SURVEYS · QUESTIONNAIRE DESIGN

Pilot Testing Surveys: Methods for Catching Problems Before Launch

A pilot test runs a survey on a small scale before the main fieldwork. This article explains what pilots can reveal, how they fit with other pretesting methods, how to plan and analyse one and what to check before launch.

ModQu research article · 1,436 words · 16 references (APA 7th edition)

Every survey contains mistakes when it is first written. Some are obvious once spotted: a skip instruction that sends respondents to the wrong question, a missing answer option, a typing error in a scale label. Others are subtle: a question that most people interpret differently from the designer, a section that takes far longer than expected or a screen that does not display properly on a phone. Once the main survey is in the field, these mistakes are expensive or impossible to fix. A **pilot test**, in which the survey is run on a small scale under realistic conditions before the main launch, is the last and often most effective chance to catch them.

This article explains what pilot testing can reveal, how it fits with other pretesting methods, how to plan and analyse a pilot and what to check before launching the main survey.

What a pilot test is

A pilot test administers the survey, or a close version of it, to a small sample of people similar to the target population, using the same mode, procedures and technology planned for the main study. When the pilot reproduces the main study as closely as possible, from invitation to final data file, it is sometimes called a **dress rehearsal**. The aim is to find out whether the questionnaire and procedures work in practice, not to produce results.

Pilot testing is one of several pretesting methods, each of which finds different kinds of problem (Presser, Couper, et al., 2004). No single method is sufficient, and the most thorough approach combines several in sequence (Presser, Rothgeb, et al., 2004).

