

The Reliability and Validity of Objective Measures of Customer Service: “Mystery Shopping”

Published in *Australian Journal of Market Research* January 2000

John Dawes
Byron Sharp
Marketing Science Centre
University of South Australia
North Terrace Adelaide
South Australia 5000
Australia

<http://www.marketing.unisa.edu.au/>

The Reliability and Validity of Objective Measures of Customer Service: “Mystery Shopping”

Abstract

The purpose of this paper is to examine the reliability and validity of an “objective” method of measuring customer service frequently utilised by some market research organisations. This method is often referred to as “mystery shopping”. Using four quarters of data from a large mystery shopping program involving over 200 outlets we examine inter-rater reliability, convergent validity and criterion validity. We then examine the stability of various service performance factors in predicting service quality across four different surveys over a twelve month period. Finally, we examine the validity of the process of objective service quality measurement at individual store level. We found that it is indeed possible to create a mystery shopping instrument that exhibits a high degree of reliability. We also found that mystery shopping scores show positive signs of validity, including that the various aspects of service performance showed a consistent relationship with overall service quality. However, we found that there is considerable variation in store scores between waves and since the potential sampling error involved in mystery shopping is high, it is impossible to determine whether this is real change or not. Sampling error aside, the degree of variation in service performance that we report throws serious doubt on the meaningfulness of reporting an average level of service quality for a store. These results have implications for the way that mystery shopping results are presented and interpreted by market research organisations. In particular, we conclude that mystery shopping surveys should seldom be used to assess changes in service performance at individual store level.

Keywords: Mystery shopping, customer service, reliability, validity

Introduction: Objective Service Quality Measurement

In this article we examine measurement issues in mystery shopping research, which is used to appraise the quality of stores' service delivery.

The importance of providing acceptable levels of customer service has doubtless been appreciated by businesspeople for as long as trading has occurred. However, in recent years the academic, and popular management and marketing press has witnessed a surge in interest on the topic (see Rust, Zahorik, and Keiningham 1995) in line with the tremendous growth in the service sector of the economies of developed countries. Much of the research focus has been on operationalising, testing and refining measures of customer perceptions of service quality. Pioneering work in this area was conducted by Parasuraman Zeithaml and Berry (1988) who developed a measurement tool based on an expectancy-disconfirmation paradigm. This approach has been heavily criticised on the basis that it confounds service quality with satisfaction (Cronin and Taylor 1994), among other criticisms. In response to this criticism a performance-based model (Cronin and Taylor 1994) has been proposed which is suggested to be a more appropriate operationalisation of service quality. Performance-based simply means a higher score is better and expectations are not considered or measured. In this study we utilise an approach based on a performance based model, specifically for retail sites which provide not only "service" but also physical "product" (Dabholkar, Thorpe, and Rentz 1996). However, since we use quite a different approach, using expert raters rather than consumer sentiment, we purposefully do not review the large body of literature on customer based service quality/satisfaction.

Our instrument creation follows the evidence presented in Buttle (1996) that instruments designed to measure service quality may require customisation for specific industries. This does not necessarily mean that results lack generalisability because as will be seen, the items used are what many people would accept as basic components of good service. In addition, our primary purpose was not to identify or validate service quality components *per se* but rather to demonstrate, through an examination of their statistical properties, the extent of reliability and validity of an approach to measuring objective service quality, namely mystery shopping¹.

Objective Measures

The difference between perceived and objective quality is an important one in the literature on product quality. To the scientist or technician, objective quality is something that can be measured by tests. It is a level of performance against some standard (Riesz 1980) which in itself is ultimately subjective (Maynes 1976), such as energy efficiency, amount of defects, even number of features or type of ingredients. Here we use the term "objective" in relation to service quality to mean that the resultant scores are relatively independent of the person providing the rating and the time the rating occurred (see Ehrenberg and Shewan 1953).

¹ Also known as *shadow shopping*, or *phantom shopping*.

In contrast to the product quality literature where there is a long history of utilisation of objective product quality and product feature ratings (e.g. Morris and Bronson 1969; Jacoby and Olson 1985; Kamakura and Russell 1993) there is a dearth of research using objective measures of service quality. Perhaps one reason for this lack of use is a fear that objective measures may lack validity because it may be more difficult to clearly rate features of service delivery than it is to rate product features. Another concern may be that objective measures may bear little relation to customer assessment of service quality. It may be that raters briefed to closely observe and score the service experience would notice aspects that real customers would not.

In spite of these fears, there are a number of attractive features of objective measures of service performance and quality. Objective assessments are potentially useful to management as they can provide overall evaluation of service encounters, but also allow survey work to focus on in particular aspects that are of managerial interest. Particular aspects of service staff performance, outlet appearance, or merchandising can be examined, for example, provision of full and correct information on pricing, refund policy or warranty. Measurement can focus on specific instances of poor performance or exceptions to policy. For instance, an organisation may initiate a training regime to improve the friendliness, and product knowledge of its staff. It may then monitor the efficacy of this regime by periodic mystery shopping surveys. Since customer perceptions of service quality may lag changes in service performance (see Bolton and Drew 1991), this approach offers a more immediate assessment of the impact of any training effort which is designed to improve the quality of service that staff provide to customers. For this reason, it is not uncommon for reward or incentive systems for service staff to be linked to objective service quality measurement results.

