SURVEYS · QUESTIONNAIRE DESIGN

Ranking Questions Versus Rating Questions: What Each Reveals

Rating questions ask how important each item is; ranking questions ask which matters most. This article explains how the two formats differ, what the research shows about their results and how best-worst scaling combines their strengths.

ModQu research article · 1,488 words · 16 references (APA 7th edition)

A company wants to know what customers value most in a delivery service: speed, price, reliability, tracking or friendly drivers. It can ask customers to rate each feature on a scale from "not at all important" to "extremely important," or it can ask them to rank the features from most to least important. Both approaches seem reasonable. They often lead to different conclusions.

The difference arises because the two formats measure different things. **Rating questions** measure how important each item is on its own. **Ranking questions** measure how important items are relative to each other. This article explains the strengths and weaknesses of each format, summarises the research comparing them and describes best-worst scaling, a method that combines some of the advantages of both.

The problem with rating everything

Rating questions are easy to write and answer. Each item is judged on its own, respondents can work quickly and the results can be analysed with familiar statistics. The main weakness is that respondents can rate every item as important. In the delivery example, many customers will rate speed, price and reliability as "very important" or "extremely important." The ratings may be accurate, since all three do matter, but they do not reveal which one the company should prioritise.

This tendency is called **non-differentiation**. It is partly genuine, because many things do matter, and partly a form of **satisficing**, in which respondents give the same answer to a series of items to reduce effort (Krosnick, 1991). Respondents also differ in how they use rating scales: some use the top of the scale for almost everything, while others are more sparing, and these styles differ across countries and groups (Baumgartner & Steenkamp, 2001). As a result, differences in ratings between people or groups may reflect scale use rather than differences in what they value.

The strengths and costs of ranking

Ranking questions force respondents to make trade-offs. Only one item can be ranked first, so the format reveals priorities directly. It also removes differences in scale use, since every respondent produces the same set of ranks. These properties have made ranking the traditional choice for measuring values. Milton Rokeach's influential value survey

asked respondents to rank eighteen terminal values, such as freedom and family security, in order of importance (Rokeach, 1973). Ronald Inglehart's widely used index of materialist and post-materialist values asks respondents to choose the most and second most important of four national goals (Inglehart, 1977).

Ranking has costs. It is harder and slower than rating, especially when the list is long, because respondents must compare every item with every other. In practice, people tend to rank the first few and last few items with care and the middle ones more arbitrarily. Ranking also prevents respondents from saying that two items are equally important, even when they are. And it reveals nothing about absolute importance: an item ranked last may still be very important to the respondent, or not important at all.

Ranked data also raise statistical problems. Because each respondent's ranks must add up to the same total, the ranks are not independent of each other. If one item moves up, another must move down. This property, known as ipsativity, complicates the use of standard methods such as factor analysis and correlation (Ovadia, 2004).

What the research shows

The most influential comparison was carried out by Duane Alwin and Jon Krosnick, who randomly assigned respondents in a national survey to rate or rank a set of qualities that parents might want in their children, such as obedience, honesty and good sense (Alwin & Krosnick, 1985). The average ordering of the qualities was similar in both formats: the qualities seen as most important by raters were also the ones ranked highest. But the underlying structure of the answers differed. Ratings were more strongly affected by a general tendency to rate everything high or low, while rankings captured trade-offs between particular values more clearly. A follow-up study found that the two formats also produced different correlations with other variables, so conclusions about who holds which values could depend on the format (Krosnick & Alwin, 1988).

Other studies have reached more favourable conclusions about ratings. Research on the measurement of personal values found that ratings could serve as a practical substitute for rankings in some marketing applications (Munson & McIntyre, 1979), and one study comparing the two formats found that ratings showed stronger validity in predicting related attitudes (Maio et al., 1996). Researchers have also tested hybrid procedures. One effective approach asks respondents first to identify the most and least important items and then to rate all of them, which reduces non-differentiation while keeping the advantages of rating (McCarty & Shrum, 2000).

In cross-national research, where differences in scale use are a particular concern, ranking has been found to reduce some response biases, including the tendency of respondents in some countries, or answering in a second language, to give more extreme or more positive ratings (Harzing et al., 2009).

