The Canadian Journal of Program Evaluation Vol. 16 No. 2 Pages 41–53
ISSN 0834-1516

DO DISTINCT SERVQUAL DIMENSIONS EMERGE FROM MYSTERY SHOPPING DATA? A TEST OF CONVERGENT VALIDITY

Michelle Lowndes
John Dawes
University of South Australia
Adelaide, Australia

Abstract: Service quality is commonly thought to encompass five generic dimensions: responsiveness, assurance, tangibles, empathy, and reliability. These dimensions form the basis for service measurement tools such as SERVQUAL. Research in this area using tools such as SERVQUAL has predominantly focused on customer perceptions of quality. However, another approach used by many organizations is to send trained raters into the service environment, posing as customers, to evaluate service levels. This approach is often called "mystery shopping" and is very commonly used in both private- and public-sector organizations. This study examines whether the accepted service quality dimensions derived from customer perceptions studies are reflected in service quality evaluations using mystery shopping. It finds that the dimensions that emerge from mystery shopping data resemble SERVQUAL dimensions. Furthermore, a replication found that those dimensions are reasonably stable over time. The findings suggest that data from mystery shopping surveys can exhibit convergent validity.

Résumé: On pense généralement que la qualité du service se mesure en fonction de cinq paramètres: adaptation, assurance, éléments tangibles, compassion et fiabilité. Les instruments de mesure du service, comme SERVQUAL se fondent sur ces dimensions. La recherche en ce domaine, utilisant des instruments comme SERVQUAL, a porté principalement sur les perceptions de la qualité chez les consommateurs. Cependant, une autre approche utilisée par de nombreuses organisations consiste à envoyer sur place des inspecteurs qualifiés qui se font passer pour des clients afin d'évaluer le service. L'approche du «client mystère», souvent appelée «évaluation mystère», est très commune dans les secteurs privé et public. Cet article vise à déterminer si les paramètres acceptés de la qualité du service découlant des études menées auprès des consommateurs sont prises en compte dans

les évaluations mystères. La conclusion est que les paramètres qui ressortent des données concernant l'évaluation mystère ressemblent à celles de SERVQUAL. De plus, une reprise de l'étude a confirmé que ces paramètres sont raisonnablement stables. Les résultats suggèrent que les données provenant des enquêtes sur l'évaluation mystère peuvent avoir une validité convergente.

Customer service is an important issue for all types of organizations. For business organizations in a competitive environment, it is obviously necessary to pay attention to the level of service provided. Customers are free to choose alternatives, and if they perceive that they will get better service at a competitor, they may well transfer some of their custom.

Customer service is also increasingly important for public-sector organizations. Governments today are more willing to deregulate markets and so allow privately owned companies to compete against former government monopolies. Service levels become vital to the former monopoly if it is to retain its customers. Even where a government organization does have a monopoly granted by the government, it may be held accountable to specified customer service levels. Alternatively, key performance indicators for public organizations may be published and available for the scrutiny of customers or other stakeholders (the study of PennDOT is a case in point; see Poister & Harris, 2000). There are thus few organizations that can afford not to consider how well they are providing service to their customers.

Service quality and customer satisfaction are also very significant research streams in the academic literature. Many research studies have been conducted to identify supposedly generic aspects or components of service quality (e.g., Parasuraman, Zeithaml, & Berry, 1985, 1988) and the links between service quality, satisfaction, and customer buying behaviour or its proxies (for a recent review, see Zeithaml, 2000). Much effort has been expended to develop questionnaire instruments measuring service quality levels as perceived by customers.

For managers, the advances in service quality research have been useful. Managers can use service quality measures to benchmark and track customer service, and obtain reports that identify stronger or weaker points in service delivery. For example, a manager might find that the organization is strong on responsiveness but weaker in terms of empathy. For all this usefulness, it should be said that there are some limitations to the use of customer perceptions of serv-

ice quality. First, surveys of customers rely on the respondent's memory of previous service encounters. The customer may have a less than perfect recollection of recent encounters, and so the data may be biased. Second, customer perceptions are by definition subjective. Even when two customers experience the same level of service, one might be pleased and the other displeased. It can be argued that what matters in the end is what the customer thinks, but the preceding point highlights that there may also be value in obtaining more impartial evaluations of the service that is offered. Third, customers may not think to a great degree about the service they received, decreasing their ability to supply accurate responses that mirror the service levels offered.

