INVESTIGATION · MARKET AND CONSUMER INVESTIGATION

Mystery Shopping as Investigative Research

Mystery shopping sends trained observers to experience a service as ordinary customers would. This article explains how the method works, what it can and cannot establish, how it has been used to investigate discrimination and service quality, and the design and ethical standards that make its findings credible.

ModQu research article · 2,900 words · 25 references (APA 7th edition)

Most research on customer experience asks customers what happened. Mystery shopping finds out directly. A trained observer, posing as an ordinary customer, visits a shop, telephones a call centre, applies for a loan, books a doctor's appointment or uses a website, and records what happens against a set of defined criteria. Because the staff being observed do not know they are being assessed, the method can capture what actually happens in ordinary service encounters, not what staff do when they know they are being watched or what customers remember afterwards.

The method is used widely in retail, hospitality, banking and the public sector to measure service standards. In a more rigorous form, known as audit or correspondence testing, it has become one of the most powerful tools for investigating discrimination. In health research, unannounced "standardised patients" have revealed large gaps between what clinicians know and what they do. This article explains how mystery shopping works as a form of investigation, what it can establish, how to design studies that produce reliable evidence, and the ethical questions raised by research that relies on deception.

What mystery shopping is

Wilson (1998, 2001) describes mystery shopping as a form of participant observation in which researchers act as customers to collect information about the service experience. The method has three defining features. The observer experiences the service directly, rather than relying on others' reports. The observer follows a planned scenario, so that each visit tests the same things. And the people providing the service do not know that the customer is an observer, at least not at the time.

These features make mystery shopping an investigative method rather than simply a survey technique. It tests claims, such as "our staff offer every customer a product demonstration" or "our advisers explain fees before a customer signs", by observing whether the claimed behaviour actually occurs. It can also reveal behaviour that organisations do not know about or would prefer not to disclose, such as misleading sales practices or unequal treatment of customers.

Uses of mystery shopping

Measuring service standards

The most common use is to check whether staff follow service standards. A retailer might specify that staff greet customers within a set time, ask about their needs, offer relevant products and explain the returns policy. Mystery shoppers visit stores and record whether each step occurred. The results are used to compare branches, identify training needs and reward good performance. Service quality research provides frameworks for deciding what to measure, such as reliability, responsiveness, assurance, empathy and the physical setting (Parasuraman et al., 1988).

Regulatory investigation

Regulators use mystery shopping to investigate whether firms comply with consumer protection rules. A financial regulator might send shoppers to banks to see whether advisers assess customers' needs before recommending products, or whether they disclose the risks of an investment. Consumer agencies use test purchases to check whether shops sell age-restricted products, such as alcohol or tobacco, to minors. In these settings, mystery shopping produces evidence that can support enforcement action, which raises the standard of rigour required.

Testing for discrimination

Audit studies apply the mystery shopping method to the question of whether people are treated differently because of characteristics such as race, sex, age or disability. Pairs of testers who are matched on every relevant characteristic except the one being tested apply for the same job, enquire about the same apartment or negotiate for the same car. Differences in how they are treated provide direct evidence of discrimination.

Housing audits in the United States, in which matched pairs of Black and white testers enquired about advertised housing, provided evidence that minority home seekers were shown fewer properties and given less information (Yinger, 1986). Ayres and Siegelman (1995) sent trained testers to negotiate for new cars at Chicago dealerships using an identical bargaining script, and found that Black and female testers received significantly higher final offers than white men. Pager (2003) used matched pairs of job applicants, one of whom reported a criminal record, and found large effects of both a record and race on callbacks.

Correspondence studies use written applications instead of in-person testers. Bertrand and Mullainathan (2004) sent fictitious résumés in response to job advertisements in Boston and Chicago, randomly assigning names that were typical of white or Black Americans. Résumés with white-sounding names received about 50 per cent more callbacks. A meta-analysis of field experiments on hiring in the United States found no change in the level of discrimination against Black applicants over the twenty-five years studied (Quillian et al., 2017). Reviews of the method set out its strengths and limits in detail (Riach & Rich, 2002; Bertrand & Duflo, 2017; Neumark, 2018).