Best-worst scaling

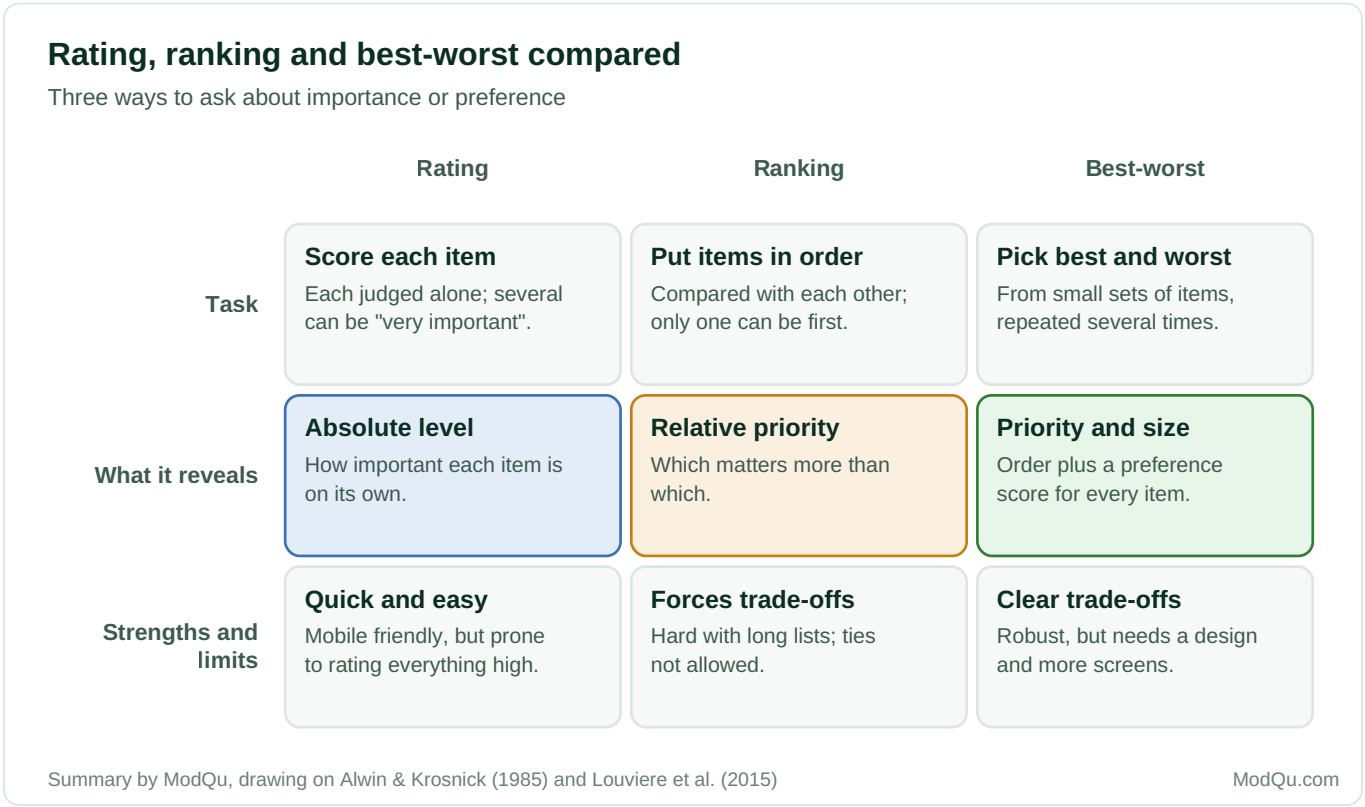


Figure 1. Rating, ranking and best-worst scaling compared. Summary by ModQu, drawing on Alwin and Krosnick (1985) and Louviere et al. (2015).

A third approach, **best-worst scaling**, also known as maximum difference scaling or MaxDiff, was developed to keep the trade-offs of ranking while avoiding its burden (Finn & Louviere, 1992). Respondents see a series of small sets of items, typically four or five at a time, and for each set choose the most and least important. Across many sets, drawn from an experimental design so that every item appears several times alongside different others, the choices produce a score for every item on a common scale (Louviere et al., 2015).

Best-worst scaling has several advantages. Choosing the best and worst from a small set is easy and natural. Respondents cannot rate everything as important, so the method discriminates well between items. Differences in scale use do not arise, which makes the method attractive for comparisons between countries and groups. The resulting scores show both the order of importance and the size of the gaps between items. The main costs are the need for a proper experimental design, a longer questionnaire with several screens and analysis methods that are less familiar to many users.

An example

Return to the delivery company. Suppose 1,000 customers rate five features on a five-point scale. Speed, price and reliability each receive "very" or "extremely important" from around 80% of customers, tracking from 60% and friendly drivers from 45% (illustrative numbers). The ratings suggest three equally important features and two less important ones, but they do not tell the company where to invest.