In addition to such assessment of one's own organisation, the technique is sometimes used to obtain intelligence or benchmarking information on the operations or marketing profiles of competitors. It is also used by organisations to monitor product knowledge, customer service or selling skills of their own staff or the staff of resellers, for instance a computer manufacturer may survey retailers to determine which brands are actively recommended. While there is little published work on the topic, World Wide Web sites such as the U.S. National Mystery Shopper Directory direct employment inquiries to over 450 organisations which conduct research in the United States. In the U.K. Dawson & Hillier (1995) found over half of the firms in a sample of 88 organisations who used *some* form of market research, also undertook mystery shopping. It is not unreasonable to infer that world wide mystery shopping expenditure runs into many hundreds of millions of dollars annually.

Research questions

Mystery shopping and other approaches to measuring objective service quality are apparently widespread and presumably management decisions are made based on the results. However, as stated previously, the lack of published work on the subject suggests that little is known about objective measures of service quality. A literature search through scholarly journals yielded only three published works on the topic: Dawson & Hillier (1995), Morrison et al (1997) and Wilson (1998). Dawson & Hillier (1995) primarily discussed ethical issues, Morrison et al (1997) examined cognitive issues affecting accuracy, and Wilson (1998) provided a general overview

of the research method. There appears to be little work on the topics of scale construction, reliability or validity of mystery shopping instruments or the process itself. Therefore research organisations may be using this technique, and presenting results on the results of surveys which may fail acceptable measurement criteria. Another issue is the extent to which such transaction based measures can generalise to “usual” levels of service. Mystery shopping surveys typically utilise only small samples, up to a few visits to any particular outlet or even only a single visit. In our experience, research organisations often present the results of such mystery shopping surveys by ranking outlets in order of the performance level identified. A store which performed at a certain level during one or several transactions is instantly classified as an over or under performer for the period. However, whether this is justified is questionable.

This paper addresses these issues by considering the following basic questions:

1. Can reliable mystery shopping instruments be created ? Reliability is a necessary, but not sufficient condition for validity (e.g. Peter 1981). We find that a high degree of reliability is, at least, possible with a carefully designed questionnaire and well trained interviewers.
2. How valid is mystery shopping ? We find that mystery shopping scores can certainly exhibit convergent and criterion validity.
3. How stable are the components of objective service quality ? By components we mean the individual aspects of service quality that collectively comprise an overall evaluation. If the important aspects vary from survey to survey, then it would be difficult for managers to know what aspects of service delivery are important to concentrate on improving. We find a high level of stability across surveys in terms of which are the important components of objective service quality.
4. How accurate are mystery shopping scores, considering that individual store level results from any survey are subject to sampling error ? Identifying single instances of good or poor service may have little managerial significance if such instances give no indication of the “usual” level of service provided. Can managers reasonably infer that a single survey or even series of surveys provides a reasonable indication of the level of average service performance of an individual retail outlet ? We find that the level of sampling error is very large, sufficient for a store to vary from being one of the very best to one of the very worst performers from one survey to the next.
5. Sampling error aside, how variable is retail service quality ? We find that there is considerable variation in the quality of service encounters at any one store. This throws doubt on the value of ascertaining and reporting any store’s average level level of service quality.

Implications of these findings are discussed at the end of the paper.

The measurement instrument

In our empirical research, we examined service quality in a retail setting, surveying the service provision of retail outlets all retailing the same product range. The measurement instrument was developed following two focus group discussions with both regular and occasional users of the product category. This was a low priced, disposable entertainment service for which the consumer may frequently ask simple questions of the seller relating to options and costs. The service is sold through numerous retail establishments who in effect act as “agents”, and like many services is consumed at the moment of production. The focus group discussion centred around determining the issues that were salient to consumers in their perception of what constituted good or bad service for this particular product category. The spectrum of service issues that were canvassed followed the five service quality factors presented by Parasuraman, Zeithaml & Berry (1988) namely tangibles, reliability, responsiveness, assurance and empathy.

Some aspects of customer service for this product appeared to be best measured as categorical in nature, such as whether the seller could provide a correct answer to a customer enquiry. Others were best measured as a matter of degree, such as friendliness. Therefore, a mixture of metric and categorical scales were used. The items are shown in Table I.

INSERT Table I HERE

The list shows not only items intuitively expected to be associated with customer perceptions of service, but two other items which the client organisation wished to include in order to measure the selling effectiveness of the retailers. This is not problematic as it will be shown that these items “drop out” of the later regression analysis examining the most significant predictors of overall assessment of objective service quality.

Survey methodology

Approximately four hundred and fifty sites were surveyed three times in each survey round. This process was repeated four times over twelve months. Some sites were not included in every survey round because they dropped below the minimum sales threshold used for inclusion in the mystery shopping survey. Subsequently some sites were dropped from the analysis because they did not feature in all four rounds. This is not considered to have biased the results because the remaining sites in the sample still had a very wide variation in sales volume.

In each survey round the team of raters was reallocated so that raters rated different outlets in each round. The team of raters totalled approximately 60 in each round.