Health care

In health research, unannounced standardised patients, who are trained actors presenting with scripted symptoms, have been used to measure the quality of clinical care. A systematic review found that the method was feasible and valid for assessing actual practice (Rethans et al., 2007). Das et al. (2012) sent standardised patients to health providers in India and found that providers spent little time with patients, asked few of the recommended questions and often gave incorrect treatment. Mystery shopper studies have also been used to measure access to care: Rhodes et al. (2014) had callers posing as new patients telephone primary care practices in the United States and found that appointment availability differed sharply by insurance type. Rhodes (2011) discusses the value of such studies and the ethical questions they raise.

What mystery shopping can and cannot establish

Mystery shopping has distinctive strengths. It measures behaviour rather than recall or opinion. It controls the customer side of the encounter, so that differences in outcomes can be attributed to the service provider. And it can detect practices that would never appear in surveys or official records.

Mystery shopping compared with other ways of measuring service

Each method answers a different question

	What it measures	Main strength	Main limitation
Customer survey	Customers' perceptions and satisfaction	Large samples of real customers	Recall and opinion, not behaviour
Mystery shopping	What staff do in a defined scenario	Direct, standardised observation	Few visits per site; observer judgment
Audit or correspondence test	Differential treatment of matched customers	Strong evidence of discrimination	Covers only the first stages of a process
Transaction data and recordings	What happened in all transactions	Complete coverage	Shows outcomes, not always the reasons

Based on Wilson (2001), Riach & Rich (2002) and Neumark (2018) ModQu.com

Figure 1. Mystery shopping is strongest when it is used to observe specific behaviour in a standardised way, and is usually best combined with other evidence. Diagram by ModQu.

The method also has clear limits.

Small samples per site. Most programmes make only a few visits to each branch or employee per period. A single visit captures one encounter with one member of staff at one time, and may not represent typical service. Conclusions about individual branches or staff based on one or two visits are unreliable.

Observer accuracy. Mystery shoppers must remember and record what happened, often after the visit, since taking notes during it would reveal their role. Research on the accuracy of mystery customers has found that their reports are affected by memory limitations and by the way questions are framed (Morrison et al., 1997). Finn and Kayandé (1999), analysing the reliability of mystery shopping scores, found that a substantial share of the variation in scores reflected differences between shoppers and visits rather than real differences in service, and that many visits per site would be needed to compare outlets reliably.

Artificial scenarios. The shopper follows a script, which may not reflect what real customers ask or how they behave. A scenario that is unusual or poorly performed may itself change the service received.

Limited scope of audit studies. Correspondence studies measure callbacks or responses to initial enquiries, not final outcomes such as job offers, wages or loan terms. They show whether discrimination occurs at the first stage of a process, which is important but not the whole picture (Neumark, 2018).

Designing a credible study

Define what is being tested

As in any investigation, the first step is a clear statement of the questions. "How good is our service?" is too vague to test. "Do advisers ask about the customer's existing borrowing before recommending a loan?" is specific and observable. Each item on the evaluation form should describe a behaviour that a trained observer can reliably judge as present or absent.

Design the scenarios

Scenarios should be realistic, consistent with what genuine customers do, and standardised so that every visit tests the same thing. In discrimination testing, testers in each pair must be matched on everything relevant except the characteristic being studied, and the order in which they visit should be varied. Correspondence studies typically randomise the assignment of characteristics, such as names, to otherwise equivalent applications.

Train and check the observers

Observers must be trained to follow the scenario, to remember and record details accurately, and to avoid behaviour that would reveal their role or influence the outcome. Their reports should be completed immediately after each visit. Where possible, accuracy should be checked by comparing reports with audio recordings, receipts or other evidence, and by having more than one observer assess the same encounter. Agreement between observers can be measured with statistics such as Cohen's (1960) kappa, which corrects for agreement expected by chance.

Sample enough visits

The number of visits should match the conclusions to be drawn. Comparing regions or chains requires fewer visits per site than comparing individual branches or staff. Results should be reported with an indication of their uncertainty, and decisions about individual staff should not rest on a single visit.

