrument prior to full-scale administration lends credibility to your proposed work and accountability to you as a researcher, which could also potentially increase the probability of obtaining research funding.

Pretesting also serves as practice administration and a way to evaluate respondents' understanding of the concepts under study as well as the quality of their interviews (Converse and Presser 1986). All respondents should understand the concepts and ideas *in the same exact way*. In the following example, *housework* may be interpreted differently by men and women and by individuals with and without children. Some might argue that childcare is included in housework, while others might argue that it is not:

1. Considering your work and commute schedules, would you say the division of your housework is fair?

    1) No, I do way too much of the work.

    2) No, I do somewhat more than I should.

    3) No, my partner does way too much of the work.

    4) No, my partner does somewhat more than she/he should.

    5) Yes, it is fair enough.

Choosing not to pretest a questionnaire poses a potentially serious threat to the accuracy of the survey questions and resulting data. It is better to pretest a questionnaire on even just one person rather than field the survey without pretesting. The rule of thumb is to test the survey on at least 12 to 50 people prior to pilot testing or full-scale administration (Sheatsley 1983; Sudman 1983). This is a cost-, energy-, and time-efficient number of people—a large enough number that many will note the same problems with the survey questions. Inclusion of more than 50 "test respondents" *may* lead to the identification of more problems, but there comes a point of diminishing return, as the financial and time costs of further pretesting outweigh the benefits of discovering more relatively small issues and inconsistencies in the instrument.

The time involved in pretesting a survey (and the posttest assessment) depends largely on the length of the survey. The following section discusses who should participate in a survey pretest (i.e., the sample of testing participants), how the pretest should be carried out, how to collect pretest data, and what to do with feedback from the survey pretest.

Expert-Driven Pretests

Researchers sometimes call upon experts in a given field to identify problems with questions or response options in a survey (Presser and Blair 1994). For instance, a child behavior specialist may help pinpoint measurement issues in a newly developed child behavior checklist. As noted in Chapter 5, expert-driven pretests are crucial when assessing the face validity and construct validity of a measurement.

You might ask experts to pretest your survey items by going through the entire survey themselves, and, rather than asking them to provide an individual assessment of each item, ask them to rate the items on a Likert scale such as that outlined below. The idea is not to collect the experts' opinions and beliefs but to get their judgment of how well each questionnaire item truly reflects the construct you intend it to measure (Jansen and Hak 2005). Soliciting such expert appraisals of each and every survey question, using Likert-type scale items such as those below, can be an extremely valuable strategy for identifying problems and fine tuning items to collect optimal measurements:

1) Very strongly represents the construct
2) Somewhat strongly represents the construct
3) Unsure
4) Somewhat weakly represents the construct
5) Very weakly represents the construct

Experts are important not only for cross-checking the substantive aspects of the survey but for improving the overall style of the instrument as well. With their finer knowledge of the breadth of the given field, experts can tell you if all questions in the survey are relevant and necessary, or if some may be cut to shorten the questionnaire length and reduce respondent burden. They can also help decide if the survey flows seamlessly from one question to the next, thus following a logical and intuitive layout that again reduces respondent burden and improves the quality of your data (Olson 2010).

Finally, whether you have access to experts in the relevant field(s) or not, you can familiarize yourself with some of the extant expert research in the field, as well as with previous surveys on the same topic, to compare your newly designed measures to those in the established literature.

Respondent-Driven Pretests

Administration of the pretest survey to friends and colleagues is encouraged. However, the most useful pretesting is often done on a small subsample of the sample population, so that your pretesters fit the cultural and demographic profile of the larger sample to be surveyed later (Ferketich, Phillips, and Verran 1993). At the same time, you want your pretest group to encompass some variation within the broader profile, to ensure enough variety to notice any potential issues across the entire range of your questionnaire.

For example, if you plan to survey fundamentalist Christians, make sure to test individuals within different subgroups of that population. Fundamentalist Christians belong to different social classes and have varying ages and levels of education. If the terminology used in the questionnaire is widely understood only to younger populations, this oversight will be salient in the pretest of older individuals. Thus, a researcher will know to edit this terminology in the full-scale survey. If possible, pretest the study on multiple people within the various important subgroups of your sample too, so that their views can be confirmed by others in their subgroup.