These limitations suggest that other methods of ascertaining the adequacy of customer service levels would be desirable. One supplementary method is called "mystery shopping." Mystery shopping is used for a variety of purposes, including monitoring of an organization's own customer service and the levels of service offered by competitors, and evaluating if and how resellers are promoting a supplier's goods (Wilson, 1998a, 1998b). Mystery shopping is widely used in both the private and public sector. Even institutions such as hospitals use mystery shopping, sending into the facility researchers posing as patients requiring treatment (Millstead, 1999). Another example of the ubiquity of the technique is the use of mystery shopping to determine whether financial institutions engage in fair lending practices ("OTS to Test 'Mystery Shopping'," 2000).

The focus of research in this article is on the monitoring of the organization's own service quality. In that context, monitoring involves sending trained shoppers posing as customers into the service environment; immediately following the service encounter these shoppers rate various service aspects according to internally specified criteria. This process should ideally result in an "impartial" evaluation of a specific service encounter, although whether the process is reliable and valid are questions that have to date received little attention in the literature (see Dawes & Sharp, 2000).

As with the use of customer perceptions of service quality, there are also some fundamental issues to consider in using mystery shopping. First, those who use or commission mystery shopping would presumably like it to measure variables that are important to customers. It would be all too easy to measure things that the organization feels are important but to which customers are relatively

indifferent. Therefore, there is a good rationale for basing mystery shopping questionnaires or checklists as much as possible on the "SERVQUAL/SERVPERF" dimensions: tangibles, responsiveness, empathy, assurance, and reliability (Cronin & Taylor, 1994; Parasuraman et al., 1988).

Another issue relating to measurement is the composition of the questionnaire items. A basic tenet of measurement is that any measure is subject to some degree of error (Peter, 1979). Therefore, it is generally accepted that multiple measures of a construct are desirable so as to minimize the effect of error from any one measure (Churchill, 1979). To this end, multiple-item scales are commonly used; for example, four questions are used to measure the latent construct of tangibles in the SERVQUAL scale (Parasuraman et al., 1988). Mystery shopping questionnaires or checklists, if they are to gauge the extent of latent constructs such as responsiveness, should probably also use multiple measures.

There is also a fundamental difference between the conceptualization of service quality in terms of customer perceptions, and the process by which service quality is measured using mystery shopping. Service quality is conceptualized as an attitude formed from repeated service encounters, and so, according to the literature, it is not the same as an impression of one encounter. Mystery shopping, on the other hand, endeavours to obtain an evaluation of a single service encounter, ideally as unaffected as possible by previous encounters. Because mystery shopping focuses on one encounter, it is plainly not the same as service quality viewed as an attitude developed over time. However, if a mystery shopping survey is undertaken using the same sort of measurement instruments as are used to tap customer perceptions of service quality, will the same underlying dimensions that comprise customer-perceived service quality emerge? This becomes a question of convergent validity. Convergent validity is a desirable psychometric property of measurement instruments such as questionnaires, which refers to the extent that items designed to tap a particular construct actually "load together" or converge on that construct:

> If a construct were hypothesised to have 3 dimensions, a factor analysis of a purported measure which produces 3 meaningful factors could be interpreted as supportive evidence for convergent validity. (Peter, 1979)

If an analysis found that the same service quality dimensions do emerge, this research would produce positive evidence for the convergent validity of the mystery shopping process. If not, it would suggest that the elements of service quality that tend to closely covary according to customer perceptions do not do so when mystery shopping is used. This may have important implications for the commercial use of mystery shopping, and also the interpretation of surveys based on customer perceptions of service quality. For instance, if previous research has shown that particular questionnaire items form a common factor when administered to customers but do not when used by mystery shoppers, this might suggest that the results from customer surveys are affected by halo bias. That is, respondent's scores for a particular question might be distorted by impressions of other aspects of the object or person (see Fisicaro & Vance, 1994). If mystery shopping data form a factor structure quite different from what would be expected from previous research using the SERVQUAL or SERVPERF instruments, the convergent validity of the mystery shopping process might be questioned. In either case, the results from this research are likely to be of interest to academic researchers, market research practitioners, and those who commission or are interested in monitoring their customer service levels.