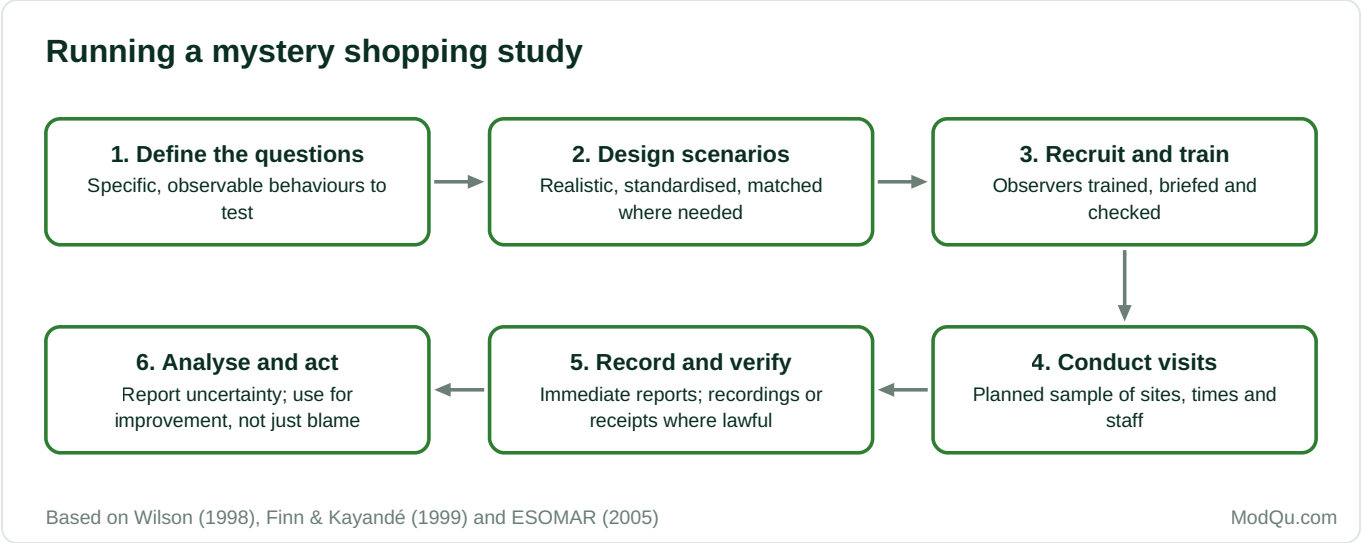


Figure 2. The stages of a mystery shopping study. Reliability depends most on clear questions, trained observers and enough visits for the conclusions drawn. Diagram by ModQu.

Report the uncertainty

Mystery shopping results are often reported as simple percentages, such as "70 per cent of visits met the standard", without any indication of how precise the figure is. With small numbers of visits, the uncertainty is large. If 14 of 20 visits to a branch met a standard, the observed rate is 70 per cent, but a 95 per cent confidence interval calculated by the Wilson method runs from about 48 to 85 per cent. With 140 of 200 visits, the same 70 per cent would have an interval of about 63 to 76 per cent. Agresti and Coull (1998) showed that intervals of this kind perform better for proportions than the simple textbook formula, especially with small samples. Reporting intervals, or at least the number of visits behind each figure, prevents readers from treating small differences between branches as meaningful when they may reflect chance.

Analyse fairly

Results should be analysed in ways that respect their limits. Averages across many visits are more informative than individual scores. Patterns across sites or times are more reliable than single observations. Where the study tests for discrimination, statistical analysis should account for the paired design and for differences between testers.

Digital and remote mystery shopping

Much service now takes place by telephone, chat, email and websites, and mystery shopping has followed. Telephone and online scenarios have practical advantages: they can be recorded where the law allows, which removes many problems of observer memory, and they can be run at larger scale and lower cost than visits. Regulators and researchers have used telephone mystery shopping to test what callers are told about products, prices and appointments, as in studies of access to medical care (Rhodes et al., 2014).

Online scenarios can test whether websites display required information, how easy it is to cancel a subscription, whether prices change during the purchase process, and how customer service responds to written complaints. Accessibility testing applies the same logic to disabled customers, with testers who use screen readers or other assistive technology recording whether they can complete common tasks. Correspondence studies have long used written applications, and the growth of online hiring and housing platforms has made such tests easier to run at scale (Neumark, 2018).