Each survey comprised three separate visits, on separate days of the week, over a 10 working day period, at different times of the day, requesting different products within the small product range. The shoppers were briefed with carefully constructed inquiries and filled out the questionnaire immediately after the shopping encounter. The sellers (retail sites) were aware that their principal undertook this research but were unaware when it would take place, or the form of inquiries. During the process there were no reports of the seller guessing the inquiry was anything other than

genuine. On each occasion the shopper made a purchase to ensure that the shopping encounter was realistic. The seller was not told at the conclusion of the encounter that it was a service audit.

Basing the set of questions on expressed consumer sentiment as well as feedback from people who were experienced in the industry provided us with some confidence as to the face validity of the items. We now address the issue of the reliability of the mystery shopping process via an examination of interrater reliability.

Q.1 Reliability

Validity can be defined as the extent to which a device measures what it is intended to measure (eg. Churchill 1979). A necessary but not sufficient precondition to achieve validity is a high level of reliability, the degree to which ratings reflect true scores, or variation, in the phenomena under study (Guilford 1954). In terms of interrater reliability, reliability can be operationalised in two ways. For ratings based on a metric scale, reliability is the degree to which ratings of different judges are proportional when expressed as deviations from their means (Ebel 1951). Raters may also use categorical items in which case the relevant statistic is agreement - the extent to which different judges make the same judgements about a rated subject (Tinsley and Weiss 1975). The degree of agreement can also be translated into a comparable estimate of reliability as will be shown shortly.

In order to obtain data suitable for examining rater reliability, we carried out 60 shopping encounters using pairs of raters who made an inquiry and purchase together, and then filled out their questionnaires separately. We used four pairs of raters who each surveyed 15 outlets, for a total of 60 outlets being surveyed in this way. To minimise any effects due to chance similarity in rating lenience/severity between raters we rotated the pairs as shown in Table II.

INSERT Table II HERE

We examined interrater reliability for the metric scale items using the Intraclass Correlation Coefficient (1979) which is based on an analysis of variance model. Interrater agreement for categorical questions was examined via the proportion of agreement adjusted for the proportional reduction in loss from using the combined rating. We avoided using the popular measure of agreement adjusted for chance, *Kappa* (Cohen 1960) because *Kappa* is too conservative a measure and is inappropriate for most marketing research applications (Rust and Cooil 1994). To address this shortcoming, Perreault & Leigh (1989) developed a reliability measure called PRL (Proportional Reduction in Loss) for nominal data which adjusts for the number of raters and the number of categories. Essentially, an item with more categories and/or more judges with a given level of agreement will exhibit a higher PRL score than one the same agreement with fewer categories and/or judges. The results of the ICC and PRL analyses are shown in Table III and Table V.

INSERT Table III AND Table V HERE

As can be seen the overall degree of reliability of ratings is high. Both the ICC and PRL measures are comparable to Cronbach's alpha (Rust and Cooil 1994), (in fact Cronbach's alpha is one type of ICC) which has a widespread rule of thumb of 0.7

being acceptable for exploratory analysis and 0.9 for experimental research. Table III and IV show that all the items with the exception of Q 12 are over 0.7. This process itself can be seen to be useful in identifying problem items, however, overall the questionnaire appears to have adequate to high levels of reliability.

One possible limitation of this test was that the four raters who undertook the reliability testing were among the more experienced and best trained in the larger field team. While this might be a shortcoming in that the results are less generalisable, it gives a degree of confidence that no collusion occurred between the raters which would have boosted the degree of agreement in scores. Even though the raters may have been experienced interviewers, it nevertheless demonstrates that high levels of interrater reliability are at least possible.

Having established that the instrument, and a sample of personnel were capable of achieving generally high levels of interrater reliability, we now address the issue of convergent validity.

Q.2 Convergent Validity

Convergent validity refers to whether scores from certain variables correlate with other variables designed to measure the same construct (Campbell and Fiske 1959). In this case the other measure was a “global” or overall service quality score provided by the mystery shopper. As we used a combination of metric and categorical variables, we used cross-tabulations for the categorical variables (after also re-coding the global scores into three categories), and correlations for the metric variables. The results are shown in Table VII. They show that eleven out of the thirteen variables had a strong association with the global assessment of service quality.

INSERT Table VII HERE

Criterion Validity

Criterion validity (sometimes called predictive validity) refers to how well the scores from a test correlate with some other criterion of interest. For instance, scales measuring market, or customer orientation are often correlated against organisational performance (e.g. Deshpandé and Farley 1998). Service quality is presumed to be related to sales, and the rationale for service quality improvements is usually that they will lead to *increased* sales for the organisation. We assessed criterion validity by examining the association between the objective service quality scores and sales performance using correlations. The correlation was positive, albeit not very high (Pearson $r = 0.18$) and significant at $p < 0.001$. Objective service quality measurement appears to exhibit criterion validity, as well as interrater reliability and convergent validity.

Q3. Stability of service quality components

We then wanted to assess the stability of the most important predictors of overall service quality. To do so we created a regression model using the stepwise method, which includes and rejects independent variables in the regression model according to

their F-values (Norusis 1993). This procedure was used on each set of survey results (n = approximately 1,300 shopping encounters in each set, averaged to produce 400 cases for each of four survey waves) to determine which variables were significant predictors. To facilitate the use of categorical variables, these were coded as dummy variables (see Hair et al. 1995).