The next section briefly describes the development of constructs and measures of service quality as perceived by customers. The constructs discussed form a basis for testing the convergent validity of mystery shopping.

SERVICE QUALITY

The development and accompanying debate surrounding the SERVQUAL and SERVPERF measures is well documented, and so will be only briefly recounted here.

In the early 1980s Parasuraman, Zeithaml, and Berry (1985) conducted an extensive program of qualitative research with businesspeople and consumers to explore the concept of service quality. They concluded that service quality as perceived by customers depends on the *gap* between their expectations and the level of service that was provided. They also identified ten components of service quality, including constructs such as competence, courtesy, credibility, and security. This list was later reduced to five constructs: tangibles, responsiveness, empathy, assurance, and reliability (Parasuraman et al., 1988), forming the basis of a scale called SERVQUAL. The structure of the constructs, the notion of a gap, and the linkage

between service quality and satisfaction have been vigorously debated in the literature since that time (e.g., Asubonteng, McCleary, & Swan, 1996; Chong, Kennedy, Riquier, & Rungie, 1997; Cronin & Taylor, 1992; Dabholkar, Thorpe, & Rentz, 1996). There has been a fairly convincing argument to suggest that customer-perceived service quality is based simply on "performance" rather than on a gap between performance and expectations, with the performance-only scale termed SERVPERF (Cronin & Taylor,1 992, 1994). However, there is at least *some* degree of agreement about the five generic factors that comprise service quality, although it is acknowledged that the precise composition on an instrument may depend on the industry under study.

RESEARCH OBJECTIVES AND METHODOLOGY

As stated earlier, the broad purpose of this work is to test the convergent validity of data gathered from the mystery shopping process. The method to be used is exploratory factor analysis. More specifically, the purpose is to assess whether the service quality dimensions that emerge from factor analysis on mystery shopping data based on the SERVPERF instrument are the same as those found in the customer perceptions literature. This research therefore acts as a validity and reliability check on the SERVPERF instrument as used in mystery shopping. Furthermore, the research sought to test the generalizability of the results by using two separate sets of data.

Questionnaire

As with many other replications adopting the SERVQUAL instrument, the attributes were slightly modified to fit the industry tested. The industry was retailing, and the organization in question was a government-owned enterprise (GOE), which nevertheless competes with private-sector organizations for many aspects of its product and service mix. Due to the nature of the mystery shopping program and the particular type of organization evaluated, some items that appear in the SERVQUAL scale were not included, specifically those pertaining to reliability, which was not assessable in this study. Measurement of reliability requires the shopper to determine how the organization performs relative to what it promises or undertakes to do (Parasuraman et al., 1988), and this was not possible from a single service encounter. Therefore, we would expect to find only four of the five SERVPERF/SERVQUAL dimensions: assurance, tangibles, empathy, and responsiveness.

Data Collection

Data were obtained from a commercially funded market research project. It comprised ratings for 85 retail outlets of a national Australian retail organization. Some of the outlets were mystery shopped on two occasions (by a different shopper each time). Interviewer Quality Control Australia (IQCA) trained and accredited interviewers were employed to pose as customers and, immediately following, to rate the aspects of service quality encountered on 11-point (zero to ten) scales. The interviewers were extensively briefed on the interpretation and use of the measurement scales. They were also provided with prepared "enquiry scenarios," to make their encounters similar to those of a normal customer. Data were collected in 1997 and again in 1998, with a total of 142 observations in each year. The 1998 data used a different team of mystery shoppers. The list of items used by the mystery shoppers is shown as Appendix 1.

