SURVEYS · QUESTIONNAIRE DESIGN

# Adaptive Testing Principles Applied to Survey Questionnaire Design

Computerised adaptive testing chooses each question according to the respondent's earlier answers, so that fewer questions are needed for the same precision. This article explains the principles, how they have been applied to surveys and what is needed to use them.

ModQu research article · 1,330 words · 14 references (APA 7th edition)

In a conventional questionnaire, everyone answers the same questions in the same order. This is simple and makes answers easy to compare, but it is inefficient. A question about severe difficulty walking tells us little about a marathon runner, and a question about mild tiredness tells us little about someone already known to be exhausted. **Computerised adaptive testing** takes a different approach: each question is chosen according to the answers already given, so that every respondent receives the questions that are most informative for them (Weiss, 1982; Wainer, 2000). The result is often the same precision with far fewer questions.

The method was developed for educational and ability testing, but its principles apply to any survey that measures a trait or attitude with several items. This article explains how adaptive testing works, how it has been applied to surveys and what organisations need before using it.

## The principles

Adaptive testing rests on item response theory, which describes each question by parameters such as its difficulty, or the level of the trait at which it best separates respondents, and its discrimination, or how sharply it separates them (Lord, 1980; Embretson & Reise, 2000). Because all items in a calibrated **item bank** are placed on the same scale, a respondent's level can be estimated from any subset of items, and scores from different subsets can be compared.

The adaptive procedure is a loop, shown in Figure 1. The questionnaire starts with an initial estimate of the respondent's level, often the population average. It then chooses the item that would give the most information at that estimate. After the respondent answers, the estimate is updated, and the next item is chosen for the new estimate. The loop continues until a stopping rule is met, typically when the standard error of the estimate falls below a set value or a maximum number of items has been asked (Weiss & Kingsbury, 1984; Thompson & Weiss, 2011).

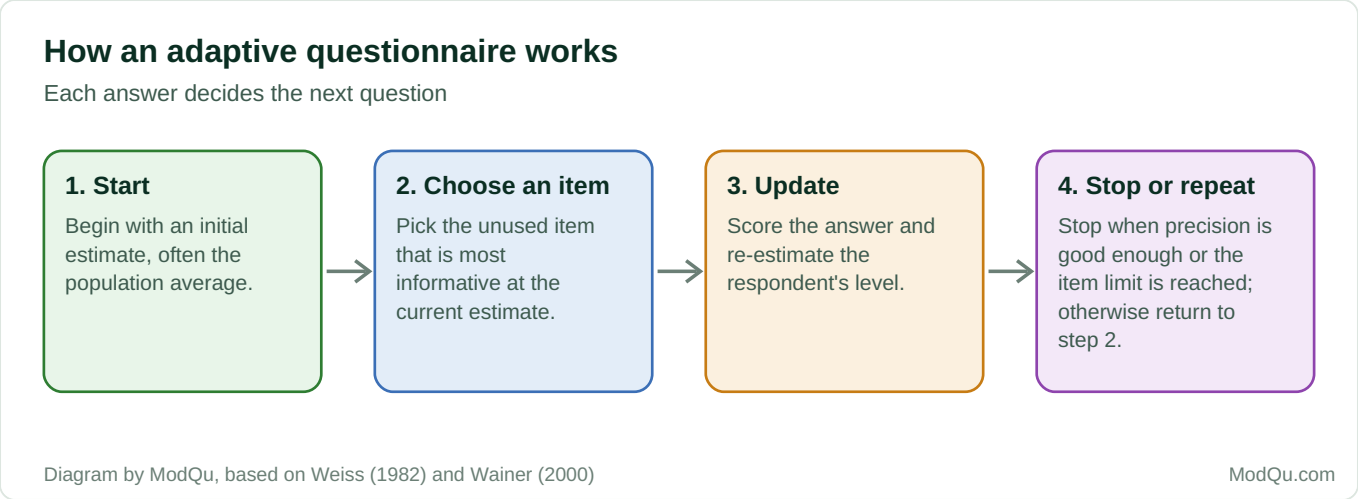



Figure 1. The adaptive questionnaire loop. Diagram by ModQu, based on Weiss (1982) and Wainer (2000).

This approach changes some familiar rules of measurement. In classical test theory, longer tests are more reliable and everyone must take the same items for scores to be comparable. In adaptive testing, a short tailored test can be more precise than a long fixed one, and different respondents can answer different items yet receive comparable scores (Embretson, 1996).

## Applications in surveys

### Health and quality of life

The most extensive use of adaptive methods in survey-style measurement is in health. The Patient-Reported Outcomes Measurement Information System (PROMIS) built large calibrated item banks for outcomes such as fatigue, pain and physical function, and offers adaptive versions that typically use a small number of items per domain (Cella et al., 2010). An early adaptive test for depression showed that a small number of adaptively chosen items could reach the precision of much longer fixed questionnaires (Fliege et al., 2005). In mental health assessment, an adaptive test drawn from a bank of over 600 items achieved high agreement with the full bank while asking each person only a small fraction of the items (Gibbons et al., 2008).