We checked for possible bias in the results caused by individual raters by examining the average scores for the predicted variable, by rater. There were several that were markedly below the mean. However, these were all raters who undertook small numbers of ratings, (typically 0.03% of the sample) and their average scores appeared to be largely a function of the geographic area they were allocated to (non-metropolitan and lower socio-economic areas tended to score poorly, regardless of rater). In addition to this diagnosis, a sample of raters were coded as dummy variables to determine if the rater him/herself could possibly be a significant predictor variable. This was not evident, resulting in R²s of 0.00.

We checked the data for deviations from the normal requirements for regression. These are linearity, constant variance, independence, normality of error term distribution, and absence of multicollinearity (Hair et al. 1995). Most were satisfied. The exception was independence. We detected positive autocorrelation in the residuals, with the Durbin-Watson test indicating this autocorrelation was significant at p<0.05. This signified that there was a mild “hangover” effect, with the score given to a particular outlet by a particular rater being partially predictable from the score given by that rater to the previous store. Autocorrelated residuals can lead to overestimations of a model’s predictive ability (Mendenhall and Sincich 1996). To remedy this we introduced an autoregression term, shown in equation 1:

$$Y_t = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6Y_{t-1} + \text{error} \quad (1)$$

Introducing the autoregression term reduced the level of autocorrelation in the residuals to an acceptable level as indicated by the Durbin-Watson statistic. It also resulted in minor improvements to the R² for each wave of between 0.01 to 0.08. The adjusted R² with or without the autoregressive term was over 0.70, indicating that the subset of variables are excellent “predictors” of overall service quality.

To address the question of the stability of these “predictors”, we can examine the regression coefficients. Standardised coefficients are used because they are directly comparable even if different measurement scales are used, as was the case here. The standardised coefficients from the regression analyses using equation (1) showed that the same five indicators plus the autoregression term were significant predictors of overall service performance ratings, over the four waves of data collection. Furthermore, they tend to retain their relative importance. Results are shown in Table VIII.

INSERT Table VIII HERE

Suitability of Beta values

Standardised coefficients, or beta values, can be influenced by the level (Allison 1977) or variance in the data (Schumacker and Lomax 1996) and for this reason it has

been argued that they are inappropriate for comparison over multiple data sets (Schumacker and Lomax 1996).

We examined whether the beta values might be being affected by such changes. The questionnaire item with the greatest movement in beta value over time was Q6 (showing interest). The standard deviations for Q6 were 1.4, 1.2, 1.3, and 1.3 for waves 1-4. An F-test indicated there were significant differences in variation over the four waves ($P < 0.004$). However this change is very small (0.02 from wave 1 to wave 2) in absolute terms, and we conclude that the variance is small enough not to unduly affect the beta values. The mean scores for this question were 5.1, 5.3, 5.1, and 5.1 for the four waves. A post hoc test indicated a significant difference between wave 2 and the other waves, but this difference was also very small in absolute terms (0.02). We conclude that the beta weights were not unduly affected by either changes in variance or level.

The results show that a consistent set of questionnaire variables (leaving aside the autoregressive term) accounts for over 70% of the variance in overall evaluation of the shopping encounter. Furthermore, these variables exhibit reasonable stability over the four rounds in their relative importance in predicting the overall evaluation. For instance, Q6 (interest) tends to be the most important or second most important variable and Q5 (ending the exchange with well-wishing) tends to retain its rank as the least important variable in terms of beta weight. This stability is confirmed by taking the regression coefficients derived from the Wave 1 data and using them to create a predicted service quality assessment for the *other* three waves of data. The resultant R^2 's were 0.71, 0.74 and 0.70 for waves 2, 3, and 4 respectively, showing that the original regression model (ie., calculated from wave 1) has the ability to predict overall service quality assessments in later data sets (ie. waves 2, 3 and 4). This is a pleasing result given that researchers such as Ehrenberg (1993) have noted the abysmal record of least-squares based methods in developing empirical generalisations. It has been suggested that they have not been helpful in identifying empirical generalisations in marketing because they direct emphasis towards development of new models rather than refining/testing existing models (Lindsay and Ehrenberg 1993; Ehrenberg 1995). Our results show that regression can still be a useful procedure for developing empirical generalisations but only, as Ehrenberg advocates, replications are employed and researchers *look* for consistency in results.

This stability in coefficient values is graphed in Figure 1 for easier interpretation. It shows that while each variable is somewhat mobile over time, there is reasonable stability in terms of ranking from most to least important. The autoregressive term in particular declines over the four waves, perhaps indicating some learning effects from mystery shoppers who participated in the four rounds, and possibly improvements in briefing and training. We interpret the stability of service components as predictors of overall evaluations as positive evidence that the mystery shopping instrument was reliable and valid.

INSERT Figure I HERE

Q.4 Level of Individual Store level Accuracy

Our examination has shown that the measurement instrument exhibited a good degree of reliability, and convergent & criterion validity and in use. But how can the results of such mystery shopping surveys be used? In the author's experience the most common way in which mystery shopping data is used is for the outlets that were surveyed to be listed with their scores for the period, and possibly compared to previous periods. The implication of this is that the score from an outlet represents its "level of quality" for the period, and changes in scores are interpreted as improvements or declines in service quality. Is this justified? To answer this question we considered the variance in scores and the issue of sampling error. The scores were the predicted values generated by the SPSS regression procedure, for each wave. Predicted scores here are in effect, a weighted composite of the significant predictors of overall service quality.