Finally, when you administer the pretest, include an additional introduction to the questionnaire that once again thanks your participants and highlights the special importance of the pretesting process. For example, an introduction might read as follows:

> *In an effort to collect high-quality data on residential mobility patterns in the United States, we are developing a questionnaire to assess rates of and experiences with residential relocation. We greatly appreciate your willingness to participate in a preliminary assessment of this survey. After you have finished filling out the questionnaire, we will ask you to provide feedback on your understanding of the individual items in the survey. We would like to thank you in advance for this assistance.*

Collecting Pretest Data

When collecting the pretest data, it is important to use the same administration technique that will be used in the full-scale survey. If the full survey is to be conducted via phone interviews, for example, then you would not want to employ face-to-face interviews for the pretest—this would change the entire dynamic of data collection and cause you to overlook salient issues, and it could introduce extraneous issues that would not come up in the full survey. In general, when analyzing pretest results, you should pay particular attention to those more complex questions and items that were difficult for you to develop (such as, perhaps, multidimensional questions), as these items by definition contain more "moving parts" that could introduce problems. In addition to this, there are four more-specific strategies to follow to conduct a valid and reliable pretest assessment of your survey: behavior coding, cognitive interviews, individual debriefing, and group debriefing. We turn to each of these strategies in the sections below.

Behavior Coding

In **behavior coding** pretest assessment, researchers themselves administer the survey and ask respondents to take it in their presence (DeMaio, Rothgeb, and Hess 1998). Researchers watch as the respondent progresses through the questionnaire, noting behaviors of the respondent that may indicate problems with the survey, such as hesitation, confusion, and frustration. When the respondent has finished taking the survey, researchers also note which items were skipped, if and where any responses were erased or crossed out, and if there are any mistakes or other physical traces of confusion or miscommunication on the instrument itself. For example, on the item in Table 6.1, answers appear to have been provided backwards (reversing the poles of the scale) and then corrected. This could indicate a problematic scale, whose logic or specific points are not as clear as they could be:

Table 6.1 Religiosity Scale

	Agree		→		*Disagree*
Religion is important in my life…	1	⊗2	3	**(4)**	5
I regularly attend religious services…	1	2	**3**	4	5
I feel connected to a higher power…	⊗1	2	3	4	**(5)**

The respondent in the example above may have initially thought she was rating the "correctness" of the statements, rewarding more or fewer points for more or less correct statements. Should such an intuition prove common among pretesters, the researcher may want to reverse the scale as it appears on the instrument, or, at the least, add or clarify item instructions.

Cognitive Interview

In a **cognitive interview**, the researcher encourages pretest respondents to think out loud and voice their ongoing mental reactions, essentially narrating their thought processes while they take the survey (DeMaio, Rothgeb, and Hess 1998). For instance, thinking aloud, a respondent may respond to a question about how many times he has relocated in the last year by stating "Well, we moved from our primary house to our summer home in Sarasota at the end of June, and then back to our primary house in September. I guess that means we moved twice if we count the times we moved to our summer home. It would be zero if that doesn't count." This is a valuable revelation for the attentive researcher. The confusion highlighted in this answer is an indication that the question is not clearly worded, and should be changed or made more specific to more clearly measure residential relocation, so that it can be understood by everyone the same way.

Individual Debriefing

In an **individual debriefing assessment**, researchers debrief respondents after they have completed the survey, explicitly to gather feedback and reactions to specific questions (not just those eliciting respondent comments, as in the example above), the survey design, and the survey process (DeMaio, Rothgeb, and Hess 1998). Each question is reviewed and discussed individually with the respondents, and particular attention is paid to elements such as these:

- Question wording and language
- Comprehensive measurement
- Mutually exclusive measurement
- Additional comments

Once respondents have completed the survey, the researcher reviews each survey question with each respondent individually, asking them to remark about what they *believe* they are being asked and if they found anything confusing or misleading about the survey question. For example, with respect to the question, "How many times have you relocated in the past 12 months?" a respondent might respond, "I said twice because my family moves to Florida in the summer and then back to New York in the fall. I wasn't

sure if that counts since we still maintain the same primary residence." Again, this might prompt researchers to reconsider their question, perhaps even with help from the respondent, to accurately measure the concept intended. This might lead to a clearer question: "In the last 12 months, how many times have you and your immediate family undergone a permanent change of residence?"