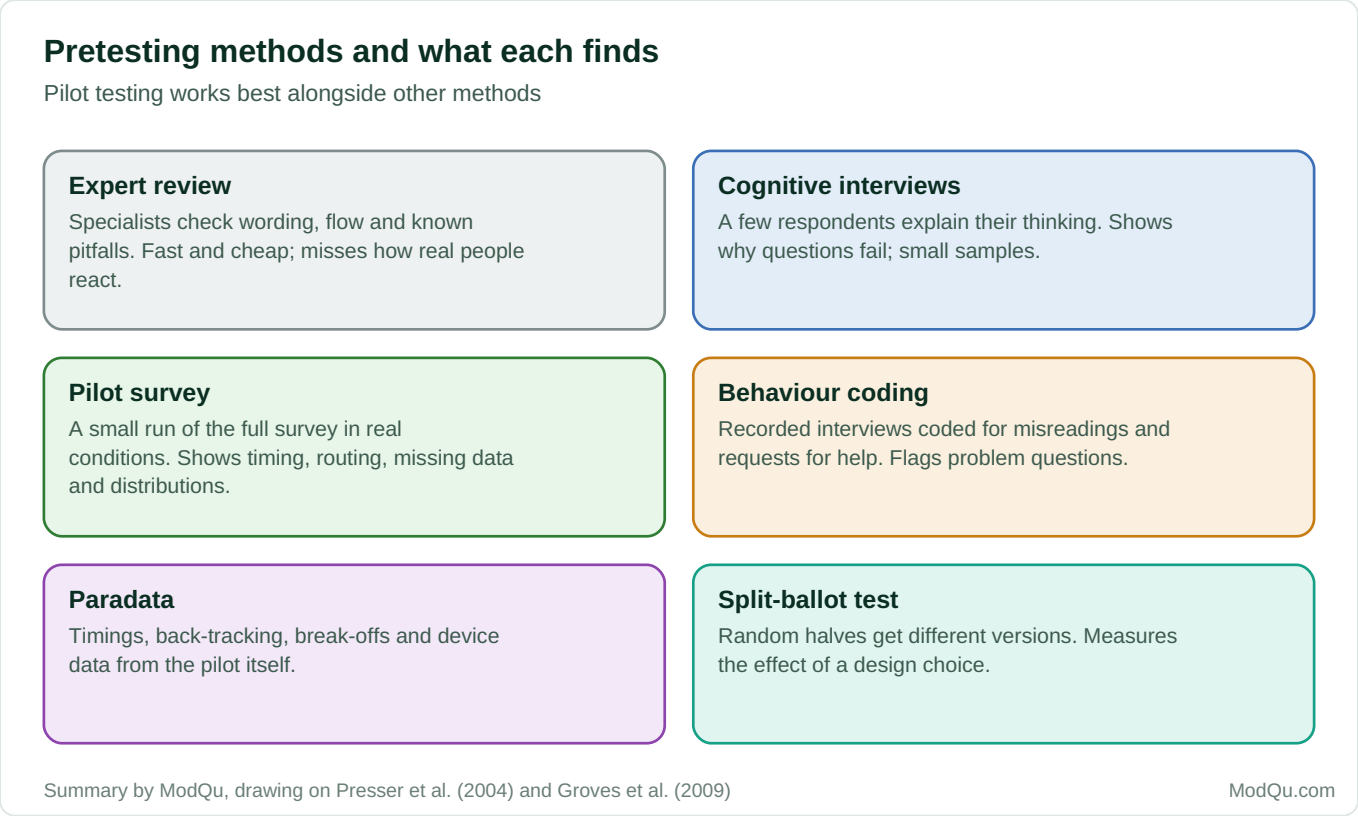


Figure 1. Pretesting methods and what each is best at finding. Summary by ModQu, drawing on Presser et al. (2004) and Groves et al. (2009).

Expert review and cognitive interviews are usually done first, while the questionnaire is still being drafted. Cognitive interviews are especially good at showing why respondents misunderstand questions (Willis, 2005). The pilot comes later, once the questionnaire is close to final, and tests the whole survey as a working system.

What a pilot can reveal

Problems with questions

Pilot data show how each question performs in a realistic setting. Warning signs include:

- **High item nonresponse**, which may mean a question is confusing, sensitive or not applicable to many respondents.
- **Many "don't know" or "other" answers**, which may point to unclear wording or missing answer options.
- **Unexpected distributions**, such as almost everyone choosing the same answer, which may mean the question does not discriminate or the options do not fit.
- **Heaping or implausible values** in numeric answers.
- **Write-in answers** that reveal categories the designer did not anticipate.

In interviewer surveys, pilot interviews can be recorded and coded for signs of difficulty, such as interviewers changing the wording, respondents asking for clarification or giving answers that do not fit the options. This technique, called **behaviour coding**, flags questions that cause problems in actual interviews (Oksenberg et al., 1991; Fowler & Cannell, 1996). In computer-assisted surveys, the time taken to answer each question can serve a similar purpose: unusually long response times often indicate difficulty (Bassili & Scott, 1996).

Problems with flow and technology

A pilot tests whether the questionnaire works as a whole. Skip patterns and routing can be checked by examining whether respondents received the questions they should have. Web surveys must display correctly across browsers and devices, and pilots often reveal layouts that are hard to use on small screens. Observing people as they use computer-assisted questionnaires, an approach borrowed from usability testing, can expose navigation problems that designers did not anticipate (Couper, 2000).

Length and burden

A pilot gives the first reliable measure of how long the survey takes. Designers almost always underestimate length. Longer questionnaires reduce participation and the quality of answers, especially near the end (Galesic & Bosnjak, 2009), so a pilot that shows the survey runs ten minutes over its target is a clear signal to cut.

Process data

Modern data collection systems record **paradata**: information about the process of answering, such as timestamps, changed answers, back-tracking, break-off points and the device used (Kreuter, 2013). Pilot paradata show where respondents slow down, where they go back to change answers and where they abandon the survey. A spike in break-offs at a particular question is one of the clearest signs that something is wrong with it.

Procedures

Pilots also test everything around the questionnaire: invitation letters and emails, reminders, incentives, sample management and the processes for coding and cleaning data. Response rates in the pilot give an early indication of what to expect in the main study, though small pilots give imprecise estimates.

Planning a pilot

How many respondents?

Pilot samples are usually small. For checking questionnaire function and spotting major problems, a few dozen completed surveys often suffice. Larger pilots, of one hundred or more, are needed to estimate response rates or item distributions with useful precision, or to compare alternative versions (Hertzog, 2008). The sample should include the full range of people who will be surveyed, especially groups likely to have difficulty or to follow unusual routes through the questionnaire.

Who should take part?

Pilot respondents should come from the target population, but they should not normally be included in the main sample, because having already answered the survey changes how they respond. Colleagues and friends are useful for initial checks of technical function, but they are no substitute for real respondents (Converse & Presser, 1986).

Add debriefing

A pilot can include extra questions at the end asking respondents whether any questions were unclear, difficult or uncomfortable, and why. In interviewer surveys, interviewers can be debriefed about questions that caused problems. These debriefings often explain the patterns seen in the data (Fowler, 1995).

Test alternatives

When designers are unsure between two versions of a question, the pilot can randomly assign respondents to each. With a large enough pilot, this **split-ballot** design shows whether the choice makes a difference.