Variation in scores

We found that there was considerable variation in scores for individual outlets from wave to wave. Figure 11 illustrates this using a randomly selected sample of six outlets. As can be seen, a high score in one round is often followed by a much lower score in another round, and vice versa. We found this to generally be the case. In fact the average amount of variation in scores for any outlet over the *four* rounds approaches the average amount of variation for *all* outlets in any *one* round. The average standard deviation for all outlets in a wave was 10.1. The average standard deviation for an outlet over the four waves was 7.8. This demonstrates how much the scores for any outlet vary from one wave to another. Therefore it may be very unwise to categorise a particular outlet as a "good" or "poor" performer based on one round of mystery shopping. This is unless the number of observations is very large, which is cost prohibitive and not normal practice in the authors' experience.

INSERT Figure II HERE

Replication

This finding relating to variation in scores was surprising. In case it was a non generalisable artifact of our data we replicated the analysis on an unrelated data set. The new data set comprised 5 individual observations over a 12 month period for each of 40 retail outlets. These outlets were in an entirely different service industry to the first study, and the field team that carried out this survey had no members which were involved in the first study. The findings were similar. The average variation in scores for any particular outlet over the five rounds was approximately equal to the average variation for all outlets in any one round (11.2 cf. 11.9).

Sampling error

We showed that on average there is considerable variation in scores for any particular outlet over time. A more serious issue is whether these changes can simply be attributable to sampling error. Each wave of mystery shopping surveys analysed in

this paper comprised three observations. This is a very small number on which to base statistical inference, though it is probably large compared to many commercial mystery shopping surveys. We wished to see if this small number of observations would be adequate to identify how stable or variable service quality is over time. If three observations were not found to be enough, we would combine the results from waves 1&2 and compare them to the combined results for waves 3&4.

We calculated the standard deviation of scores for the three observations for each outlet. This was done across all the waves. The average standard deviation was 13.1. This enabled us to calculate a confidence interval in the mean scores for each outlet, for one wave of mystery shopping results. This used the basic formula (Berenson and Levine 1996) for a 95% confidence interval:

$$\mu_k \pm (1.96)\sigma_x / \sqrt{n} \quad (2)$$

Where μ_k is the mean score, 1.96 is the Z value corresponding to an area of (1-.05)/2 from the centre of a normal distribution, σ_x the standard deviation and n is the sample size.

From this we computed the 95% confidence interval to be +/- 15 scale points. Therefore if the mean scores for an outlet over two periods fall outside this (large) interval we can conclude at a 95% level of confidence that its level of service quality is different from one wave to another.

We examined the 228 outlets that were monitored over the four waves of data collection and the differences in scores for each of those outlets between waves 1&2, 2&3 and 3&4. The number of significant differences in service quality is shown in Table X.

INSERT Table X HERE

This number of cases is under what would be expected by chance. At a 95% confidence level we would expect $228 \times 0.05 = 11$ significant differences simply due to chance. Therefore comparisons between two sets of three observations provides no evidence that service quality is variable over time. Of course, many stores could have gone through real changes in the average quality of their service provision, but the lack of statistical power provided by only 3 observations makes it impossible to distinguish this real change from sampling variation.

Mindful that three observations are indeed a small number we aggregated the scores for waves 1&2 and also aggregated scores for waves 3&4. Aggregating meant we were now comparing two sets of six observations. The confidence interval using formula (2) was now +/- 11.5 points.

We compared the two aggregated scores for each of the 228 outlets to identify if any significant differences in performance were evident.

The results are shown in Table XI.

INSERT Table XI HERE

The number of cases displaying significant differences is again within sampling error.

So we have shown that there is substantial individual level variation from one survey to another, but this variation could easily be due to sampling error. For all practical purposes it is impossible to determine whether an individual store's improvement or decline in service provision from one survey to the next is real or not. This renders the comparison of results at the individual store level almost valueless. It would certainly be wrong for store managers to use changes in mystery shopping stores to assess staff or their own performance. An exception may be using the results to identify exceptions to pre-specified minimum standards that do not require managers to generalise about staff or store performance.

Q.5 Variation in Service Quality at Individual Site Level

Sampling error aside, how much variation is there in service performances? Even if a complete census was taken of every single service encounter at a store (thereby removing all sampling error) would there still be considerable variation, encounter to encounter? If this was the case it would be rather meaningless to talk about an outlet's average or general level of service quality.

To address this question we report the variation in scores exhibited by stores within any one survey wave in Table XII. This simply shows how little, or how much service quality can vary within any particular retail outlet (in our study) within the space of ten days. The range refers to range of points from lowest rating to highest rating for the three observations.

INSERT Table XII HERE

Table XII shows that most of these retail outlets exhibit *large* variation in reported service levels within any one wave. 59% of outlets have a range of scores beyond 20 points. The effective scale range is approximately 90 points with the lowest scores computed from the regression coefficients being 9 and the highest being 99. This suggests that to talk of a retail outlet's level of (average) service quality over a period might be meaningless. The old French saying seems appropriate: *there is no such thing as a great wine, only great bottles.*

Of course, it is partly because of this high variability in service delivery that firms wish to measure service quality. Unlike products, for which standardisation is easier, standardisation of service quality is difficult and requires on-going management.