Such individual-item debriefing may evoke a greater range of feedback or detail than a cognitive interview. However, there is also a danger in overusing this technique, as it may encourage respondents to retrospectively over-think items that were actually properly answered and unproblematic during the actual completion of the survey.

Group Debriefing

In **group debriefing assessment**, researchers bring test respondents together after the survey for a focus group discussion of the instrument (Vogt, King, and King 2004). They can then read the individual questions aloud and assess the test participants' reactions. Often, group debriefing will help researchers assess the magnitude of confusion for certain items, so that they do not over- or underreact to issues raised by any one individual respondent. For example, even though only one pretester might mention a specific item or scale in an individual interview, sharing this issue in a group debriefing might remind others that they too found it problematic. It could also encourage them to be more confident and forthcoming about an issue they were not sure they should raise or could not clearly articulate themselves. And even if one respondent's issue is not echoed by the group, this process confirms it as mere idiosyncrasy, making it easier for researchers to simply note and then move on.

Important Issues to Identify in Pretests

When analyzing pretester responses, researchers are likely to find issues converging on some common themes. Be attuned to the possibility of each of the general problems below, and pay special attention should one or two prove much more common than the others, as these may lead you to rethink the organization or design of the instrument.

Unclear Directions

Are all directions clearly articulated? Are ambiguities centered on a single item that needs reworking, a counterintuitive scale, a larger group of questions, or on all of the instructions in general? In the following example, there are no directions for respondents to mark all that apply. Respondents might therefore mark only the most important option:

2. Which of the following causes you a lot of stress?
   - ☐ My friends
   - ☐ My partner
   - ☐ My job/finances
   - ☐ My family
   - ☐ My children

Skipped Items

Are there multiple items in the survey (not part of a skip pattern) that were missed or avoided? Look for patterns among skipped items—do they relate to similar, difficult content? Personal information? Complex instructions? Or are they perhaps clustered toward the beginning, middle, or the end of a survey, indicating patterns in respondents' attentiveness at various points?

Refusal or Inability to Answer

"I don't know" and "N/A" are response options that should ideally be selected for only a very few items. These options often indicate an inability or refusal to answer a question. Thus, an inordinate number of "I don't know" or "N/A" responses implies there may be a problem with a question. Consider the example below:

3. How do you feel about the systemic reform of immigration policies that will assist lawmakers with adequately addressing delays in visa processing and the enforcement of contemporary immigration laws?
   1) Strongly agree
   2) Agree
   3) Neutral
   4) Disagree
   5) Strongly disagree
   6) Not applicable

When asking a long or difficult question, many individuals may skim the question and select "not applicable" if the question is too complex, takes too much time, requires too much thought, or must be read multiple times. The question above may need to be simplified or completely redesigned with introductory information for inclusion in the final survey.

"Other" Responses

Did respondents mark "other" on a frequent basis? If so, this may indicate that there are additional response options that may have been overlooked and need consideration. For example, a question asking about an individual's political party affiliation that lists only Democrat, Republican, and other as the options may lead to too many "other" responses, given that there are other common political party possibilities (Libertarians, independents, Greens, etc.).

Assuming that party affiliation is a major variable in the study, a researcher could, for example, be overlooking a potentially large group of Libertarians who have repeatedly marked "other." It could be problematic to merge their responses with the responses of other survey takers who mark "other"—such as independents, respondents who are undecided, and members of other political denominations like the Green Party, as these groups may have little in common other than not being members of one of the major parties. Including an open-ended response line for "other" to "please specify" in the pretest survey allows respondents to indicate precisely what they meant with their "other" responses. When analyzing the pretest results, if a given answer seems common among respondents, it might warrant inclusion as a separate, additional categorical response option.

