SURVEYS · QUESTIONNAIRE DESIGN

Survey Experiments: Testing How Question Design Affects Answers

A survey experiment randomly assigns respondents to different versions of a question or questionnaire. This article explains how split-ballot and other survey experiments work, what they have taught survey research and how to design, analyse and interpret them.

ModQu research article · 1,229 words · 15 references (APA 7th edition)

How can a survey designer know whether one wording is better than another, or whether the order of questions changes the answers? Opinions and rules of thumb are not enough, because effects that seem obvious often turn out to be small, and effects that seem trivial can be large. The most reliable way to find out is to test: to give different versions to different respondents, chosen at random, and compare the results. This is a **survey experiment**.

This article explains how survey experiments work, what they have revealed about question design and how to design and interpret them.

How a survey experiment works

In the simplest design, the **split-ballot experiment**, the sample is divided at random into two or more groups, and each group receives a different version of a question. Because the groups are formed at random, they are alike on average in every respect except the version they received. Any difference in their answers larger than chance can therefore be attributed to the difference between the versions (Schuman & Presser, 1981).

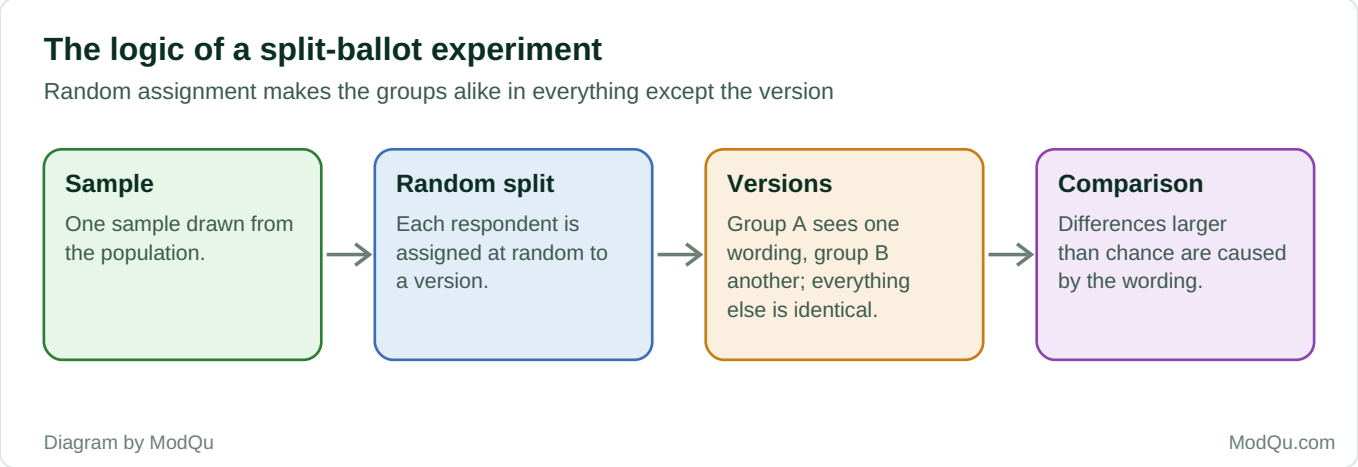


Figure 1. The logic of a split-ballot survey experiment. Diagram by ModQu.

Combining random assignment with a sample drawn from a defined population gives survey experiments a particular strength. Laboratory experiments establish cause and effect but usually with small, unrepresentative groups; ordinary surveys describe populations but struggle to establish causes. A survey experiment on a probability sample can do both (Mutz, 2011).

What experiments have shown about questions

Some of the best-known findings in survey methodology come from split-ballot experiments. In a classic example from 1940, respondents were asked whether the United States should "forbid" or "not allow" speeches against democracy. Far more said the country should "not allow" such speeches than said it should "forbid" them, although the two wordings appear logically equivalent (Rugg, 1941).

Systematic programmes of experiments later examined many aspects of question design: open versus closed questions, offering a middle option, offering "don't know", balanced versus unbalanced wording and the order of questions (Schuman & Presser, 1979, 1981). In Britain, experiments tested the effects of wording opinion questions in different ways and of offering a middle option (Kalton et al., 1978). More recent experiments have compared formats for web surveys, such as check-all-that-apply versus forced-choice lists (Smyth et al., 2006), and the effect of stated survey length on participation (Galesic & Bosnjak, 2009).

Beyond question wording

Survey experiments are also used to study attitudes and decisions, not only questions. In a **framing experiment**, respondents receive the same information described in different ways, and their reactions are compared. A famous example showed that people chose differently between two health programmes depending on whether outcomes were described as lives saved or lives lost (Tversky & Kahneman, 1981). Framing experiments are widely used to study how the presentation of issues affects public opinion (Chong & Druckman, 2007).