Conclusions

This research has important implications for providers of mystery shopping research. We have outlined an approach for examining the reliability of a mystery shopping instrument and have shown that good levels of inter-rater reliability are possible. Also we have shown that the component items in mystery shopping instruments can display acceptable levels of convergent and criterion validity; and stability in comparative importance over multiple surveys. However, the validity of the process itself, that of rating and ranking individual outlets from a survey of only a few service encounters, is doubtful. The results suggest that research organisations, and their

clients, should interpret the results of mystery shopping surveys with some caution. It is not advisable to generalise about the extent to which certain outlets are good performers from mystery shopping surveys which are really only measures of individual performances. The degree of service provided by retail outlets in this study was quite variable for most outlets.

Clearly a prudent approach is to aggregate mystery shopping results across time periods, and even better, across stores/employees/agents. Rather than ranking individual service providers and comparing changes in rank or score over time, it would be better to concentrate on aggregate scores, perhaps for regions or countries, and evaluate their

- absolute level - how well are we doing ?
- degree of variation - how consistent is our delivery across outlets ?
- components - what are we strong/weak on ? what really matters ?
- and changes in these statistics over time.

Another very appropriate use for mystery shopping may be to monitor “minimum standards” for the organisation. For instance, a service organisation may set a standard that no customer should wait for more than 3 minutes to be served or that sales staff must mention certain pre-specified product benefits when serving clients. It is not necessary to generalise about such results at the individual store level. The research provider can then simply see if such standards are ever breached and give appropriate feedback to the client.

Recommendations for future research

Our study at least partially throws doubt on the validity of the mystery shopping process. Further research is needed to better describe the limitations of the technique and possible ways of overcoming these limitations. One avenue to explore is the extent to which customers’ perceptions of service performance correlate with expert ratings which use the same criteria. This would provide more evidence as to whether mystery shopping really measures what it is intended to measure.

The study also suggests that service quality may be best conceptualised as something that varies considerably from transaction to transaction. Should this variation be of greater concern to marketing managers than the average service quality of a store or group of stores ? It is possible that there is asymmetry in the consequences of high and low service quality. An excellent service performance may often have little effect whereas a really bad service encounter might very often result in customer defection or bad word-of-mouth. If this were true, management would be wise to worry less about improving the average level of service performance and worry more about reducing (downward) variability. These are questions which require empirical investigation.

Finally, there is also the obvious question of whether and by how much, changes identified by objective service quality assessment correspond to customers’ subjective assessment of overall service quality. In other words, how much do customers notice

changes in service quality ? Such research would provide managers with further evidence as to the utility of mystery shopping programs. Replication and extension research is also required in other service categories. Our research examined a fairly low involvement entertainment service, research is needed in traditional durable product stores, financial services and other settings where the total service is consumed on site, for example restaurants.

References

- Allison, Paul D. (1977), "Testing for Interaction in Multiple Regression," *American Journal of Sociology*, 83 (No. 1), 144-153.
- Berenson, Mark L. and David M. Levine (1996), *Basic Business Statistics - Concepts and Applications*, Sixth ed. Englewood Cliffs, New Jersey: Prentice Hall, Inc.
- Bolton, Ruth N. and James H. Drew (1991), "A Longitudinal Analysis of the Impact of Service Changes on Customer Attitudes," *Journal of Marketing*, 55 (January), 1-9.
- Buttle, Francis (1996), "SERVQUAL: Review, Critique, Research Agenda," *European Journal of Marketing*, 30 (No. 1), 8-32.
- Campbell, Donald T. and Donald W. Fiske (1959), "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix," *Psychological Bulletin*, 56 (No. 2, March), 81-105.
- Churchill, Gilbert A., Jr (1979), "A Paradigm for Developing Better Measures of Marketing Constructs," *Journal of Marketing Research*, 16 (February), 64-73.
- Cohen, Jacob (1960), "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20 (1), 37-46.
- Cronin, J. Joseph, Jr. and Steven A. Taylor (1994), "SERVPERF Versus SERVQUAL: Reconciling Performance-Based and Perceptions-Minus-Expectations Measurement of Service Quality," *Journal of Marketing*, 58 (January), 125-131.
- Dabholkar, Pratibha A., Dayle I. Thorpe and Joseph O. Rentz (1996), "A Measure of Service Quality for Retail Stores: Scale Development and Validation," *Journal of the Academy of Marketing Science*, 24, 3-16.
- Dawson, Janet and Jill Hillier (1995), "Competitor Mystery Shopping: Methodological Considerations and Implications for the MRS Code of Conduct," *Journal of the Market Research Society*, 37 (No. 4, October), 417-428.
- Deshpandé, Rohit and John U. Farley (1998), "The Market Orientation Construct: Correlations, Culture, and Comprehensiveness," *Journal of Market Focused Management*, 2 (No. 3), 237-239.
- Ebel, Robert L. (1951), "Estimation of the Reliability of Ratings," *Psychometrika*, 16 (No. 4), 407-424.
- Ehrenberg, A.S.C. (1995), "Empirical Generalisations, Theory, and Method," *Marketing Science*, 14 (No. 3, Part 2 of 2), G20-G28.
- Ehrenberg, Andrew S.C. and John A. Bound (1993), "Predictability and Prediction," *Journal of the Royal Statistical Society Association*, 156 (Part 2), 167-206.
- Ehrenberg, A.S.C. and J.M. Shewan (1953), "The Objective Approach to Sensory Tests of Food," *Journal of the Science of Food and Agriculture*, 4 (October), 482-490.

