# AUSTRALASIAN JOURNAL OF MARKET RESEARCH

**JANUARY 2000**

**VOLUME 8, NUMBER 1**

**ISSN 1325-4634**

## CONTENTS

<table>
<thead>
<tr>
<th>Article Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responding to the Problem of Nonresponse</td>
<td>3</td>
</tr>
<tr>
<td><em>Philip Gendall</em></td>
<td></td>
</tr>
<tr>
<td>An Evaluation of Political Polling in a Postal Vote</td>
<td>19</td>
</tr>
<tr>
<td><em>Malcolm Wright</em></td>
<td></td>
</tr>
<tr>
<td>The Reliability and Validity of Objective Measures of Customer Service: “Mystery Shopping”</td>
<td>29</td>
</tr>
<tr>
<td><em>John Dawes, Byron Sharp</em></td>
<td></td>
</tr>
</tbody>
</table>
The *AJMR* is the official journal of the Market Research Society of Australia (MRSA), and the Market Research Society of New Zealand (MRSNZ). It is published twice yearly.

All members of the MRSA receive each copy free and all full members of the New Zealand Society receive complimentary copies.

Subscribers may obtain copies for $30.00 annually, or $35.00 for overseas subscribers ($5.00 extra for Air Mail per issue).

Original papers are invited. There is an editorial policy to ensure that a mix of theoretical and practical papers are presented. Papers should be of some relevance to at least a segment of the population of practicing market researchers. Purely academic marketing papers are probably more suitable elsewhere. All papers are refereed by independent assessors.

Contributors of papers will receive 5 copies of the issue in which their article appears.

Manuscripts are to be typed and double spaced. Diagrams and figures should be of professional quality. In writing articles, authors should consider the style of articles in previous issues.

On acceptance of an article, authors will be asked to supply their article on a 3.5" disk. The word processor of preference is Microsoft Word for Windows. However, conversions can be made from other packages.

The *AJMR* is an open publication. All expressions of opinion are published on the basis that they are not to be regarded as expressing the official opinion of the Market Research Society of Australia Limited or the Market Research Society of New Zealand. The MRSA/MRSNZ accepts no responsibility for the accuracy of any of the opinions or information contained in this publication and readers should rely upon their own enquiries in making decisions.

All papers for submission should be sent to

Professor Lester W. Johnson  
Editor, *AJMR*  
Monash Mt. Eliza Business School  
P.O. Box 2224  
Caulfield Junction, Vic 3161  
Australia

All business enquiries and requests for additional copies should be sent to

The Market Research Society of Australia  
P.O. Box 697  
North Sydney, NSW 2059  
Phone 02 9955 4830  
Fax 02 9955 5746

Copies of articles in the *AJMR* may be made for personal or classroom use, without charge and with the publisher’s consent. Copying for any other purpose must first be approved by the publisher.

The Society recognises the contribution of Lester Johnson (Editor) and the Monash Mt. Eliza Business School, Monash University.

Published by the Market Research Society of Australia Limited

Copyright 2000.

Editor - Lester Johnson  
Production - Lynette Francis
RESPONDING TO THE PROBLEM OF NONRESPONSE

Philip Gendall
Department of Marketing
Massey University

Introduction

It is generally asserted and commonly believed that survey response rates are falling and have been declining for decades. The reasons seem self-evident: more women are working, society is more urbanised and lifestyles have changed; people are busier and more mobile, they have answering machines and caller ID on their telephones, they are more fearful of crime, are annoyed by telemarketing and suspicious of SUGGING (selling under the guise of research), FRUGGING (fund raising under the guise of research) and CUGGING (voter canvassing under the guise of research).

This perception, that response rates are declining, may be true but the time series information available presents a more complex and less certain picture. A 1995 review of 56 time series around the world showed 22 declines in response rates, 16 with no change, 14 with variable trends, and four with rising response rates (Smith, 1995). While declines clearly greatly outnumber increases the pattern is hardly one of general decline. To complicate matters, the time series studied are not representative of all surveys. They are based on studies with results in the public domain and contain few commercial polls and no market research.

Similar research on response trends in official surveys (Labour Force Surveys) over time and in different countries has produced similar results. In some countries there is a clear downward trend, in others the response rate has remained stable while refusals have increased and non-contacts decreased (indicating more fieldwork effort). In New Zealand, two mail surveys of the general public on the topic of religion, conducted in 1991, and again in 1999, had response rates of 66% and 65%, respectively.

Thus it is difficult to quantify the scale of the nonresponse problem. However, anecdotal evidence suggests that commercial survey response rates in Australia and New Zealand are declining (particularly for telephone surveys) and, where response rates are not declining, more effort is required to maintain them at previous levels.

What is the problem?

Why should we be concerned about declining response rates? The answer is, of course, because as response rates fall, the potential for nonresponse bias increases. In other words, there is an increasing risk that our sample of respondents is not representative of the population in which we are interested because those who respond are different to those who did not.

However, it is important to note that a low response rate increases the potential for nonresponse bias. Nonresponse bias is not inevitable with a low response rate, because it depends on the level of nonresponse and on the extent to which respondents and nonrespondents differ on the variables of interest.
Defining a "response rate"

There is sometimes confusion over the term "response rate". This is because there are different ways of calculating the response rate for a survey and these give different results. The American Association for Public Opinion Research (AAPOR) has produced a standard definition of a survey response rate for RDD telephone surveys and in-person household surveys:

The number of completed interviews with reporting units divided by the number of eligible reporting units in the sample. (AAPOR, 1998)

Depending on how partial interviews are considered and how cases of unknown eligibility are handled, six different response rates can be calculated under this definition. However, the basic formula is as follows:

\[
\frac{\text{Completed interviews}}{\text{Completed interviews} + \text{Refusals} + \text{Noncontacts}}
\]

Clearly the key determinants of response rate are refusals and noncontacts; both of these are considered later when measures to increase response rates are discussed. First, however, the question of whether there is an acceptable level of nonresponse for a survey needs to be addressed.

A Minimum Acceptable Response Rate

Several researchers have examined the pattern of responses to survey questions at different response rates. In one study, Hosie (1995) compared survey estimates for a mail survey at different response rates. The survey involved 1900 members of the New Zealand general public and addressed health, social and environmental issues.

Hosie calculated interim survey estimates for the 178 questions in the mail survey at several different response rates (21%, 30%, 45%, 51% and 60%) and compared them with the final estimates (at a 70% response rate). She found, at 51% response rate, only seven questions for which the absolute difference between the interim estimate and the final estimate was greater than two percentage points, and no cases in which the absolute difference was greater than 3%.

Hosie's conclusion was that these observed differences would be unlikely to change the interpretation of the survey's results and consequently that a response rate of 50% could be regarded as minimum acceptable response rate for most practical purposes.