Little or No Response Variation

How much variation is there across test respondents? Assuming that you have a heterogeneous test sample, the responses should differ across surveys, for the most part. If they don't, there may be a problem with the question itself, or, if the attribute measured really is universally shared, then the question may not be a necessary component of the survey at all. For example, in a study of adolescent risk behavior, the following question would likely have very little variation:

4. Before age five, how many times did you consume alcohol?
    1) Never
    2) 1–3 times
    3) 4–5 times
    4) 6 or more times

Since we know that an overwhelming majority of individuals did not consume alcohol before the age of five, most respondents will be inclined to select "never." When analyzing pretest data, a researcher should be mindful of questions with very little variation in responses. A lack of variation in responses can be an indication that a question is not relevant enough to warrant inclusion in the final survey.

Easily Misinterpreted Questions

Are there double-barreled questions in the survey or other questions that could possibly be misinterpreted? The following question is an example of a double-barreled question that was identified in the pretest of an open-ended survey:

5. Who do you feel closer to, your friends or your partner? Who's more fun? Tell us who and why.

The pretest of this open-ended question would allow researchers to note the double-barreled nature of the question. For instance, when respondents replied with "my partner," researchers would be unable to determine if the response indicated that an individual was closer to the partner, if the partner was more fun, or both.

Other misleading and confusing types of questions are those that are posed with negative wording. If a negative clause is necessary, try to avoid adding additional ones. For example, the double negative in the following question would be very likely to confuse a respondent:

6. Is it true that *not* ending the war will lead to *negative* economic outcomes?

Perhaps even more confusing are complex questions posed with double and triple negatives in the wording that would take multiple attempts to interpret: "Would you *disagree* that *not* ending the war would have a *negative* influence on the economy?" See Chapter 4 for more examples of confusing items and ways to improve them.

Sensitive Questions

Pretesting helps researchers determine whether or not respondents will be overly sensitive to specific questions, causing respondents to hesitate, hold back, or skip survey items. A pretest design can also help determine where sensitive and private questions work best within the overall layout of the survey.

Inconsistent Scales

The pretest allows researchers to verify that all scales are standardized to include the same number of points. For instance, if a question asks individuals to rate their concern with current environmental issues on a 1–7 scale where 1 = "very concerned," it is unwise to follow this up with a scale where 5 = "very concerned." In this case, respondents may condition their response to the scale from the first question and unintentionally report the opposite answer in the second question.

A pretest also helps researchers ensure that respondents are able to differentiate the points on a scale clearly. If a 9-point scale has too many options and does not

allow the respondent to differentiate between them, a smaller 5- or 7-point scale may be warranted.

Order of Response Options

Are responses influenced by the order of questions or response options? The order of possible responses in the following question presents this issue:

7. Which of the following cause you a lot of stress these days? (Please check all that apply.)

   ☐ Job
   ☐ Pets
   ☐ Financial situation
   ☐ Household chores
   ☐ My health
   ☐ Relationships with friends
   ☐ Relationships with family members
   ☐ Relationship with my partner
   ☐ Not enough time to spend with people I care about
   ☐ Not able to exercise enough
   ☐ Too little personal time to myself
   ☐ Day-to-day responsibilities that come with raising kids
   ☐ Other
   ☐ None of the above

Let us assume that "job" is really the item causing our respondent the most stress. If this option is presented at the top of the list, the respondent will likely have no trouble finding and checking the appropriate box. However, if the response is listed closer to the end of the list, the individual may be influenced to choose other preceding list items first, identifying health, friends, and household chores as sources of stress. By the time he reaches the "job" option, he has been distracted by other thoughts and may no longer consider this to be a significant cause of stress. This bias toward marking early response options to the exclusion of later ones is a common phenomenon. A possible safeguard against this is to randomize the list so that the order is different across each survey. (Of course, this is a much more practical approach in computer assisted surveys, which can be randomized much more efficiently than can pen-and-paper surveys.) Or, it may be more effective to shorten or collapse the list or redesign the question altogether.

Computer-Based and Technical Problems

Are skip patterns correctly designed and implemented? If you used a progress bar on a computer-based survey, is it functioning properly throughout the instrument? Are survey filters working correctly? Difficulties here may require simple reprogramming of the instrument, or they may encourage an alternative method of administration.

Other Pretesting Issues

Even if pretest survey items are not skipped, answered incorrectly, or identified as problematic in follow-up interviews, researchers should be attuned to a few more issues and opportunities this valuable tool presents.