Data Analysis

Exploratory factor analysis was used to determine the underlying factor structures. The choice of exploratory over confirmatory factor analysis was made on the basis that there is less than unanimous agreement on the precise dimensionality of service quality, and some inconsistency regarding exactly which question items belong to which service dimension (e.g., Buttle, 1996). Therefore, it would be rather difficult to decide in advance how to construct a confirmatory factor model.

Exploratory factor analysis also offers the analyst some control over the resultant factor structure. In this case, it was reasonable to assume that a four-factor solution would fit the data (as the variables were adapted from four of the SERVQUAL factors). Therefore, the procedure prespecified four factors. This was also appropriate on the basis that the sample size was modest, with 142 observations in each data set. As the data set comprised 18 variables, this suggested that a smaller rather than a larger number of factors would be desirable. The technique used was principal-axis with varimax rotation.

RESEARCH FINDINGS

The analysis revealed a four-factor solution, but also showed that two variables, pertaining to wait times, did not have a loading on

any factor of higher than 0.23. These two variables were therefore discarded, and the analysis was repeated. Table 1 summarizes the loadings on each factor. Factor loadings of approximately 0.50 and above are circled to facilitate the reader's interpretation of the factor structure. The left column indicates which aspect of service quality the particular item is meant to measure.

Table 1
Factor Solutions

Key*	Question	1997 data Factors 1	2	3	4	1998 data Factors 1	2	3	4
T	Physical facilities visually appealing	**0.82**	0.03	0.01	-0.03	**0.86**	0.07	-0.07	0.17
T	Modern-looking equipment and fixtures	**0.75**	0.08	-0.05	-0.01	**0.85**	0.01	-0.12	0.20
T	Presentation of merchandise excellent	**0.85**	0.24	0.05	-0.14	**0.91**	0.07	-0.10	0.18
T	Store layout made things easy to find	**0.58**	0.31	0.08	-0.01	**0.76**	0.16	-0.03	0.10
T	Easy for customers to move around in the store	**0.58**	0.30	0.10	0.02	**0.65**	0.13	-0.02	-0.03
T	Interior was neat and clean	0.34	**0.41**	0.16	-0.26	**0.73**	0.12	-0.07	0.27
T	Employees well dressed	0.31	**0.89**	0.12	-0.16	0.49	0.28	-0.14	**0.74**
T	Employees neat in appearance	0.26	**0.88**	0.11	-0.14	0.42	0.31	-0.14	**0.77**
E	Staff were polite	0.10	0.30	**0.50**	-0.03	0.17	**0.84**	-0.20	0.41
E	Person who served gave a pleasant smile	0.12	0.21	**0.67**	0.04	0.03	**0.46**	-0.04	0.17
E	Given a pleasant parting remark	0.16	0.25	**0.69**	-0.07	0.05	**0.80**	-0.25	0.21
A	Asked pertinent questions	0.07	-0.10	**0.63**	-0.13	0.17	**0.77**	-0.13	-0.05
A	Additional information	-0.10	-0.01	**0.50**	-0.10	0.16	**0.80**	-0.13	0.01
A	Did not know what my needs were	0.15	0.08	**-0.42**	0.31	-0.04	**-0.53**	0.27	-0.12
R	Not individual attention	-0.05	-0.20	-0.13	**0.84**	-0.11	-0.30	**0.77**	-0.15
R	Not personal attention	-0.10	-0.14	-0.17	**0.96**	-0.13	-0.33	**0.83**	-0.06

*T = Tangibles E = Empathy A = Assurance R = Responsiveness

The 1997 data show a four-factor solution, but the structure is somewhat different from the SERVQUAL/SERVPERF composition. Factor 1 in 1997 is comprised of tangibles variables: the physical outlet, presentation, and layout. Factor 2 is also a tangibles factor but pertains to the appearance of staff and the cleanliness of the outlet. Factor 3 is an assurance and empathy factor. "Assurance" refers to employee knowledge and their ability to instill trust and confidence, and "empathy" refers to individualized attention. Factor 4 is a responsiveness factor (willingness to help and offer prompt service).