Similar studies (see Dolsen & Machlis, 1991; Johnson, 1983; Merkle, Bauman & Lavrakas, 1993; and Garland & Tweed, 1998) produced similar conclusions for mail, personal and telephone surveys. However, Hosie's analysis of a telephone survey of payphone usage and awareness was much less supportive of the idea that there is little practical difference between final survey results and interim results at 50% response rate. In this part of her study some important variables had differences of

\[\text{1 Only two differences between interim estimates at 51% and final estimates at 70% were statistically significant (p<.05), the proportion living in a country home, underestimated by 2.0 percentage points, and the proportion of Methodists, overestimated by 1.1 percentage points. Hosie's use of absolute differences was based on the rationale that significance is dependent on sample size and the fact that a significant difference may not be of any practical importance. Nevertheless, the practical significance of an absolute difference does depend on the level of the estimate concerned. In other words, a difference of two percentage points is more important for an estimate of 5% than one of 75%. Thus neither the size of the absolute difference between interim and final estimates or the statistical significance of the difference is a completely satisfactory measure of the effect of response rate on estimates.}\]
between 5% and 6%. Furthermore, even if estimates at 50% response and final response are identical, this does not mean that there is no nonresponse bias if the final response rate is less than 100%. Brown (1994), for example, reports an estimate for the number of self-employed people in Britain, based on a survey with a 60% response rate, which was less than half the number reported by the Government Statistical Service.

Nevertheless, further evidence that efforts to substantially improve response rates may not produce any meaningful reduction in nonresponse bias was provided by an American experiment conducted by the Pew Research Center (1998). The Pew Research Center compared two identical telephone surveys, one using standard polling methodology – 1000 adults contacted by phone over a five-day period – the other a more rigorous approach involving an eight-week field period, an advance letter and incentive, repeated attempts at refusal conversion, and a strictly random respondent selection method. The overall response rate was 42% for the 'standard' survey, 71% for the 'rigorous' survey.

Despite the difference in response rates, the two samples were nearly identical and the results of the two polls were very similar. For more than 85 questions concerning media use, lifestyle and a range of social and political issues, only five questions showed statistically significant differences between the two surveys (and in a few instances, significant differences between the two samples seemed to reflect actual changes in public opinion between June, when both surveys began, and August, when the rigorous survey was completed). For most questions (75 out of 85) the responses given by each sample differed by four percentage points or less, and the average difference was 2.7 percentage points.

The one area in which the two polls did differ significantly was racial issues. Respondents in the standard sample were significantly more sympathetic towards blacks and other minorities. This suggests that, while typical media polls may be adequate for most topics, more effort to convert reluctant respondents is required for accurate results from surveys dealing with contentious social issues such as race.

At best, a 50% response rate can be regarded as a rough rule of thumb for a minimum acceptable response rate in survey research. For many surveys (particularly mail surveys) and many variables (particularly attitude and opinion variables) estimates obtained at 50% response rate will be adequate for practical purposes. But this does not apply to all surveys or to all variables in a particular survey. Sometimes the direction, if not the extent, of nonresponse bias can be predicted from knowledge of the survey topic and the population samples. However, the only certain way to reduce the potential for nonresponse bias is to increase response rates.

Increasing Response Rates For Mail Surveys

Despite the often-repeated claim that mail surveys have low response rates, it is possible to consistently achieve response rates of 60% or better for mail surveys of the general public, specific consumer groups and businesses, regardless of the topic investigated. (It is true that many mail surveys do have low response rates but this is because of failure on the part of those conducting them rather than an inherent characteristic of mail surveys.) However, to achieve response rates of this magnitude requires at least two reminder letters, and a
questionnaire and reply-paid envelope included with each reminder.

Relatively small (i.e., 50c or $1) prepaid monetary incentives also increase mail survey response rates. Furthermore, because these incentives also increase the speed of response, they can effectively pay for themselves by reducing the number of reminder letters subsequently required.

More reminders will increase the response to a mail survey but at a diminishing rate. Experience shows that two reminders are usually a reasonable compromise between an acceptable response rate and reasonable survey cost. Alternatively, there is some evidence (Gendall, Hoek & Brennan, 1998) that using a $1 prepaid incentive and one reminder can produce a similar response rate to the same survey with no incentive and two reminders at roughly the same cost. The practical advantage of the former approach is that it reduces the time the survey is in the field by several weeks.

Many other techniques have been used in attempts to increase mail survey response rates. The problem in reviewing them is that the results they produce are invariably inconsistent. Sometimes they seem to work, sometimes they do not. Consequently, unlike reminders and prepaid incentives their effectiveness cannot be generalised. Another problem is that much of the research which purports to test these techniques is either methodologically flawed or based on surveys with very low response rates. What we really want to know is how effective these techniques are when applied to well-conducted surveys. Bearing this in mind, an assessment of the effectiveness of these techniques is as follows.

Pre-notification: Often claimed to be effective, but there is no convincing evidence that pre-notification as such is more effective than an additional contact with respondents. In other words, an advance letter and, say, one reminder, is no more effective than two reminders without an advance letter. The critical factor is the total number of contacts. However, pre-notification does have one practical advantage; it identifies ‘gone-no-addresses’ (GNAs) and probably some refusers. This allows the researcher either to substitute for the GNAs or to save money by not sending questionnaires to GNAs and refusers.

Non-monetary incentives: Pens, stickpins, teabags and discount coupons have all been used as non-monetary incentives, but are less effective than prepaid cash incentives.

Cash draws, lotteries, donations to charity: May sometimes be effective but there is no evidence of any consistent impact on response rate for incentives offering rewards contingent on the return of the questionnaire.

Personalisation: Addressing the questionnaire to the respondent by name, hand-addressing the envelope or signing the covering letter personally seem to have little effect in well-conducted surveys of the general public. Personalisation may be effective in surveys of businesses, but even this is not certain.

Questionnaire length: Other things being equal, shorter questionnaires are probably more effective than longer ones. However, questionnaires of more than 30 pages are often successfully used (and number of pages is not a good measure of questionnaire length anyway).

Cover letter appeals: Some researchers believe response rates can be influenced by
the type of appeal in the covering letter (e.g., help the sponsor, or social utility), but the evidence is not persuasive.

**Other design features**: Other facilitation techniques that have been tried include stamps instead of franking on the outer envelope, different classes of mail, promising anonymity, specifying a response deadline, coloured questionnaires, and different questionnaire cover designs. There is no evidence that any of these techniques consistently increases mail survey response rates.

Despite the general ineffectiveness of techniques apart from repeated contacts and monetary incentives, it may be appropriate and effective to use other techniques in specific situations. For example, a guarantee of anonymity is likely to improve the response rate for a survey on a sensitive subject like sexual behaviour. Thus, although we know what works in general, for a particular survey a researcher needs to make a judgement about the design elements, based on the topic in question, the population surveyed and the organisation undertaking the survey. This may simply mean modifying the tone of the covering letter or its appeal, or the level of language and complexity of the questionnaire, but it could also include decisions on the quality of stationery, or whether an incentive is used.

**Increasing Response Rates For Face-to-Face and Telephone Surveys**

The response rate for face-to-face and telephone surveys depends critically on the interviewers' ability to persuade potential respondents to cooperate. Research has shown that there is considerable variation in response rates among interviewers. Some physical characteristics and attributes of interviewers, such as gender, manner of dress, and vocal characteristics, can make a difference to response rates in certain circumstances (for example, see Fowler & Mangione 1990; Morton-Williams 1993). However, methods for increasing response rates for surveys using interviewers generally involve improving the performance of interviewers in the few crucial moments when they first contact a respondent.