- Guilford, J.P. (1954), *Psychometric Methods*. Bombay: Tata - McGraw Hill.
- Hair, Joseph F., Rolph E. Anderson, Ronald L. Tatham and William C. Black (1995), *Multivariate Data Analysis*, Fourth ed. New Jersey: Prentice Hall International.
- Jacoby, Jacob and Jerry C. Olson (1985), "Perceived Quality," Lexington: Lexington Books.
- Kamakura, Wagner A. and Gary J. Russell (1993), "Measuring Brand Value with Scanner Data," *International Journal of Research in Marketing*, 10, 9-22.
- Lindsay, R. Murray and A.S.C. Ehrenberg (1993), "The Design of Replicated Studies," *The American Statistician*, 47 (3), 217-228.
- Maynes, S. E. (1976), "The Concept and Measurement of Product Quality," *Household Production and Consumption*, 40, 529-59.
- Mendenhall, William and Terry Sincich (1996), *A Second Course in Statistics: Regression Analysis*, Fifth ed. New Jersey: Prentice-Hall International, Inc.
- Morris, Ruby Turner and Claire Sekulski Bronson (1969), "The Chaos of Competition Indicated by Consumer Reports," *Journal of Marketing*, 33 (July), 26-43.
- Morrison, Lisa J., Andrew M. Colman and Carolyn C. Preston (1997), "Mystery Customer Research: Cognitive Processes Affecting Accuracy," *Journal of the Market Research Society*, 39 (No. 2, April), 349-361.
- Norusis, Marija J. (1993), *SPSS for Windows - Advanced Statistics Release 6.0*. Chicago, USA: SPSS Inc.
- Parasuraman, A, V Zeithaml and LL Berry (1988), "SERVQUAL: A Multiple-item Scale for Measuring Consumer Perceptions of Service Quality," *Journal of Retailing*, 64 (Spring), 12-40.
- Perreault, William D., Jr. and Laurence E. Leigh (1989), "Reliability of Nominal Data Based on Qualitative Judgments," *Journal of Marketing Science*, 26 (May), 135-148.
- Peter, J. Paul (1981), "Construct Validity: A Review of Basic Issues and Marketing Practices," *Journal of Marketing Research*, 18 (May), 133-145.
- Riesz, Peter C. (1980), "A Major Price-Perceived Quality Study Reexamined," *Journal of Marketing Research*, 17 (May), 259-262.
- Rust, Roland T. and Bruce Cooil (1994), "Reliability Measures for Qualitative Data: Theory and Implications," *Journal of Marketing Research*, 31 (February), 1-14.
- Rust, Roland T., Anthony J. Zahorik and Timothy L. Keiningham (1995), "Return on Quality (ROQ): Making Service Quality Financially Accountable," *Journal of Marketing*, 59 (April), 58-70.

Schumacker, Randall E. and Richard G. Lomax (1996), *A Beginner's Guide to Structural Equation Modeling*. New Jersey: Lawrence Erlbaum Associates, Publishers.

Shrout, Patrick E. and Joseph L. Fleiss (1979), "Intraclass Correlations: Uses in Assessing Rater Reliability," *Psychological Bulletin*, 86 (No. 2), 420-428.

Tinsley, Howard E.A. and David J. Weiss (1975), "Interrater Reliability and Agreement of Subjective Judgments," *Journal of Counseling Psychology*, 22 (No. 4), 358-376.

Wilson, Alan M. (1998), "The Use of Mystery Shopping in the Measurement of Service Delivery," *The Service Industries Journal*, 18 (3, July), 148-163.

TABLES AND FIGURES

Table I Questionnaire items

Friendliness	metric
Interest	metric
Enjoyable encounter	metric
Confidence in answering a question	metric
Positive reflection on the principal	metric
Cleanliness and tidiness of the outlet	metric
Immediate service or wait	categorical
Form of greeting	categorical
Friendly verbal exchange	categorical
Seller ending exchange with well wishing	categorical
Mentioning of other products by the seller	categorical
Prompting for purchase/cross selling	categorical
Product knowledge	categorical
Technical problems	categorical
Global measure of satisfaction with the overall purchase experience. This item was used as the dependent variable in tests of predictive validity of the questionnaire	metric, rating out of 100.

Table II Rater Rotation

Raters 1 & 2 - 15 outlets each
Raters 1 & 3 - 15 outlets each
Raters 2 & 3 - 15 outlets each
Raters 2 & 4 - 15 outlets each

Table III Interrater Reliability values

Item	ICC value
Q. 4	0.90
Q. 6	0.90
Q. 7	0.86
Q. 11	0.76
Q. 12	0.55
Q. 15	0.81

Table V Reliability - categorical questions

Item	Average Proportion Agreement	# categories	Reliability
Q. 1	0.87	7	0.90
Q. 2	0.83	3	0.85
Q. 3	0.67	5	0.75
Q. 5	0.87	3	0.89
Q. 8	0.98	2	0.99
Q. 9	1.0	3	1.0
Q. 10	0.95	3	0.98
Q. 13	0.10	4	1.0