Digital methods raise their own design questions. Automated testing tools can visit websites thousands of times, but they may not experience the service as a person would, and platforms may detect and treat them differently. Online scenarios still need careful scripts, realistic customer profiles and checks that the observations reflect what real customers see.

A short example

The following hypothetical case shows how a regulator might use the method. A financial regulator receives complaints that some firms selling retirement income products recommend high-fee products without asking about customers' circumstances. It designs a scenario in which a shopper aged sixty-four, with a modest pension pot and a small mortgage, asks for advice on what to do with the pension. The scenario specifies the shopper's background in detail and lists the questions a competent adviser should ask under the regulator's rules.

Trained shoppers, recruited to fit the profile, make two hundred visits and calls to forty firms. Meetings are recorded where the law and the firms' own practices allow, and shoppers complete structured reports immediately afterwards. Analysts check a sample of reports against the recordings to confirm their accuracy. The study finds that in a large share of encounters advisers recommended a specific product before asking about the shopper's other debts and income needs, and that the problem was concentrated in a small number of firms with sales incentives tied to particular products.

The regulator publishes the aggregate findings without naming firms, writes to the firms with the worst results requiring them to review past sales, and changes its rules on adviser incentives. Because the evidence came from standardised scenarios, firms could not argue that the customers in the study had been unusual, and because the reports were checked against recordings, they could not easily dispute what was said.

Ethical questions

Mystery shopping relies on deception: the staff who are observed believe they are serving a real customer. Professional codes for market research generally accept this, provided that certain conditions are met. Guidance from research industry bodies says that staff should normally be told in advance that mystery shopping may take place, even if they are not told when; that the method should not be used to gather information for disciplinary action against individuals unless staff have been informed of this possibility; that the time taken from staff should be minimised and, where appropriate, paid for; and that personal data should be handled in accordance with data protection law (ESOMAR, 2005; Market Research Society, 2023; Regulation (EU) 2016/679, 2016).

Discrimination audits raise further questions, since the businesses tested have not consented and correspondence studies consume the time of employers who respond to fictitious applications. Researchers generally justify these costs by the importance of the information and the lack of any other way to obtain it, while keeping the burden on participants as small as possible. Kimmel (2007) and Kimmel et al. (2011) discuss the conditions under which deception in research can be justified, including the absence of alternative methods, minimal harm and appropriate debriefing.

In health care, unannounced standardised patients use clinicians' time and may affect care for other patients. Rhodes (2011) argues that such studies can be justified when they address important questions about access or quality that cannot be answered in other ways, and that ethics review and careful design are essential.

There is also the question of reactivity. If staff know that mystery shopping takes place, they may behave differently with any customer they suspect of being an observer. Research on the Hawthorne effect, the idea that people change their behaviour when they know they are being studied, finds that such effects exist but vary greatly in size and are not well understood (McCambridge et al., 2014). Programmes that are well known within an organisation may measure staff's ability to spot shoppers as much as their normal behaviour.

Combining mystery shopping with other evidence

Mystery shopping is most informative when its findings are read alongside other sources. Customer surveys show how customers feel about the service, while mystery shopping shows whether specific behaviours occurred; when the two disagree, the disagreement is itself useful. A branch with high satisfaction but poor mystery shopping scores may have loyal customers who forgive lapses, or scenarios that do not reflect what its customers actually need. A branch with good scores but poor satisfaction may follow procedures that customers do not value.

Operational data adds a third view. Sales records, complaint logs, call recordings and transaction data cover every customer rather than a sample of visits, and can show whether the patterns found by shoppers appear in ordinary business. If shoppers find that advisers rarely explain fees, and complaint data show a cluster of complaints about unexpected charges in the same branches, the two sources reinforce each other. If the sources point in different directions, the investigator should ask which is more likely to be wrong and why.

This kind of triangulation follows the general principle that conclusions supported by several independent methods are more reliable than conclusions resting on one. It also guards against the particular weaknesses of each method: the small samples and observer error of mystery shopping, the recall and opinion of surveys, and the missing context of transaction data.

Using the findings

Mystery shopping findings are most useful when they lead to improvement rather than blame. An observed failure to follow a procedure may reflect inadequate training, unrealistic workload, unclear standards or incentives that reward other behaviour. Organisations that use mystery shopping only to discipline staff tend to produce defensive behaviour and staff who look for shoppers rather than serve customers. Those that use it to identify where systems and training fall short get more lasting change.