**Interviewer training**

To obtain cooperation from reluctant respondents requires effective social skills, to perceive and adapt to individual situation. Snijkers, Hox & De Leeuw (1996) identified eight factors that describe the tactics experienced interviewers use to obtain cooperation in a face-to-face survey. Among these are, projecting a positive image by social skills, tailoring the introduction, and maintaining communication. 'Tailoring' involves the use of different approaches – in words, behaviour and strategies – for different people. For example, in poor areas interviewers may 'dress down', and vice versa in rich areas. In both cases the same compliance principle – similarity leads to liking – is engaged, but in a different way (Groves & Couper, 1998). 'Maintaining communication' (or 'maintaining interaction') means that successful interviewers avoid a hard refusal by stepping back and keeping the opportunity open to contact the respondent again.

While similar strategies are used by both face-to-face and telephone interviewers, some characteristics of telephone

---

2 The strongest predictor of interviewer response rates is experience (years in the job), but this is confounded by the fact that successful interviewers are more likely to remain in the job.
interviewing mean that it requires extra effort from interviewers. Telephone interviewers have fewer cues, fewer means and less time to tailor than face-to-face interviewers. Telephone interviewers cannot use smiles or gestures, they can only sound enthusiastic and convincing. They cannot show or hand over their official ID or introductory letter to establish their credibility. Furthermore, whereas in face-to-face interviews initial interaction may last for five minutes, in most telephone interviews the decision to cooperate or refuse is made within a minute. Nevertheless, tailoring and maintaining interaction increase the likelihood of cooperation in telephone surveys (Couper & Groves, 1996).

Research has shown that refusals in telephone surveys occur when interviewers cannot answer questions adequately and that a 'conversational' introduction, in which interviewers are allowed to use their own words, produces fewer refusals than scripted standard introductions (reported in Hox, De Leeuw & Snijken, 1997). Research has also shown that particularly non-effective persuasion techniques are repeating the respondent's refusal ("You're not interested"), asking why the respondent won't cooperate ("Why won't you participate?") and countering the respondent's reason for refusing ("Fifteen minutes isn't that long" as a reaction to "That will take too long"). Such reactions make it easy for the respondent to simply repeat their refusal (Dijkstra, 1999). The most effective strategy is to explicitly accept the respondent's objections, but to turn them into positive arguments. For example, "Yes it takes some time, but your opinions are important and we would like to have a clear picture of them". Another effective technique involves making a small initial request ("Let's just try the first few questions"). The purpose of interviewer training should be to learn these effective persuasion techniques and to unlearn non-effective ones (Dijkstra, 1999).

In general, telephone interviewers who speak rapidly, clearly and loudly, and sound confident and successful, have lower refusal rates. And, while a prescribed introductory script is less successful than a 'conversational' one in which interviewers are allowed to use their own words, the introduction should include the survey length, topic and sponsor, and an assurance that the interviewer is not trying to sell anything.

**Interviewer motivation and attitudes**

Apart from training in persuasion techniques in general, the more familiar an interviewer is with the particular survey they are working on, the better their technique, and consequently their response rate, is likely to be (Sturgis & Campanelli, 1995). The key implication of this is that interviewer briefings should concentrate on familiarising interviewers with the nature of the survey they will be administering and on giving them appropriate responses to the sorts of questions they will be asked by reluctant respondents.

Interviewers' performance can also be influenced by the expectations created among interviewers by what they are told and how they are paid (see, for example, Conway, 1992). Thus it may be possible to improve response rates by rewarding interviewers who have fewer non contacts or refusals, by creating incentives for refusal conversions, and by emphasising in interviewer instructions and training that getting a good response rate is very important. However, as Conway (1992) points out, pay schemes and interviewer instructions need to be carefully considered.
because they can have unintended and, sometimes, undesirable consequences for survey results. Furthermore, it is sometimes difficult to devise a fair bonus system to encourage high response rates because interviewing areas vary, and it is also important not to sacrifice data quality for a high response rate.

The effects of interviewer morale are hard to quantify but most field managers believe it plays an important part in achieving a good response. If interviewers are interested in a survey, convinced that it is of value, and that what is being asked of respondents is reasonable, they are more likely to be enthusiastic and convincing in their work.

Fieldwork management

For probability samples call-backs at different times and on different days will reduce the number of noncontacts and increase the response rate. At least three call-backs is standard practice for most reputable face-to-face surveys. However, Brown (1994) suggests that even non-probability designs, such as quota samples, should involve call-backs. This would require that interviewers were free to work the hours necessary to revisit addresses and that their remuneration package encouraged them to do so.

Groves & Couper (1998) suggest that certain physical characteristics of households can be used to schedule calls so that noncontacts are reduced. For example, they recommend that households with physical impediments to access, such as gatekeepers, should be called first because they generally require more calls to first contact. Similarly, in telephone surveys, when an answering machine is encountered, more frequent calling might be warranted to find a time when the household is not using the machine to trap calls.

Apart from stipulating the minimum number and pattern of calls at an address before it is abandoned, interviewers can be instructed to organise their contact addresses geographically and to coordinate their call-backs and new calls in the same area. Experience in the British General Household Survey has shown that interviewers may make as many as ten calls at an address with little additional expense by sensible scheduling of calls; as a result the noncontact rate is less than 3%, compared with the normal average of around 10% (Morton-Williams, 1993).

The optimum number of diallings for a telephone survey is difficult to determine. On social surveys up to ten calls may be necessary to achieve an acceptable contact rate; for other types of survey where the requirements are less stringent, up to seven calls may be sufficient. The number of calls required can be reduced by making calls at times when people are more likely to be at home; generally, this is in the evenings and on Sunday mornings for surveys of the general public.

Because it is relatively cheap and easy in telephone surveys to make many attempts by redialling a number, noncontact rates are typically much lower than refusal rates, (though the increasing incidence of answering machines is increasing telephone survey noncontact rates). Consequently the difference between face-to-face and telephone survey response rates is mainly due to the higher rates of refusals in telephone surveys. These refusals tend to occur much earlier in the interaction than in face-to-face surveys; less than half of telephone refusals typically come from the selected person, most occur either before
selection or are made on behalf of the selected person. One way to address this problem is to shorten the respondent selection procedure. This means using either the 'next or last birthday' or 'youngest male/oldest female' method of respondent selection, rather than a method (such as the Kish grid) based on a full listing of household members. In fact, there is some evidence that the 'youngest male/oldest female' may produce slightly higher response rates than the 'last/next' birthday method because of the greater efficiency of selecting among household residents who are home at the time of the call (Keeter & Fisher, 1997).

Reissuing noncontacts and refusals to another, usually more experienced, interviewer can increase overall survey response by 4% to 5%, though more commonly the increase is 2% to 3% (Morton-Williams, 1993). Extending the length of the fieldwork period will also increase response rates for both face-to-face and telephone surveys.