Table VII Measures of association between service quality components and global score

Variable	Type	Association with overall service quality assessment
Friendliness	metric	$r=0.73, p<0.001$
Interest	metric	$r=0.78, p<0.001$
Enjoyable encounter	metric	$r=0.82, p<0.001$
Confidence in answering a question	metric	$r=0.52, p<0.001$
Positive reflection on the principal	metric	$r=0.38, p<0.001$
Cleanliness and tidiness of the outlet	metric	$r=0.40, p<0.001$
Immediate service or wait	categorical	$\chi^2=57, df 12, p<0.0001$
Form of greeting	categorical	$\chi^2=283, df 4, p<0.0001$
Friendly verbal exchange	categorical	$\chi^2=303, df 8, p<0.0001$
Seller ending exchange with well wishing	categorical	$\chi^2=312, df 4, p<0.0001$
Mentioning of other products by the seller	categorical	$\chi^2=23, df 2, p<0.0001$
Prompting for purchase/cross selling	categorical	$\chi^2=6, df 2, p=0.21$
Product knowledge	categorical	$\chi^2=182, df 4, p<0.0001$
Technical problems	categorical	$\chi^2=2, df 2, p=0.46$

Table VIII Standardised Regression (Beta) coefficients

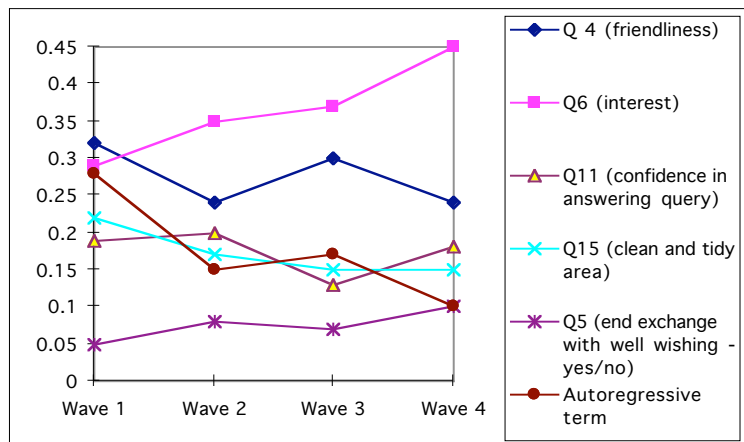
(Standard errors in parentheses)

Wave Wave 2 Wave 3 Wave 4
1

	Beta	Beta	Beta	Beta
Q 4 (friendliness)	.32 (.03)	.24 (.04)	.30 (.04)	.24 (.03)
Q6 (interest)	.29 (.03)	.35 (.03)	.37 (.04)	.45 (.03)
Q11 (confidence in answering query)	.19 (.01)	.20 (.02)	.13 (.01)	.18 (.01)
Q15 (clean and tidy area)	.22 (.02)	.17 (.02)	.15 (.01)	.15 (.01)
Q5 (end exchange with well wishing - yes/no)	.05 (.01)	.08 (.01)	.07 (.01)	.10 (.02)
Q16 LAGGED	.28 (.02)	.15 (.01)	0.17 (.02)	.10 (.01)
Adj. R² with autoregression term	0.80	.73	0.77	.76
Adj. R² without autoregression term	0.72	0.72	0.74	0.75

Note: all coefficients significant at $P < 0.01$.

Figure I Beta values for the four survey rounds (average of 3 encounters ea. round)



The top five predictors of overall service quality scores remain reasonably stable in rank over four survey rounds

Figure II Variance in scores from wave to wave - six randomly selected outlets

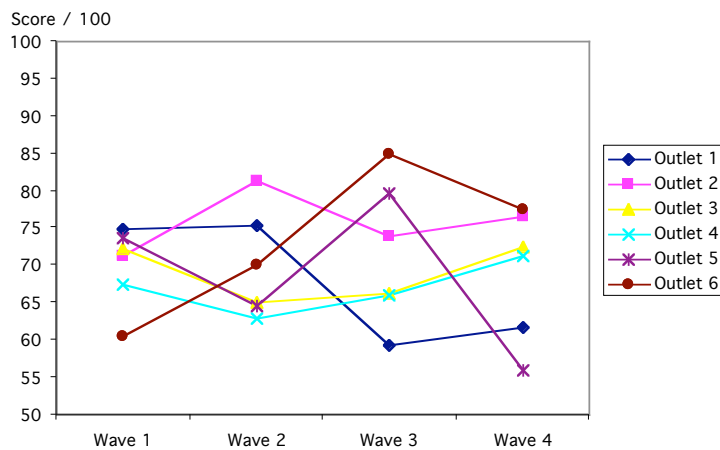


Table X Significant Differences

Waves	Number of outlets displaying differences significant at P<0.05
1 & 2	5
2 & 3	7
3 & 4	1

Table XI Significant Differences - aggregating survey waves

	Number of outlets displaying differences significant at P<0.05
Comparison between aggregated scores for waves 1&2 compared to aggregated scores for 3&4	4

Table XII Variation in scores for outlets in any one survey

Range	Percentage of outlets
1 to less than 5 points	3
5 to less than 10 points	10
10 to less than 20 points	28
20 to less than 30 points	28
30 to less than 40 points	18
over 40 points	13