For investigators, mystery shopping is a reminder that the most direct way to find out what happens is often to observe it. Its findings are strongest when the questions are specific, the scenarios are realistic, the observers are trained and checked, the sample is large enough, and the conclusions stay within what the evidence can support.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52*(2), 119–126. https://doi.org/10.1080/00031305.1998.10480550

Ayres, I., & Siegelman, P. (1995). Race and gender discrimination in bargaining for a new car. *The American Economic Review, 85*(3), 304–321.

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments* (Vol. 1, pp. 309–393). North-Holland. https://doi.org/10.1016/bs.hefe.2016.08.004

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Das, J., Holla, A., Das, V., Mohanan, M., Tabak, D., & Chan, B. (2012). In urban and rural India, a standardized patient study showed low levels of provider training and huge quality gaps. *Health Affairs*, *31*(12), 2774–2784. https://doi.org/10.1377/hlthaff.2011.1356

ESOMAR. (2005). *ESOMAR world research codes and guidelines: Mystery shopping studies*.

Finn, A., & Kayandé, U. (1999). Unmasking a phantom: A psychometric assessment of mystery shopping. *Journal of Retailing, 75*(2), 195–217. https://doi.org/10.1016/S0022-4359(99)00004-4

Kimmel, A. J. (2007). *Ethical issues in behavioral research: Basic and applied perspectives* (2nd ed.). Blackwell.

Kimmel, A. J., Smith, N. C., & Klein, J. G. (2011). Ethical decision making and research deception in the behavioral sciences: An application of social contract theory. *Ethics & Behavior, 21*(3), 222–251. https://doi.org/10.1080/10508422.2011.570166

Market Research Society. (2023). *MRS code of conduct*.

McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology, 67*(3), 267–277. https://doi.org/10.1016/j.jclinepi.2013.08.015

Morrison, L. J., Colman, A. M., & Preston, C. C. (1997). Mystery customer research: Cognitive processes affecting accuracy. *Journal of the Market Research Society, 39*(2), 349–361.

Neumark, D. (2018). Experimental research on labor market discrimination. *Journal of Economic Literature, 56*(3), 799–866. https://doi.org/10.1257/jel.20161309

Pager, D. (2003). The mark of a criminal record. *American Journal of Sociology, 108*(5), 937–975. https://doi.org/10.1086/374403

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–40.

Quillian, L., Pager, D., Hexel, O., & Midtbøen, A. H. (2017). Meta-analysis of field experiments shows no change in racial discrimination in hiring over time. *Proceedings of the National Academy of Sciences, 114*(41), 10870–10875. https://doi.org/10.1073/pnas.1706255114

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). (2016). *Official Journal of the European Union, L 119*, 1–88.

Rethans, J.-J., Gorter, S., Bokken, L., & Morrison, L. (2007). Unannounced standardised patients in real practice: A systematic literature review. *Medical Education, 41*(6), 537–549. https://doi.org/10.1111/j.1365-2929.2006.02689.x

Rhodes, K. V., Kenney, G. M., Friedman, A. B., Saloner, B., Lawson, C. C., Chearo, D., Wissoker, D., & Polsky, D. (2014). Primary care access for new patients on the eve of health care reform. *JAMA Internal Medicine, 174*(6), 861–869. https://doi.org/10.1001/jamainternmed.2014.20

Rhodes, K. V. (2011). Taking the mystery out of "mystery shopper" studies. *New England Journal of Medicine, 365*(6), 484–486. https://doi.org/10.1056/NEJMp1107779

Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal, 112*(483), F480–F518. https://doi.org/10.1111/1468-0297.00080

Wilson, A. M. (2001). Mystery shopping: Using deception to measure service performance. *Psychology & Marketing, 18*(7), 721–734. https://doi.org/10.1002/mar.1027

Wilson, A. M. (1998). The use of mystery shopping in the measurement of service delivery. *The Service Industries Journal, 18*(3), 148–163. https://doi.org/10.1080/02642069800000037

Yinger, J. (1986). Measuring racial discrimination with fair housing audits: Caught in the act. *The American Economic Review, 76*(5), 881–893.