**Advance letters**

There is some evidence that an advance letter can boost the response rate for face-to-face and telephone surveys (Morton-Williams, 1993). Most interviewers are certainly in favour of them and this may be the main reason why they are effective. Advance letters are most effective when they are brief and general, rather than detailed, and when the gap between arrival of the letter and arrival of the interviewer is short. The only disadvantage of advance letters, apart from their cost, is that refusers appear to be more adamant after receiving one, and they cannot be used in RDD telephone surveys.

**Incentives**

The results of experiments using incentives in face-to-face and telephone surveys have been inconsistent. However, a review of such experiments by Gebler, Singer, Raghunathan & McGonagle (1996) concluded that incentives are effective in increasing face-to-face and telephone survey response rates, without affecting data quality, though the effect diminishes as the amount of the incentive rises. As for mail surveys, prepayment is more effective than promised payment and cash incentives more effective than gifts or promises of a donation to charity (see Kropf, Scheib & Blair, 1999).

The main problem with financial incentives in face-to-face and telephone surveys is in administering them. However, when a survey requires a lot of work on the part of respondents, such as when respondents are asked to keep a diary, or when the respondent is a doctor, lawyer or other business professional, a financial incentive is appropriate and effective.

**Weighting of Data**

Another way of dealing with the problem of nonresponse is to attempt to correct for it by weighting the responses of those who do respond. One method of weighting is to establish trends from the results of several waves of responses and to adjust for nonresponse by continuing these trends. This method assumes that nonrespondents are more like late responders than early responders. However, there is little or no evidence to support this assumption.

Weighting survey data by known population parameters such as age and sex is also quite common. Gendall & Davis (1993)
showed that differences between estimates based on one call and all calls in a face-to-face survey virtually disappeared when the first call sample was weighted so that its age-sex distribution matched that of the survey population. However, in a similar study, Gijsberts (1993) concluded that differences between weighted and unweighted results disappeared when rounded off. He attributed this lack of effect to the weak relationship between opinions and socio-demographic characteristics.

Another alternative is weighting by at-home behaviour. Ward, Russick & Rudelius (1985) had mixed results when they weighted 'one-call' data by 'at-home' behaviour; in eight out of 12 comparisons, weighted results tended to be more accurate than unweighted data. However, weighting increases the variance of survey estimates and Ward et al concluded that the small gains in accuracy they achieved by weighting were more than offset by the greatly decreased precision of their estimates.

Weighting of survey data is an integral part of a strategy proposed by Bradburn (1992) for dealing with nonresponse. Bradburn has suggested that, rather than make the maximum effort affordable to complete as many sample cases as possible, researchers should combine low-cost and high-cost data collection methods. For example, starting data collection with a low-cost method applied to all cases, such as trying each number twice, then selecting a sample of the remaining cases and concentrating the remaining resources on getting a high response rate for these cases. When these subsampled cases are combined with the original sample cases they have to be weighted because their probability of selection is different from the cases given the inexpensive treatment. Bradburn argues that incorporating information on hard-to-get respondents in this way would reduce potential nonresponse bias.

However, if opinions or attitudes or other variables of interest are not related to the weighting variables used, then weighting by these variables may actually make matters worse than not weighting at all. In fact, Maxted has described attempts to correct biased samples by using age-gender weights as having as much practical value as embalming a corpse in the hope of its resurrection (Maxted, 1990). This analogy is a little harsh; weighting can mitigate the effects of nonresponse in some circumstances, but its effects are unpredictable and it will typically increase the error margins of estimates. Consequently, weighting is not a satisfactory solution to the problem of low response rates.

Pre-reCRUIted Panels

Another strategy for dealing with nonresponse is the use of large pre-recruited panels of willing respondents from which demographically representative samples are drawn. This approach has been stimulated

---

3 For example, assume women are easier (on average but not uniformly) to find and interview than men because child care responsibilities mean they are more likely to be at home. Weighting by gender overweights men who are most like women in this respect and underweights women who are most like men.

4 Imputation is another post survey adjustment procedure for dealing with nonresponse. Hot-deck imputation creates a data set with complete data for all respondents by substituting the data from a respondent case that resembles the nonrespondent on certain known variables. However, imputation is more often used for replacing missing data within a respondent record than for replacing missing sample cases.
by the observation that the results achieved from such samples are often similar (or identical) to those from 'scientific' surveys.

This technique involves trolling massive mailing lists, inviting recipients to join a 'panel' of people willing to respond to subsequent surveys. Approximately 1% to 2% of those contacted agree to join the panel, which is subsequently weighted by a few standard demographics. Thus the panel is in effect a very large (70,000 or 80,000) quota sample with a very low response rate.

Panels constructed in this way are biased in predictable ways; they severely under-represent the homeless and the highly mobile. Consequently they are manifestly unsuitable for studying some important social issues (poverty, for example). However, for many other issues samples selected from these panels appear to be as effective as samples selected by 'normal' survey methods (Putnam & Yonish, 1999).

This technique raises some critical issues for survey researchers. Critics of the approach argue that, since the samples selected are not probability samples, estimates from them cannot be extrapolated to the population with any known degree of certainty. Proponents argue that, with refusal rates reaching 40% to 50% in many surveys and overall nonresponse being even higher, the reality is that so-called 'scientific' surveys fall far short of this description. They also point out that the real issue is nonresponse bias rather than nonresponse per se. If nonresponse bias is not related to what we are measuring, then it is not a problem. However, the problem is knowing when this occurs and, whatever else we might say about response rates, they are an indication of potential nonresponse bias.

Nevertheless, the debate over panels has highlighted the fact that sample bias is not the only source of error in surveys, and since total survey budgets are finite, a legitimate question is whether some resources should be shifted to combating other types of error. Concentrating on random sampling and response rate, at the exclusion of other non-sampling error, may not be the optimal approach.

E-mail and Web Surveys

It is often difficult, if not impossible, to calculate the response rate for on-line surveys because the population of potential respondents is unknown. The evidence to date is that e-mail and Web surveys have produced lower response rates than comparable mail or telephone surveys. Nevertheless, some e-mail surveys of particular populations have performed better than mail, and on-line surveys are invariably faster and cheaper than their conventional counterparts (see Schaefer & Dillman, 1999; Weible & Wallace, 1998). However, the obvious problem with on-line surveys is that they are unavoidably and systematically biased as far as the general population is concerned.

For some purposes the fact that the Internet population tends to be significantly younger, better educated, more affluent and contains more males than females may not be a problem. However, on-line polls are more easily biased by self-selection because Internet users decide for themselves whether or not they even want to be included in the sample, much less the final group of participants.

Two American studies, one a pre-election poll conducted in November 1998, the other a poll on the millennium conducted
in April 1999, compared identical telephone and on-line surveys (Flemming & Sonner, 1999). The Internet samples for these studies not only reflected the known differences between Internet users and the American population, they exaggerated them. The result was that, while on some questions the on-line survey results were comparable to identical national telephone surveys, on many others they were very different. More importantly, there were no predictable patterns to these differences. Thus, while e-mail and the World Wide Web may provide a potential solution to the response rate problem for specific groups, it does not offer the same potential for surveys of the general public, at least, not yet.\(^5\)

It seems reasonable to assume that on-line surveying will follow a similar pattern to telephone surveying. Initially, doubts were expressed about the validity of telephone surveying because of lack of universal coverage. Once telephone penetration had increased to more than 90%, this reservation disappeared. In the early years, the telephone and telephone surveys were novelties and this virtually guaranteed a good response rate. However, familiarity, the introduction of telemarketing, answering machines, caller ID and so on have seriously reduced the effectiveness of the telephone as a survey medium. A similar history seems likely for on-line surveying, but with a much shorter 'honeymoon' period. If this is correct, then on-line surveys will provide only a temporary respite from the problem of declining response rates.