Check and Improve Respondent Recall

Pretests are a good way to determine whether or not recall is too strenuous for retrospective questions. Questions may need to be more clearly defined, and respondents may need to be given specific time references to help them recall events that occurred in the past.

Clarify Complex Concepts

Researchers can ask pretest respondents to define specific concepts in order to help design questions for those concepts in the full-scale survey. For example, a researcher might ask test respondents to "define what the term *family values* means to you." Since "family values" means a number of different things to different people, a researcher can assess the different meanings individuals associate with this term.

Track Question Response Timing

Web-based surveys allow for the assessment of response latency, the timing of respondents' completion of individual questions, which is helpful to determine if certain questions can be streamlined for quicker response. Some survey software may have additional features to track the number of clicks a respondent makes on a given page (Heerwegh 2003).

Assess Adequacy of Space for Reponses to Open-Ended Questions

The pretest allows researchers to assess whether or not more room is needed for open-ended responses. This is important for pen-and-paper surveys, where physical space is

important. However, this also allows for assessment of the space needed in certain web-based surveys that allow only a certain number of characters in an open-ended response field.

Assess Survey Appearance on Varying Media

Researchers should pretest each medium used to collect data. For example, can web-based surveys be taken on a smartphone without zooming in and out to view questions?

Updating Time-Specific Surveys and Multilanguage Surveys

When using or building on existing instruments, pretesting the instrument is important to assess whether the questions have stood the test of time. Respondents may draw attention to questions with outdated wording or poorly designed content. After the pretest, researchers can compare these preliminary results with those from existing studies using the results from the pretest as a cross-validation of the measure's accuracy.

Pretesting is also necessary when using surveys developed in multiple languages (McKay et al. 1996). Translation requires rigorous understanding of different words, phrases, and colloquial meanings. In addition to arranging for professional translation, researchers can use a pretesting audience to help identify inconsistencies or irregularities in language and/or cultural accessibility issues within the survey instrument (Ferketich, Phillips, and Verran 1993).

After pretesting, the major issues with questions, measurement, and design should become apparent. Researchers should then prepare a memo summarizing all of the concerns about the survey for their research team and, using data gathered from the survey assessment(s), make revisions to the survey design to improve the quality of the questionnaire and its resultant data. Usually, this produces a more complete survey instrument that is ready for fielding or pilot testing. However, depending on the amount of content revised, another round of pretesting might be in order.

Pretesting Checklist

Refer to the following checklist as a summary of the above issues and guidelines. Careful consideration of each of the elements will help you make the most of your pretest and produce the most effective survey instrument possible.

ADMINISTRATION

- ☐ How long does the survey take to complete?
- ☐ Did the time to complete the survey vary widely among the test participants?

- ☐ Are the instructions for each section clear and unambiguous?
- ☐ Did you thank the respondents for their time?

ORGANIZATION

- ☐ Do the different sections flow reasonably from one to the next?
- ☐ Are all questions necessary in order to collect information on your topic?
- ☐ Are the questions within each section logically ordered?

CONTENT

- ☐ Are the questions direct and concise?
- ☐ Are the questions measuring what they are intended to measure?
- ☐ Are the questions free of unnecessary technical language and jargon?
- ☐ Are examples and analogies relevant for individuals of other cultures?
- ☐ Are questions unbiased?
- ☐ Are there questions that make respondents feel uncomfortable, embarrassed, annoyed, or confused? If so, can these be worded differently to avoid doing so?
- ☐ Are the response choices mutually exclusive and exhaustive?
- ☐ Are all response options necessary for inclusion?

When pretesting is complete, be sure to include your use of this procedure in the methods section of your research paper, as it increases the credibility of your research.

PILOT TESTING

In a **pilot test** (also known as a *feasibility study*), the interviewers, final survey, and some stages of coding and analysis are rehearsed prior to the actual survey administration. In other words, a pilot study is a trial run of the entire study from start to finish that increases the likelihood of success for the main study. Pilot studies are conducted to test the entire research process—usually from a methodological standpoint (e.g., sampling and recruitment strategies, administration, data collection and analysis) in actual field conditions.