There is considerable stability in the factor structure when the analyses for 1997 and 1998 are compared. An exception is the variable for *interior was neat and clean,* which loaded onto the staff & cleanliness factor 2 in 1997, but in 1998 loaded onto factor 1, which is the physical facilities factor. That aside, the analysis suggests that the underlying dimensions are relatively robust over time.

The findings give somewhat mixed support for the existence of SERVQUAL-like dimensions in the measurement of service quality using mystery shopping. In particular, there are two differences between the service quality dimensions as portrayed in the literature on customer perceptions, and the results found here. Empathy and assurance form a common factor in this data, whereas SERVQUAL views them as separate factors. Also, there are two tangibles factors: one to do mainly with the physical facilities, equipment, and layout, and one pertaining to appearance — cleanliness and smartness of the outlet and its staff.

Although the factor composition is somewhat different from SERVQUAL, the factor structures are quite clear, with generally small cross-loadings across factors. This suggests that when a series of statements is presumed to measure, say, tangibles, they do appear to load onto a common factor that is justifiably thought of as comprising tangibles. Furthermore, they tend to load to only a small extent with other factors that are presumed to measure other service quality aspects. Overall, the findings suggest that mystery shopping surveys can exhibit convergent validity. Furthermore, the resultant common service factors were relatively stable over two surveys.

CONCLUSIONS AND RECOMMENDATIONS

This article put forward a rationale as to why mystery shopping instruments should be based on SERVQUAL/SERVPERF dimensions, and

also why multiple questions should be used as indicators for each of the latent service quality constructs. The findings from the analysis suggest that the factor structure of service quality from mystery shopping has some resemblance to the SERVQUAL/SERVPERF dimensions, but that (a) empathy and assurance may form a common factor, and (b) there may be separate tangibles dimensions split between physical appeal and layout, as well as cleanliness and appearance of the outlet and staff.

The dimensions also appeared to be quite stable over time. A replication survey carried out at a later date using identical methodology produced very similar factors. The findings provide positive evidence for the validity of the mystery shopping process. However, the dimensions of service quality that are tapped using the mystery shopping process are not exactly the same as those that are tapped using customer perceptions. For those using, implementing, or commissioning mystery shopper surveys, there are several implications. First, one must at least consider the multidimensional nature of service quality, and include items designed to measure constructs such as tangibles, empathy, assurance, and responsiveness. It should also be recognized that tangibles appears to comprise the physical appearance of the retail setting (i.e., the more "permanent" aspects) and physical layout, as well as variables measuring appearance (cleanliness and neatness of the setting and staff). Second, multiple items should be used to tap these constructs. This study showed that multiple items designed to measure a latent construct can exhibit convergent validity. Third, waiting times may constitute a separate service quality factor that is distinct from the accepted construct of responsiveness. It may be that consumers perceive a commonality between the components of responsiveness (i.e., showing a willingness to serve) whereas mystery shoppers distinguish between them. For instance, if wait times are somewhat extended but the staff are trying as hard as they can to cope with the situation, a mystery shopper would presumably rate certain responsiveness items highly but waiting time items lower. Consumers might not be so adept at making this distinction, given that they are often interviewed some time after their last service experience with the organization under study.

LIMITATIONS

This research has been conducted only within one industry and only within Australia. These are obvious limitations to the generalizability of the results. There may be particular industry conditions

under which the factor structure does hold and others where it does not. The study should be replicated in other industries and countries to see if there are any boundary conditions under which the dimensions do not hold. The sample size was modest, with 142 observations, but was repeated and produced quite similar results, which suggests the sample size was not a particular issue. Another limitation, it has been suggested, is that SERVQUAL factor structures can be influenced by positive/negative question wording (Smith, 1995). That could be the case here, as some of the questions were negatively phrased. To test whether the data were influenced by question wording, this research could be replicated including all positively phrased items within the instrument. Last, there were two questions that measured wait times, and these questions did not load onto any of the four factors. Wait times would be expected to load with the responsiveness factor. Future research could investigate this further, perhaps with larger sample sizes. This would enable a factor analysis without a prespecified number of factors, and perhaps enable an identification of other factors within the data. It may be the case that wait times are a separate factor from other facets of responsiveness.