Though Internet surveys are still relatively new we already know something about how to achieve good response rates with this medium. Like mail surveys, the response rate for e-mail surveys is increased by reminders. Incentives conditional on completing an on-line survey do not appear to be very effective (Brennan, Rae & Parackal, 1999), and one of the main problems with on-line surveys is respondents prematurely abandoning the survey. Other issues such as questionnaire layout and technical equipment are also important determinants of Web-survey response rates. For example, Dillman & Tortora concluded that a 'plain' Web questionnaire (without colour and HTML tables, which required less transmission time and was presented in a more conventional format) had a higher response rate than an equivalent 'fancy' version. In the future, shorter transmission times and the widespread availability of more powerful browsers may support the use of advanced page layout designs in Web surveys. However, there may always be a gap between what questionnaire designers can create on the Web and what respondents can easily access and answer. (Principles for constructing web surveys are outlined in Dillman, Tortora & Bowker (1998).)

Conclusions

Response rates may or may not be declining in general but for most commercial surveys, particularly telephone surveys, if response rates are not declining it is only because companies are expending more effort than they used to. Thus survey response rates are a problem for market researchers. The problem is the potential for nonresponse bias inherent in any survey that has a response rate of less than 100%.

For practical purposes, a response rate of 50% is a minimum acceptable response rate for many surveys. However, this should
not be taken to imply that at a 50% response rate there is negligible chance of nonresponse bias. It simply recognises that, in many surveys, there is little or no change in most survey estimates as response rates increase from 50% to what are normally regarded as acceptable levels (around 65% to 70%). However, the fact is that even at 'acceptable response rates' there is still the potential for serious nonresponse bias. How serious depends on the objectives of the survey and the relationship between nonresponse and the variable of interest.

The Internet may provide some respite from falling response rates, but on-line surveys are unlikely to provide a permanent solution to the problem. Nevertheless, there are other ways of improving survey response rates; some, such as call-backs or reminders, and monetary incentives, work for all survey modes, others are specific to different modes or situations (for example, guaranteeing anonymity in a survey on a sensitive subject). However, one thing most of these techniques have in common is that they add to either the time or cost of a survey, or both.

Consequently, responding to the problem of nonresponse requires clients to understand the simple equation of survey research (and market research in general), which is that they can have any two of speed, quality or price, but not all three. In a highly competitive market, no research company can afford to offer a service which is better, and thus more expensive, than the market demands. The onus is, therefore, on clients to decide what standard of research they are willing to pay for.

The problem is that without a standard definition of survey response rates it is impossible for clients to know exactly what they are paying for or for research companies with higher response rates to charge a premium for their work. To do this requires a standard call analysis presentation across the market research industry, based on a standard disposition of cases. The AAPOR definitions of response rates would be a logical model for this. However, clients would need to understand that the response rate for a particular survey will depend on the topic, the population surveyed, the length of the questionnaire, and so on.

The market research industry needs to speak out in defence of high quality survey research, to expose poor research, and to make a strong stand against unethical practices such as frugging, sugging and push polling. One of the heuristics people use in deciding whether to grant a survey request concerns the actor making the request. In telephone surveys in particular, a potential respondent's decision to cooperate or refuse is likely to be heavily influenced by their attitude towards market research. Thus the industry must distance itself from telemarketing, and it must continue to explain to the public the relevance and significance of market research and to emphasise assurances about the confidentiality of responses. Market researchers and clients also need to remember that the pool of potential respondents is not unlimited; respondents need to be treated ethically and considerately. This means including issues such as respondent burden as a consideration when designing surveys.

Finally, researchers and clients can take some comfort in the knowledge that standard commercial survey methodologies are relatively robust, and that well conducted surveys, with response rates around 50%, are quite adequate for many practical purposes. This does not mean that nonresponse bias is never an issue, or that we should be complacent about declining response rates,
but neither should we respond to the problem of nonresponse by over-emphasising it. The risk in doing the latter is that more important sources of survey error, such as sample coverage or question wording, will be overlooked.

REFERENCES


Surveys. *American Association for Public Opinion Research Conference, Salt Lake City, Utah, USA.*


AN EVALUATION OF POLITICAL POLLING IN A POSTAL VOTE

Malcolm Wright
Marketing Science Centre
University of South Australia

ABSTRACT

This study evaluated the performance of a political poll in a postal vote with weak party affiliations. It used “before” (n = 481) and “after” polls (n = 370) of the same voters in a Mayoral election conducted by postal vote to: (i) determine whether standard political polling still worked in a postal vote with weak party affiliations; (ii) identify the sources of error in such a poll, following the procedures used to analyse the errors in the 1992 UK general election; and (iii) to investigate whether the gap between voting intention and voting behaviour could be explained by some of the attitudes reported in the “before” poll.

The performance of the poll was at least as good as the performance of typical general election polls, validating the extension of standard political polling to postal votes with weak party affiliations. The sources of error found after the 1992 UK general election were found to apply to this Mayoral election, with similar magnitudes, but fortuitously cancelled out to give extremely accurate predictions of the candidates’ standings. The attitudes reported in the first poll by vote switchers and late deciders were unrelated to their final voting behaviour. Interestingly, a vote switching matrix showed that vote switching was far less deterministic than is often thought, and actually followed the stochastic patterns of purchase behaviour typical of consumer goods.

INTRODUCTION

Postal voting is becoming an increasingly common method of election. Absentee voters are given postal ballots in countries as diverse as Britain, India, and Bosnia-Herzegovina, while in New Zealand postal voting is used for all local council and Mayoral elections. Postal voting was also the method used to elect delegates to a recent Constitutional Convention in Australia.

Standard public opinion polls are often used with postal voters, to determine their political opinion and voting intention, and to predict electoral outcomes. These polls are assumed to be just as accurate as the usual general election polls, but in fact the circumstances can be quite different. Postal voting occurs over a longer time period than traditional voting methods, and may involve a different decision making process.

For example, the choice of a local Mayor or delegate to a constitutional convention may be comparatively independent of political party affiliation when compared to a general election. More cognition may occur before a final decision is made, especially as the ballot is likely to be provided to the voter well before it has to be submitted. Participation rates are likely to be lower, so many of those responding to a poll may not actually return the final ballot. These potential differences are sufficient to raise questions about the accuracy of standard polling techniques when applied to postal votes, especially where there are weak party affiliations.
A meaningful assessment of the accuracy of polls of postal voters must also involve a comparison with standard political polling. Standard political polling has had some very public failures in the 1990s. Crewe (1993) and Jowell, et al (1993) discussed the worrying 9% gap between the predictions of the pollsters and the general election results in Britain in 1992. Crewe (1993) demonstrated that the chances that sampling error alone was the cause of these inaccuracies to be less than 1 in 160,000. Wright (1994) subsequently noted similar inaccuracies in New Zealand and Australia. The sources of error behind these failures also need to be identified and considered for polls of postal voters.