Now suppose the same customers complete a short best-worst exercise. Reliability is chosen as most important far more often than anything else, and price is chosen as least important surprisingly often among customers who order frequently. The ratings were not wrong: customers do care about all three leading features. But the trade-off data show that a failed delivery costs more goodwill than a slightly higher price, and that frequent customers are less price-sensitive than occasional ones. That is the kind of finding that changes decisions.

The two formats can also be combined in one questionnaire. Rating questions can measure the absolute level of importance for benchmarking, and a short ranking or best-worst task can establish priorities. Reporting both, and explaining what each shows, gives decision makers a fuller picture than either format alone.

Choosing a format

The right format depends on what the research needs to know.

- **Use ratings** when absolute levels matter, when items will be compared with external benchmarks, when lists are long or when the survey must be short and mobile friendly. Consider asking respondents first to pick the most important item, which reduces non-differentiation.
- **Use rankings** when priorities are what matter, when the list is short (around five to seven items) and when respondents can reasonably compare all items at once. For longer lists, ask respondents to rank only their top three.
- **Use best-worst scaling** when clear priorities among many items are needed, especially for product features, messages or policies, and when comparisons between groups or countries are important.

Whatever the format, the order of items should be randomised to prevent position effects, and the list should be tested to make sure all items are understood and distinct (Krosnick & Presser, 2010; Dillman et al., 2014).

Conclusion

Rating and ranking questions are not interchangeable. Ratings show how important each item is on its own but often fail to separate items that are all rated highly. Rankings reveal priorities but are more demanding, cannot show absolute importance and create statistical complications. Research shows that the two formats often agree on broad orderings but can lead to different conclusions about structure and about differences between groups. Best-worst scaling offers a practical middle way for many business and policy questions. Choosing the format to suit the decision at hand, whether it requires absolute levels, relative priorities or both, is the key to getting useful answers (Sudman & Bradburn, 1982).

References

Alwin, D. F., & Krosnick, J. A. (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly, 49*(4), 535–552. https://doi.org/10.1086/268949

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

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

Finn, A., & Louviere, J. J. (1992). Determining the appropriate response to evidence of public concern: The case of food safety. *Journal of Public Policy & Marketing, 11*(2), 12–25. https://doi.org/10.1177/074391569201100202

Harzing, A.-W., Baldueza, J., Barner-Rasmussen, W., Barzantny, C., Canabal, A., Davila, A., Espejo, A., Ferreira, R., Giroud, A., Koester, K., Liang, Y.-K., Mockaitis, A., Morley, M. J., Myloni, B., Odusanya, J. O. T., O'Sullivan, S. L., Palaniappan, A. K., Prochno, P., Roy Choudhury, S., … Zander, L. (2009). Rating versus ranking: What is the best way to reduce response and language bias in cross-national research? *International Business Review, 18*(4), 417–432. https://doi.org/10.1016/j.ibusrev.2009.03.001

Inglehart, R. (1977). *The silent revolution: Changing values and political styles among Western publics*. Princeton University Press.

Krosnick, J. A., & Alwin, D. F. (1988). A test of the form-resistant correlation hypothesis: Ratings, rankings, and the measurement of values. *Public Opinion Quarterly, 52*(4), 526–538. https://doi.org/10.1086/269128

Krosnick, J. A., & Presser, S. (2010). Question and questionnaire design. In P. V. Marsden & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263–313). Emerald.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology, 5*(3), 213–236. https://doi.org/10.1002/acp.2350050305

Louviere, J. J., Flynn, T. N., & Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press. https://doi.org/10.1017/CBO9781107337855

Maio, G. R., Roese, N. J., Seligman, C., & Katz, A. (1996). Rankings, ratings, and the measurement of values: Evidence for the superior validity of ratings. *Basic and Applied Social Psychology, 18*(2), 171–181. https://doi.org/10.1207/s15324834basp1802_4

McCarty, J. A., & Shrum, L. J. (2000). The measurement of personal values in survey research: A test of alternative rating procedures. *Public Opinion Quarterly, 64*(3), 271–298. https://doi.org/10.1086/317989

Munson, J. M., & McIntyre, S. H. (1979). Developing practical procedures for the measurement of personal values in cross-cultural marketing. *Journal of Marketing Research, 16*(1), 48–52. https://doi.org/10.1177/002224377901600107

Ovadia, S. (2004). Ratings and rankings: Reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology, 7*(5), 403–414. https://doi.org/10.1080/1364557032000081654

Rokeach, M. (1973). *The nature of human values*. Free Press.

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