### Public opinion

Adaptive methods have also been applied to attitudes. Montgomery and Cutler (2013) showed how computerised adaptive testing could be used in public opinion surveys to measure concepts such as political knowledge or ideology with fewer items, and found that adaptive versions produced more accurate measurement than fixed short scales of the same length.

### Market research

A related idea is used in market research. Adaptive conjoint methods choose the next set of product profiles to show a respondent based on earlier choices, so that each respondent's preferences can be estimated with fewer questions (Toubia et al., 2003). The underlying models are different from item response theory, but the principle of choosing the most informative next question is the same.

## What adaptive testing requires

Adaptive testing is not something that can be added to an ordinary questionnaire at short notice. It requires the following:

- **A calibrated item bank.** Many items measuring one trait must be tested on a large sample and their parameters estimated. The bank should cover the full range of the trait, with items that work well at low, middle and high levels (van der Linden & Glas, 2010).
- **Evidence that one trait is measured.** Most adaptive tests assume the items measure a single dimension. If they do not, scores can be misleading.
- **Software.** The survey platform must estimate the respondent's level and select items in real time. Many standard survey tools cannot do this without custom programming.
- **Rules for item selection and exposure.** Without controls, the most informative items are used again and again, which can be a problem when items might become known or tiresome to respondents. Methods exist to spread use across the bank (Chang & Ying, 1999).
- **Stopping rules.** The designer must decide what precision is good enough and how many items respondents can be asked.

## Benefits and limitations

The main benefit is efficiency. Adaptive questionnaires can reduce the number of items needed for a given precision by half or more, which reduces respondent burden and frees time for other questions (Gibbons et al., 2008; Montgomery & Cutler, 2013). Precision is also more even across respondents, whereas fixed short forms tend to measure well only in the middle of the range.

There are limitations. Building and maintaining an item bank is expensive. Every respondent answers a different set of questions, which can complicate comparisons with older surveys that used fixed questions, and makes it harder to report results for individual items. The approach suits traits measured by many interchangeable items but not questionnaires made of distinct questions on different topics. Finally, respondents may notice that the questions vary and wonder why, so the purpose may need explaining.

## An example

An organisation measures employee well-being every quarter with a 30-item scale, and staff complain about its length. The organisation has several years of data from the full scale, collected from thousands of employees. Its analysts use these data to estimate item parameters, check that the items measure one dimension and remove items that behave

differently across departments. They then programme an adaptive version that starts every employee on a middle-level item, chooses each following item to be most informative at the current estimate and stops once the standard error reaches a set value or eight items have been asked.

In a trial in which a random half of staff receive the adaptive version and the other half the full scale, the adaptive version asks an average of six items. Its scores correlate highly with full-scale scores, and department averages are almost identical. Employees with very high or very low well-being receive items suited to their level, so their scores are measured more precisely than with a fixed short form. The organisation adopts the adaptive version and keeps the full scale for an annual check of the item parameters (illustrative example).

## Lessons for conventional questionnaires

Even when a full adaptive test is not feasible, its principles can improve ordinary questionnaires. Simple routing that skips questions irrelevant to a respondent, such as detailed product questions for non-users, applies the same logic of asking only informative questions. Two-stage designs, where a short screener determines which set of follow-up questions a respondent receives, are a simple form of adaptation. And the idea of item information encourages designers to ask whether each question in a scale adds anything for the people who will answer it (Embretson & Reise, 2000).

## Conclusion

Adaptive testing applies a simple idea: ask each respondent the questions that tell you the most about them. Built on item response theory and calibrated item banks, it can measure traits and attitudes as precisely as long questionnaires with far fewer questions. It has transformed measurement in health and is being explored in public opinion and market research. It demands investment in item banks, software and expertise, but for organisations that measure the same constructs repeatedly, that investment can pay back in lower burden and better data.

## References

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., … Hays, R. (2010). The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology, 63*(11), 1179–1194. https://doi.org/10.1016/j.jclinepi.2010.04.011

Chang, H.-H., & Ying, Z. (1999). *a*-Stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211–222. https://doi.org/10.1177/01466219922031338

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Embretson, S. E. (1996). The new rules of measurement. *Psychological Assessment, 8*(4), 341–349. https://doi.org/10.1037/1040-3590.8.4.341

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*(10), 2277–2291. https://doi.org/10.1007/s11136-005-6651-9

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D. K., Stover, A., Bock, R. D., & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59*(4), 361–368. https://doi.org/10.1176/ps.2008.59.4.361

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Montgomery, J. M., & Cutler, J. (2013). Computerized adaptive testing for public opinion surveys. *Political Analysis, 21*(2), 172–192. https://doi.org/10.1093/pan/mps057

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research, and Evaluation, 16*, Article 1.

Toubia, O., Simester, D. I., Hauser, J. R., & Dahan, E. (2003). Fast polyhedral adaptive conjoint estimation. *Marketing Science, 22*(3), 273–303. https://doi.org/10.1287/mksc.22.3.273.17743

van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer. https://doi.org/10.1007/978-0-387-85461-8

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement, 21*(4), 361–375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473–492. https://doi.org/10.1177/014662168200600408