Precisely what caused these inaccuracies was the subject of intense interest in Britain in 1992. Two companies, International Communications and Marketing (ICM) and Market and Opinion Research International (MORI) used a "before" and "after" approach. They re-interviewed pre-election poll respondents after the election and compared "before" voting intentions with "after" reports of actual voting behaviour. Similar data was also available from the Social and Community Planning Research / Nuffield College British General Election Study (BGES).

By analysing the ICM and MORI results, Crewe (1993) was able to attribute 2.4% of an 8.5% error to vote switching (a late swing) and late deciders (don't knows), 1.5% to differential support amongst non-disclosers (won't says), and 0.6% to differential support amongst unregistered voters. He ascribed the rest of the error to non-response and sampling frame bias.

Jowell, et al (1993) analysed the BGES data as well as the ICM and MORI results, and concluded that around half of the discrepancy could be explained by: vote switching; differential distribution of late deciders; and differential party support amongst non-disclosers, abstainers, and respondents who were not registered to vote. They attributed the remaining error to sample bias (rejecting the explanation of consistent lying).

Consequently, this study seeks to assess whether standard political polling techniques are applicable in a situation that is rather different from a typical general election. The situation investigated in this study is a Mayoral election, but the results are expected to apply more generally to postal votes with weak party affiliations.

In particular this study seeks to address the following questions:

- Can standard political polling techniques predict the outcome of a Mayoral election conducted by postal vote?

- What are the effects of the errors identified by Crewe (1993) and Jowell, et al (1993) on the accuracy of Mayoral poll predictions? Specifically, how is the accuracy of the poll affected by vote switching, differential turnout, late deciders, and non-disclosers?

- Can the changes between individual voters' intentions and actual voting behaviour be explained by their attitudes towards the performance of the local council, the performance of the incumbent candidate, or by the issues considered most important by voters?

**METHOD**

The "before and after" polling approach used by ICM, MORI, and BGES
was applied to a local council election in New Zealand. The election investigated was the 1995 Mayoralty race in a provincial New Zealand town (population 52,000). There were three candidates all of who were independent of any political party. One of the candidates was the incumbent Mayor (Candidate A). Postal voting for the Mayoralty took place from 4 October 1995 and continued for two weeks. The night on which returns were counted is referred to as “election night”.

An initial telephone poll of voting intentions was carried out in mid September with 481 useable responses obtained. Households were systematically sampled from the local telephone book, and respondent selection within the household was rotated to meet demographic quotas. The demographic profile of the 481 respondents was very similar to the demographic profile of the area obtained from census data.

The respondents to this “before” poll were re-interviewed immediately following the election. Of the original 481 respondents 370, or 77%, were successfully re-interviewed.

In the “before” poll, respondents were asked to:

1. evaluate the performance of the local council (using a 5 point scale from “very good” to “very poor”);

2. indicate a voting intention (with prompting of candidate names if necessary);

3. if they were undecided in their voting intention, indicate which candidate they leaned towards supporting (this is common practice in New Zealand, where up to 30% of voters may initially be undecided in general election polls);

4. to evaluate the performance of the incumbent candidate as Mayor (using a five point scale from “very good” to “very poor”);

5. to identify the most important issue in the election.

The question on the most important issues was not pre-coded. Rather, the entire range of responses was examined, and eleven categories developed to which respondents were subsequently assigned.

In the “after” poll, respondents were asked:

1. whether they cast their postal vote in the recent Mayoral elections;

2. who they finally voted for.

RESULTS AND DISCUSSION

Predictive Ability of the “Before” Poll

Table 1 demonstrates that the initial poll of 481 people was extremely accurate in its prediction of the election night results. This gives considerable credence to the practice of political polling in postal votes with weak party affiliations.

The margin of error (at the 95% confidence level) for “Decided and leaners” in Table 1 is about 5.5% for candidates A and B, and about 3.5% for Candidate C; it is somewhat greater for Decided voters. The greatest discrepancy between the poll’s results and the election night results is only
2%, well within the margin of error of the poll.

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>PROPORTION OF DECIDED AND ACTUAL VOTERS (N=481)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Decided voters (n=250)</td>
</tr>
<tr>
<td>Candidate A</td>
<td>45%</td>
</tr>
<tr>
<td>Candidate B</td>
<td>41%</td>
</tr>
<tr>
<td>Candidate C</td>
<td>14%</td>
</tr>
</tbody>
</table>

This result was obtained despite a large number of undecided voters being present in the poll. Forty eight percent of respondents were initially undecided, and when these respondents were asked who they "leaned" towards supporting, only a further 18% expressed a preference, leaving 30% still undecided. Thus, the accuracy of the "before" poll is rather surprising. There are several potential explanations.

- Undecided voters may have voted for different candidates in the same proportions as decided voters. This seems unlikely given the results of Jowell, et al (1993) and Crewe (1993).

- Undecided voters may not have voted at all, removing this source of error. This is an attractive idea, and if true it could indicate that undecided voters could be ignored in polls of elections with lower turnouts, such as local council and Mayoral elections.

- There may have been counterbalancing sources of error which fortuitously produced an accurate prediction. The subsequent analysis demonstrates that this is in fact the case.

A Comparison of the "Before" and "After" Polls

To enable an investigation of errors between the "before" and "after" polls, those respondents who were not successfully re-contacted were excluded from the data. The resulting estimates of "before" voting intention and "after" reported voting behaviour are presented in Table 2. The margin of error (at the 95% confidence level) is just over 6% for candidates A and B, and about 4.5% for Candidate C.

<table>
<thead>
<tr>
<th>TABLE 2</th>
<th>PROPORTION OF DECIDED/ACTUAL VOTERS (N=370)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&quot;Before&quot; decided &amp; leaners (n=261)</td>
</tr>
<tr>
<td>Candidate A</td>
<td>42%</td>
</tr>
<tr>
<td>Candidate B</td>
<td>45%</td>
</tr>
<tr>
<td>Candidate C</td>
<td>13%</td>
</tr>
</tbody>
</table>

Table 2 demonstrates a slightly larger gap between election night results and reported voting behaviour. The UK analyses ascribed such gaps to sampling error,
sampling frame bias, and non-response bias. It is not possible to disentangle these effects in Table 2, although as it happens the differences can still be explained by sampling error alone.

On the other hand, there are notable differences in the “before” and “after” results for candidates A and B. These cannot be explained by sampling error, sampling frame bias, or non-response bias, as exactly the same respondents are used in each poll. There must be other problems in the “before” poll, which resulted in an under-prediction of support for Candidate A, and an over-prediction of support for Candidate B. It is precisely these problems that this research seeks to investigate.

The Sources of Errors in the “Before” Poll

To help analyse the reasons for the predictive errors of the “before” poll, and in particular the effects of vote switching, differential turnout, late deciders, and non-disclosers, Table 3 presents a cross-tabulation of “before” voting intention and “after” reported voting behaviour. This cross-tabulation is presented as percentages of those expressing each voting intention; in other words, the rows will sum to 100% (plus or minus some rounding error). The equivalent raw data is provided in Appendix 1.