Acting on the results

The value of a pilot depends on acting on what it reveals. Problems should be listed, prioritised and fixed, and significant changes should ideally be tested again. A short report recording what was found and what was changed helps others understand the final design and provides evidence for the quality of the survey (Groves et al., 2009). Computer tools that flag potentially confusing wording (Graesser et al., 2006) and models that predict the measurement quality of question formats (Saris & Gallhofer, 2014) can complement the pilot but do not replace testing with real respondents.

An example

Consider a customer satisfaction survey planned for 5,000 customers of a software company. A pilot with 60 customers, run on the same platform with the same invitation, might show the following (illustrative figures). The median completion time is 14 minutes against a target of 10. Almost a fifth of respondents leave a question about "integration quality" blank, and debriefing answers show that many do not know what integration means in this context. A grid of twelve satisfaction items shows a sharp rise in break-offs on phones, where the grid is hard to read. One routing rule sends customers on a free plan to questions about billing that do not apply to them.

Each of these findings leads to a specific change: cutting or shortening sections to reach the time target, rewording or removing the integration question, splitting the grid into shorter groups that display well on small screens and correcting the routing. None of these problems would have been visible in a review of the questionnaire on paper, and all of them would have reduced the quality of the main survey data.

A pre-launch checklist

Before the main survey goes live, the pilot should have confirmed that:

1. Every question is understood as intended by the people who will answer it.
2. Answer options cover the answers people want to give.
3. Skip patterns and routing work for every type of respondent.
4. The survey displays and functions correctly on all expected devices.
5. The median completion time is within the target.
6. No question has unexpectedly high missing data or break-offs.
7. Invitations, reminders and incentives work as planned.
8. The data file contains every variable, correctly labelled and coded.

Conclusion

A pilot test is a small investment that protects a much larger one. By running the survey under realistic conditions with real respondents, it reveals problems with questions, flow, technology, length and procedures while they can still be fixed. Pilots work best as the final step in a sequence of pretesting methods that begins with expert review and cognitive interviews. Skipping the pilot to save time rarely saves anything: the problems it would have found remain in the data (Dillman et al., 2014).

References

Bassili, J. N., & Scott, B. S. (1996). Response latency as a signal to question problems in survey research. *Public Opinion Quarterly, 60*(3), 390–399. https://doi.org/10.1086/297760

Converse, J. M., & Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. SAGE. https://doi.org/10.4135/9781412986045

Couper, M. P. (2000). Usability evaluation of computer-assisted survey instruments. *Social Science Computer Review, 18*(4), 384–396. https://doi.org/10.1177/089443930001800402

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, phone, mail, and mixed-mode surveys: The tailored design method* (4th ed.). Wiley.

Fowler, F. J., Jr., & Cannell, C. F. (1996). Using behavioral coding to identify cognitive problems with survey questions. In N. Schwarz & S. Sudman (Eds.), *Answering questions: Methodology for determining cognitive and communicative processes in survey research* (pp. 15–36). Jossey-Bass.

Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation*. SAGE.

Galesic, M., & Bosnjak, M. (2009). Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public Opinion Quarterly, 73*(2), 349–360. https://doi.org/10.1093/poq/nfp031

Graesser, A. C., Cai, Z., Louwerse, M. M., & Daniel, F. (2006). Question Understanding Aid (QUAID): A web facility that tests question comprehensibility. *Public Opinion Quarterly, 70*(1), 3–22. https://doi.org/10.1093/poq/nfj012

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Wiley.

Hertzog, M. A. (2008). Considerations in determining sample size for pilot studies. *Research in Nursing & Health, 31*(2), 180–191. https://doi.org/10.1002/nur.20247

Kreuter, F. (Ed.). (2013). *Improving surveys with paradata: Analytic uses of process information*. Wiley. https://doi.org/10.1002/9781118596869

Oksenberg, L., Cannell, C., & Kalton, G. (1991). New strategies for pretesting survey questions. *Journal of Official Statistics, 7*(3), 349–365.

Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., & Singer, E. (2004). Methods for testing and evaluating survey questions. *Public Opinion Quarterly, 68*(1), 109–130. https://doi.org/10.1093/poq/nfh008

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (Eds.). (2004). *Methods for testing and evaluating survey questionnaires*. Wiley. https://doi.org/10.1002/0471654728

Saris, W. E., & Gallhofer, I. N. (2014). *Design, evaluation, and analysis of questionnaires for survey research* (2nd ed.). Wiley. https://doi.org/10.1002/9781118634646

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design*. SAGE. https://doi.org/10.4135/9781412983655