Unlike survey pretests, which are usually done on a small scale, pilot studies have no cognitive interviews or focus groups to determine which measure and concepts are appropriate in the survey questionnaire. Rather, a pilot test is systematically administered to a diverse cross-section of the sample to ensure that the entire survey schedule runs smoothly and that coding and analysis can be done properly and efficiently. A general rule of thumb is to pilot test the survey on 30 to 100 pilot participants (this

number will vary, of course, depending on the number of respondents in your entire sample; Courtenay 1978). Once pilot testing is complete, final revisions to the survey process can be made, and the survey is ready for full-scale administration.

There are several reasons for undertaking a pilot study. They help identify potential problems throughout the entire survey procedure and assess whether the project is feasible, realistic, and rational from start to finish. Prior to administering the full-scale study, pilot research helps researchers address several issues that will affect the success of the study, as outlined below.

Necessary Resources

The pilot study helps the researcher determine what resources are necessary for the full study. The researcher is able to gauge the number of interviewers, staff, data coders, and analysts that will be necessary for the full-scale study and to identify what software will be necessary for the analysis. Researchers are also able to assess whether or not any incentives provided to respondents are commensurate with the time and energy necessary to complete the survey.

Trained Surveyors/Interviewers

The pilot test serves as a means to validate the field-testing process (i.e., interviewer training). The researcher is able to observe whether or not the interviewers are knowledgeable about the survey items, able to answer questions, and competent in clarifying points about the survey and research topic. The researcher is also able to assess whether the interviewers are objective and unbiased during the interview.

Administration Procedures

Pilot testing also allows a researcher to test for possible flaws with the sampling and administration of the survey. For example, in a study asking about fairness in housework, surveying both individuals in a couple together might lead to a bias in responses, because respondents would be more inclined to respond favorably about fairness in housework when their spouses were present. A pilot study would allow for comparison of responses where individuals in a couple were surveyed individually with those where both individuals were surveyed together. In preliminary analysis, researchers might find that the answers of simultaneously surveyed individuals in a couple are biased, and they might also determine that individual surveys lead to less biased responses; thus they could choose to conduct only individual surveys in the full-scale study.

Recruitment Approaches

Pilot studies are also important for uncovering problematic features of the sample. As an illustration, a pilot study conducted by one of the authors of this text explored the use of public spaces (e.g., parks and recreation centers) among homeless individuals. The study was conducted on a small sample of homeless individuals in a public park. Inadvertently, in the process of conducting the pilot study, the researchers brought unwanted public attention to homeless individuals. This attention led to the removal of the homeless from the park due to restrictions against loitering. This unforeseen issue brought an untimely end to the pilot test and also made a segment of the research population less likely to participate in future studies for fear of legal reprimand. As a result, the researchers had to develop alternate strategies for recruitment for the full-scale study. Ultimately, they learned that a completely different research design may have been more successful than the survey they administered.

Data Analysis

The pilot test also provides raw data to test data entry and data processing procedures. Preliminary coding and analysis should be completed to test the accuracy and capability of data analysis programs, allowing for correction of critical issues at even the latest stages of data collection. In addition, researchers should be able to identify whether initial data input, management, and coding are properly, effectively, and precisely executed in a timely fashion. Even if the results of the pilot study are ultimately not meaningful, preliminary analyses and tables should be produced from these results, simply to test the viability and efficacy of the survey *process* through to the final stages.

Similar to pretesting, pilot testing has practical importance for funding and support, because it lends credibility to the research project design. Thus, like pretests, pilot tests are an important way to convince stakeholders and funding sources that the study is workable and merits financial support.

PRETEST AND PILOT TEST LIMITATIONS

Successfully pretesting and pilot testing a survey does not necessarily ensure the success of the full study. There are a number of concerns that are not addressed in the pretesting or pilot stages of survey administration, and some issues may not arise until the full-scale study is conducted. It is important to be aware of these problems before they arise.

Firstly, given the smaller number of surveys administered in a pilot study, it cannot predict, or even estimate, a response rate for the full-scale study. Even if a study has been

vetted by a well-designed and successfully executed pilot study, the full-scale study may still suffer from extremely low response rates.