TABLE 3  COMPARISON OF VOTING INTENTIONS AND REPORTED VOTING BEHAVIOUR

<table>
<thead>
<tr>
<th>Voting Intention</th>
<th>n</th>
<th>% voted Candidate A</th>
<th>% voted Candidate B</th>
<th>% voted Candidate C</th>
<th>% did not vote</th>
<th>% non-disclosers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Decided</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candidate A</td>
<td>(81)</td>
<td>75</td>
<td>6</td>
<td>2</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Candidate B</td>
<td>(90)</td>
<td>16</td>
<td>61</td>
<td>4</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>Candidate C</td>
<td>(28)</td>
<td>11</td>
<td>7</td>
<td>61</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td><strong>Leaners</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candidate A</td>
<td>(28)</td>
<td>64</td>
<td>4</td>
<td>0</td>
<td>25</td>
<td>7</td>
</tr>
<tr>
<td>Candidate B</td>
<td>(27)</td>
<td>4</td>
<td>52</td>
<td>11</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>Candidate C</td>
<td>(7)</td>
<td>0</td>
<td>14</td>
<td>43</td>
<td>29</td>
<td>14</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Late deciders</td>
<td>(96)</td>
<td>35</td>
<td>14</td>
<td>10</td>
<td>26</td>
<td>14</td>
</tr>
<tr>
<td>Non-disclosers</td>
<td>(13)</td>
<td>14</td>
<td>14</td>
<td>7</td>
<td>31</td>
<td>31</td>
</tr>
</tbody>
</table>

Table 3 provides some extremely interesting results on vote switching. It is easy to think of voting as a deterministic behaviour, with strong loyalty to a candidate, and some switching to a preferred candidate late in the campaign. This does not, however, accord with the level of vote switching demonstrated in Table 3. The bolded diagonals show that the candidates retained only 75%, 61%, and 61% respectively of decided voters, and 64%, 52%, and 43% respectively of leaning voters. Switching occurred not only between candidates, but also between those...
intending to vote and not voting, and disclosure and non-disclosure.

In fact Table 3 looks remarkably like the brand switching matrices observed for grocery brands. Furthermore, while Table 2 shows that voting support is reasonably stable at an aggregate level, Table 3 shows that it is quite variable at an individual level. This is precisely the pattern found for attitudes towards brand attributes by Castleberry, et al. (1994). Also, the patterns of support for the candidates in Table 3 follow the well known Double Jeopardy pattern (Ehrenberg, et al. 1990, McPhee 1963), with the more popular candidates retaining proportionally more of their support.

It seems that voting intentions and behaviour follow exactly the same patterns as are found in consumer attitudes towards grocery brands. Although double jeopardy has been observed in attitudes towards politicians (Ehrenberg, 1991), its presence in voting intention and behaviour is a new, and interesting, result.

Table 3 also shows that Candidate B’s voters had a lower turnout, and that the late deciders split heavily in favour of Candidate A. There do not appear to be any obvious effects from non-disclosers. The effects of all the errors arising from vote switching, differential turnout, late deciders, and non-disclosers are summarised in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>Candidate A</th>
<th>Candidate B</th>
<th>Candidate C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vote switching</td>
<td>-4.5</td>
<td>5.2</td>
<td>-0.6</td>
</tr>
<tr>
<td>Differential turnout</td>
<td>0.8</td>
<td>-1.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Late deciders</td>
<td>-3.9</td>
<td>4.8</td>
<td>-0.9</td>
</tr>
<tr>
<td>Non-disclosers</td>
<td>0.0</td>
<td>0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>Total</td>
<td>-7.5</td>
<td>8.8</td>
<td>-1.2</td>
</tr>
</tbody>
</table>

These results are similar to those of Crewe (1993) and Jowell, et al (1993). Most of the discrepancy between the “before” and “after” polls is explained by vote switching and late deciders, and other factors have a minor effect. On the other hand, the effects of vote switching and late deciders are slightly greater in magnitude than in the UK.

As it happened, the combined effects of sampling error, sampling frame bias, and non-response bias compensated for the errors arising from vote switching and late deciders. This made the “before” poll far more accurate than it otherwise might have been. Although on the surface the “before” poll appears startlingly accurate, Table 4 demonstrates that, in reality, its performance is comparable to other standard political polls.

**Can the Gap Between Voting Intention and Behaviour be Explained?**

The major errors in Table 4 arose from vote switching and late deciders. If the gap between the voting intention and behaviour of these groups could be explained, it might be possible to provide a more accurate prediction of the election outcome. In line with standard practice, respondents had
been asked to evaluate the performance of the council, the performance of the Mayor, and the most important issue in the election; so it seemed reasonable to ask whether differences on these variables may be associated with differences in subsequent voting behaviour.

Unfortunately, the short answer is that they were not. The details of this result are provided in the rest of this section.

Vote switchers, or those who switched from supporting one candidate to voting for another, were selected from the data by excluding don’t knows, non-voters, non-disclosers, and those whose voting behaviour matched their voting intention. The resulting dependent variable, reported voting behaviour, was therefore nominal with three categories (switched to Candidate A, B, or C).

A Kruskal-Wallis test was used to determine whether these three groups of switchers had varied in their assessment of the performance of the council or the Mayor. This test is similar to an ANOVA, but is applied to ordinal response categories, and uses a chi-square statistic rather than an F-test. A crosstabulation and chi-square test was also undertaken of the three groups of switchers with the most important issue (a nominal variable with 11 categories). The results (all clearly non-significant) are summarised in Table 5.

<table>
<thead>
<tr>
<th>Issue</th>
<th>Council Performance</th>
<th>Mayoral Performance</th>
<th>Main Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Kruskal-Wallis</td>
<td>Kruskal-Wallis</td>
<td>Chi-Square</td>
</tr>
<tr>
<td>n</td>
<td>36</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>( \chi^2 )</td>
<td>2.16</td>
<td>0.23</td>
<td>24.95</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>p</td>
<td>.34</td>
<td>.89</td>
<td>.13</td>
</tr>
</tbody>
</table>

A more powerful analysis was undertaken by assuming the data were metric, and using the evaluations of the performance of the council and the Mayor as independent variables in a multiple discriminant analysis (n = 36) to attempt to explain final voting behaviour amongst switchers. However the discriminant functions were non-significant, and only correctly classified 42% of cases.

Late deciders were then selected from the data by excluding non-voters, non-disclosers, and all those who expressed a voting intention in the "before" poll. The same analyses were then carried out on this group, again yielding non-significant results as can be seen in Table 6.

25
TABLE 6  VARIABLES AFFECTING LATE DECIDERS’ FINAL CHOICE

<table>
<thead>
<tr>
<th>Issue</th>
<th>Council Performance</th>
<th>Mayoral Performance</th>
<th>Main Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Kruskal-Wallis</td>
<td>Kruskal-Wallis</td>
<td>Chi-Square</td>
</tr>
<tr>
<td>n</td>
<td>57</td>
<td>57</td>
<td>57</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>2.90</td>
<td>3.87</td>
<td>20.70</td>
</tr>
<tr>
<td>df</td>
<td>2</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>p</td>
<td>.23</td>
<td>.14</td>
<td>.41</td>
</tr>
</tbody>
</table>

A multiple discriminant analysis was applied using the evaluations of the performance of the council and the Mayor (n = 57). However the discriminant functions were again non-significant, and only correctly classified 46% of late deciders to their actual voting behaviour.