REFERENCES

Asubonteng, P., McCleary, K.J., & Swan, J.E. (1996). SERVQUAL revisited: A critical review of service quality. *Journal of Services Marketing, 10*(6), 62–81.

Buttle, F. (1996). SERVQUAL: Review, critique, research agenda. *European Journal of Marketing, 30*(1), 8–32.

Chong, E., Kennedy, R., Riquier, C., & Rungie, C. (1997). The difference between satisfaction and service quality. In D. Arnott, S. Bridgewater, S. Dibb, P. Doyle, J. Freeman, T. Melwar, V. Shaw, L. Simkin, P. Stern, R. Wensely, & V. Wong (Eds.), *26th European Marketing Academy Conference*, Vol. 1 (pp. 257–269). Warwick, U.K.: University of Warwick Business School.

Churchill, G.A., Jr. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16* (February), 64–73.

Cronin, J.J., Jr., & Taylor, S.A. (1992). Measuring service quality: A reexamination and extension. *Journal of Marketing, 56* (July), 55–66.

Cronin, J.J., Jr., & Taylor, S.A. (1994). SERVPERF versus SERVQUAL: Reconciling performance-based and perceptions-minus-expectations measurement of service quality. *Journal of Marketing, 58* (January), 125–131.

Dabholkar, P.A., Thorpe, D.I., & Rentz, J.O. (1996). A measure of service quality for retail stores: Scale development and validation. *Journal of the Academy of Marketing Science, 24,* 3–16.

Dawes, J.G., & Sharp, B. (2000). The reliability and validity of objective measures of customer service: "Mystery shopping." *Australasian Journal of Market Research, 8*(1), 29–46.

Fisicaro, S.A., & Vance, R.J. (1994). Comments on the measurement of halo. *Educational and Psychological Measurement, 54*(2), 366–371.

Millstead, J.B. (1999, May/June). Mystery shopping in your organization. *Healthcare Executive*, p. 66.

OTS to test "mystery shopping" to check fair lending compliance. (2000, September). *ABA Bank Compliance*, p. 8.

Parasuraman, A., Zeithaml, V.A., & Berry, L.L. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49*(Fall), 41–50.

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*(Spring), 12–40.

Peter, J.P. (1979). Reliability: A review of psychometric basics and recent marketing practices. *Journal of Marketing Research, 16*(February), 6–17.

Poister, T.H., & Harris, R.H., Jr. (2000). Building quality improvement over the long run. *Public Performance & Management Review, 24*(2), 161–176.

Smith, A.M. (1995). Measuring service quality: Is SERVQUAL now redundant? *Journal of Marketing Management, 11*, 257–276.

Wilson, A.M. (1998a). The role of mystery shopping in the measurement of service performance. *Managing Service Quality, 8,* 414–420.

Wilson, A.M. (1998b). The use of mystery shopping in the measurement of service delivery. In G. Hogg & M. Gabbott (Eds.), *Service industries marketing: New approaches* (pp. 148–163). London: Frank Cass Publishers.

Zeithaml, V.A. (2000). Service quality, profitability, and the economic worth of customers: What we know and what we need to learn. *Journal of the Academy of Marketing Science, 28*(1), 67–85.

Appendix 1
Questionnaire Items

The physical facilities of this outlet are appealing.
This outlet has modern-looking equipment and fixtures.
The interior of the outlet was neat and clean.
The presentation of merchandise was excellent.
The employees were well dressed.
The employees were neat in appearance.
I had to wait a long time to be served. ***(deleted from the analysis)***
Staff were too busy to respond to customer requests. ***(deleted from the analysis)***
Staff were polite.
The person who served me gave a pleasant smile.
I was given a pleasant parting remark.
I did not receive individual attention. ***(reverse coded)***
I did not receive personal attention. ***(reverse coded)***
The employee who served me asked me questions pertinent to my enquiry.
The employee who served me provided me with additional information.
The person who served me did not know what my needs were. ***(reverse coded)***
The store layout made it easy to find things.
The store layout made it easy for customers to move around the store.