Another important problem that confronts researchers is contamination of the data as a result of including pretest and pilot test survey results in the final, full-scale study. Because modifications to the survey instrument may have taken place, the data collected in the pretest and pilot test of the surveys could be inaccurate or biased compared to the results of the full-scale study. For example, in the pretest, the ambiguous question, "How many times have you relocated in the last year?" might elicit a response of "twice" that was based on temporary moves. If the question were subsequently reworded to "How many times did you permanently change your residence in the last twelve months?" for the full-scale study, then including the pretest data (based on different wording) would add an incorrect response to the dataset and lower the overall quality of the survey. To avoid this, some researchers have chosen to redesign the questionnaire and readminister the revised survey to respondents who participated in the pretest or pilot study. However, this comes with its own set of complications and poses threats to the internal validity of the study. These respondents may respond differently than they would have otherwise responded had they not been conditioned by the pretesting experience; in experimental psychology, this is known as a "pretesting effect" (Richland, Kornell, and Kao 2009). For example, achievement scales may be positively influenced with a pretest: An individual completing a math problem would likely be more successful when the question is presented a second time, in the full-scale survey, after having first responded to it on the pretest. In other words, individuals who were involved in the pretest study will have experienced these questions, may have thought about them, and may be better equipped to answer when they respond to the survey in the full-scale study.

Of course, it may be unreasonable to exclude these participants from the entire study, especially in small-scale studies or with difficult-to-locate samples. In this case, comparison and discussion of the differences between the pretested groups and the full-scale group is necessary. It is also important to exercise caution when interpreting these results, and it is important to note this potential data contamination as a possible limitation of the research.

Finally, it is highly recommended to organize the timing of pretests and pilot studies to allow for analysis and revision before conducting the full-scale study. It is important to have enough time between administration of the pretest, the pilot test, and the full-scale survey so that edits to the survey can be implemented—and this can often be problematic when a pilot test does not go according to plan and major changes to the study design are necessary. The effort put into conducting a pretest and pilot study are wasted if the results are not made available in time for efficient planning of the full-scale survey.

Despite the time, financial cost, and energy invested in pretesting and pilot testing a survey, these procedures rarely warrant more than a single line in the methods section

of a research paper, if any mention at all (Presser et al. 2007). Given the strict word and space limitations required for publication, the presentation of the final research product typically receives more attention than descriptions of behind-the-scenes research. But these procedures are no less important for this fact; they allow you to learn from your methodological mistakes and to leave a detailed record of your efforts, so that you and others can avoid the pitfalls of survey development and administration in future research. For this reason, there are meta-analytic research reports (research undertaken on the process of research) that help us acknowledge these issues and bring them to the attention of the larger research community.

CONCLUSION

While it may require much time and energy for careful planning and high precision, testing the entire survey process from beginning to end is an essential process for good survey research. The sole purpose of pretesting and pilot testing early on is to reduce measurement error on a larger scale later. When the full-scale study begins, all research procedures should already be carefully checked and tested for errors and potential issues; when discovered during full-scale data collection, these glitches are much more difficult to remedy. Thus, it is important to acknowledge these problems early in the research process by conducting pretests and pilot studies. Well-organized and well-documented pretests and pilot surveys help improve the validity, reliability, accuracy, and efficiency of the full-scale study. It is better to spend a relatively modest amount of time and money uncovering potentially serious flaws in a survey instrument than to spend a serious amount of time and money engaging in potentially flawed research.

KEY TERMS

Pretest 101
Behavior coding 105
Cognitive interview 106
Individual debriefing assessment 106
Group debriefing assessment 107
Pilot test 101
Response latency 102

CRITICAL THINKING QUESTIONS

Below you will find four questions that ask you to think critically about core concepts addressed in this chapter. Be sure you understand each one; if you don't, this is a good time to review the relevant sections of this chapter.

1. What are two ways that pretesting a survey prior to administration might increase the instrument's *validity*?
2. What are two ways a pretest might improve the instrument's *reliability*?
3. What are some of the problems researchers might encounter if they *pretest* their survey but do not *pilot test* it?
4. What are some problems researchers might encounter if they *pilot test* their survey but do not *pretest* it?