In a **vignette** or **factorial survey experiment**, respondents rate short descriptions of people or situations in which several characteristics are varied at random, for example the age, sex and qualifications of a job applicant. The design shows how much each characteristic influences judgements (Auspurg & Hinz, 2015). **Conjoint experiments**, used widely in market research, ask respondents to choose between profiles that vary on several attributes, such as price, brand and features, and estimate the effect of each attribute on choice (Hainmueller et al., 2014).

Designing a survey experiment

Randomise properly

Assignment to versions must be truly random and handled by the survey software, not by interviewers or respondents. The software should record which version each respondent received.

Change one thing at a time

If versions differ in several ways, any difference in results cannot be attributed to a particular change. Factorial designs can test several changes at once, but only if each is varied independently.

Plan the sample size

Differences between versions are often modest, and detecting them requires adequate samples in each group. A power calculation before fieldwork shows how many respondents are needed to detect a difference of a given size.

Think about what is being measured

An experiment shows that two versions give different answers, but not which is more accurate. To judge accuracy, results must be compared with an external standard, such as records, or with evidence of reliability and validity.

Analysing results

The basic analysis compares the distribution of answers between versions, using a test suited to the type of question, such as a chi-square test for categories or a comparison of means for scales. Because assignment is random, simple comparisons are valid without controlling for other variables, although adding respondent characteristics to a regression model can make estimates more precise. Experiments can also show whether an effect differs between groups, for example whether a wording change affects older respondents more than younger ones. Such **interaction effects** require larger samples to detect and should ideally be specified before the data are collected, to avoid finding patterns by chance.

An example

Suppose a company wants to know whether describing a price increase as "a 5 per cent increase" or as "an extra $2 a month" changes customers' stated intention to stay. In an online survey of 2,000 customers, the software randomly assigns half to each description. If 71 per cent of the first group and 64 per cent of the second say they would stay, and the difference is well beyond what chance would produce with groups of this size, the company has evidence that the description itself affects reactions (illustrative figures). The experiment does not show which description customers would respond to in practice, but it shows that the choice matters and should be made deliberately.

Interpreting results

Survey experiments have strong **internal validity**: the difference between versions is a reliable estimate of the effect of the change for the people in the sample. Whether the result applies elsewhere, its **external validity**, depends on how the sample, setting and treatment compare with the situations of interest (Shadish et al., 2002). Critics have pointed out that a treatment delivered in a survey, such as a short news story, may differ from real-world exposure, and that effects may fade quickly (Gaines et al., 2007). Studies comparing the same experiments across different kinds of sample, such as probability samples and online convenience samples, have found that effects are often similar, which is encouraging, though not guaranteed for every topic (Mullinix et al., 2015).

In business surveys

Survey experiments are inexpensive to add to almost any online survey. A company can test two versions of a satisfaction question, two descriptions of a new product or two price levels, simply by assigning respondents at random. Experiments built into regular surveys allow evidence about design choices to accumulate over time. When a survey is changed, running the old and new versions side by side on random halves of the sample shows whether any change in results reflects real change or the new design (Groves et al., 2009).

Conclusion

Survey experiments bring the logic of randomised experiments into surveys. They have shown that small changes in wording, format and order can change answers substantially, and they allow designers to test choices rather than guess. Well-designed experiments vary one thing at a time, use proper randomisation and adequate samples and are interpreted with attention to both what they prove and where their results apply. For any organisation that relies on survey data, they are one of the most useful tools for improving both questions and knowledge.

References

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. SAGE. https://doi.org/10.4135/9781483398075

Chong, D., & Druckman, J. N. (2007). Framing theory. *Annual Review of Political Science, 10*, 103–126. https://doi.org/10.1146/annurev.polisci.10.072805.103054

Gaines, B. J., Kuklinski, J. H., & Quirk, P. J. (2007). The logic of the survey experiment reexamined. *Political Analysis, 15*(1), 1–20. https://doi.org/10.1093/pan/mpl008

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

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

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis, 22*(1), 1–30. https://doi.org/10.1093/pan/mpt024

Kalton, G., Collins, M., & Brook, L. (1978). Experiments in wording opinion questions. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 27*(2), 149–161. https://doi.org/10.2307/2346942

Mullinix, K. J., Leeper, T. J., Druckman, J. N., & Freese, J. (2015). The generalizability of survey experiments. *Journal of Experimental Political Science, 2*(2), 109–138. https://doi.org/10.1017/XPS.2015.19

Mutz, D. C. (2011). *Population-based survey experiments*. Princeton University Press.

Rugg, D. (1941). Experiments in wording questions: II. *Public Opinion Quarterly, 5*(1), 91–92.

Schuman, H., & Presser, S. (1979). The open and closed question. *American Sociological Review, 44*(5), 692–712. https://doi.org/10.2307/2094521

Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Academic Press.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Smyth, J. D., Dillman, D. A., Christian, L. M., & Stern, M. J. (2006). Comparing check-all and forced-choice question formats in web surveys. *Public Opinion Quarterly, 70*(1), 66–77. https://doi.org/10.1093/poq/nfj007

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458. https://doi.org/10.1126/science.7455683