Consequently, the attitudes towards the performance of the council and Mayor, and the identification of the most important issue do not provide an explanation of the eventual voting behaviour of vote switchers and late deciders. This is a little surprising, as these sorts of questions are commonly asked in political polls, and are expected to be useful to campaign managers.

To determine whether these variables had any explanatory power at all, a similar analysis was undertaken for all those who reported their voting behaviour (including non-voters, but excluding all non-disclosers). The results are presented in Table 7.

TABLE 7  VARIABLES AFFECTING ALL VOTERS’ FINAL CHOICE

<table>
<thead>
<tr>
<th>Issue</th>
<th>Council Performance</th>
<th>Mayoral Performance</th>
<th>Main Issue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method</td>
<td>Kruskal-Wallis</td>
<td>Kruskal-Wallis</td>
<td>Chi-Square</td>
</tr>
<tr>
<td>n</td>
<td>341</td>
<td>341</td>
<td>341</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>32.86</td>
<td>69.07</td>
<td>54.90</td>
</tr>
<tr>
<td>df</td>
<td>3</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>p</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

All comparisons reported in Table 7 are significant at the 1% level, indicating that these are reasonable variables to consider in an election of this sort, although they do not happen to explain vote switchers’ or late deciders’ behaviours. Of course there may be an effect on vote switchers and late deciders which in not detectable with the sample size used; however the sample is not unreasonably small compared to many commercial surveys of this sort, so any such undetected effect is unlikely to have much practical importance.

To complete the comparison, the variables in Table 7, together with voting intention, were included in a stepwise multiple discriminant analysis (n = 341) to
attempt explain the choices of all final voters and non-voters (non-disclosers were excluded). The dependent variables were voting intention (expanded into dummy variables), evaluation of the performance of the council, and evaluation of the performance of the Mayor. The variables entered into the discriminant analysis by the stepwise procedure were intention to vote for candidates A, B and C, and evaluation of the performance of the Mayor. These gave three significant discriminant functions, but correct classification of only 58% of cases. This is not particularly impressive, and the predictions delivered by the discriminant classifications (42%, 46%, and 12% of votes respectively) are practically identical to the use of voting intentions (a much simpler technique) as a predictor.

SUMMARY AND CONCLUSION

Despite the differences between general elections and postal votes with weak party affiliations, political polls can be successful in predicting the outcome of a postal vote for a Mayoral election. Given that polls of postal votes work for elections with weak party affiliations, they can also be expected to work for elections with stronger party affiliations. However the sources of error found in general election polls are still present; in particular: a combination of sampling error, sampling frame bias and non-response bias; vote switching between candidates; and late deciders splitting in favour a particular candidate. As these sources of error happened to compensate for each other in this study, the resulting prediction was extremely accurate.

Interestingly, common conceptions of vote switching (loyalty with some switching to a preferred candidate) appear incorrect. The data in this study demonstrate that vote switching occurs between all candidates, and also between voting and not voting, and disclosing and not disclosing. This switching appears to follow well established patterns of behaviour seen in consumer goods, including double jeopardy, and individual variation with aggregate stability.

Unfortunately, there is no obvious way to explain these patterns of vote switching, or to explain the decisions made by late deciders. Variables such as the evaluation of the performance of the council, the evaluation of the performance of the Mayor, and the identification of the most important issue, were not associated with vote switching or late deciders’ choices.

In conclusion, researchers seem to have as much justification for using standard political polling techniques in postal votes as they do in general elections. Nevertheless, vote switching and late deciders are still important sources of error, and further research is required to try to explain the direction and magnitude of these errors. Political decision makers and advisers would also do well to be aware of what the patterns of vote switching actually look like, and to adjust their promotional strategies accordingly.

REFERENCES


### APPENDIX 1 VOTE SWITCHING MATRIX - RAWDATA

<table>
<thead>
<tr>
<th>Voting Intention</th>
<th>( n )</th>
<th>Voted Candidate A</th>
<th>Voted Candidate B</th>
<th>Voted Candidate C</th>
<th>Did not vote</th>
<th>Non-disclosers</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Decided</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candidate A</td>
<td>(81)</td>
<td>61</td>
<td>5</td>
<td>2</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>Candidate B</td>
<td>(90)</td>
<td>14</td>
<td>55</td>
<td>4</td>
<td>16</td>
<td>1</td>
</tr>
<tr>
<td>Candidate C</td>
<td>(28)</td>
<td>3</td>
<td>2</td>
<td>17</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Learners</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Candidate A</td>
<td>(28)</td>
<td>18</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Candidate B</td>
<td>(27)</td>
<td>1</td>
<td>14</td>
<td>3</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Candidate C</td>
<td>(7)</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Late deciders</td>
<td>(96)</td>
<td>34</td>
<td>13</td>
<td>10</td>
<td>25</td>
<td>14</td>
</tr>
<tr>
<td>Non-disclosers</td>
<td>(13)</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>
The Reliability and Validity of Objective Measures of Customer Service: “Mystery Shopping”

John Dawes
Byron Sharp
University of South Australia
South Australia

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.

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).

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

---

1 Also known as shadow shopping, or phantom shopping.
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 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.
### TABLE I QUESTIONNAIRE ITEMS

<table>
<thead>
<tr>
<th>Item</th>
<th>Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendliness</td>
<td>metric</td>
</tr>
<tr>
<td>Interest</td>
<td>metric</td>
</tr>
<tr>
<td>Enjoyable encounter</td>
<td>metric</td>
</tr>
<tr>
<td>Confidence in answering a question</td>
<td>metric</td>
</tr>
<tr>
<td>Positive reflection on the principal</td>
<td>metric</td>
</tr>
<tr>
<td>Cleanliness and tidiness of the outlet</td>
<td>metric</td>
</tr>
<tr>
<td>Immediate service or wait</td>
<td>categorical</td>
</tr>
<tr>
<td>Form of greeting</td>
<td>categorical</td>
</tr>
<tr>
<td>Friendly verbal exchange</td>
<td>categorical</td>
</tr>
<tr>
<td>Seller ending exchange with well wishing</td>
<td>categorical</td>
</tr>
<tr>
<td>Mentioning of other products by the seller</td>
<td>categorical</td>
</tr>
<tr>
<td>Prompting for purchase/cross selling</td>
<td>categorical</td>
</tr>
<tr>
<td>Product knowledge</td>
<td>categorical</td>
</tr>
<tr>
<td>Technical problems</td>
<td>categorical</td>
</tr>
<tr>
<td>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</td>
<td>metric, rating out of 100.</td>
</tr>
</tbody>
</table>

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 (e.g. 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.

<table>
<thead>
<tr>
<th>Raters 1 &amp; 2 - 15 outlets each</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raters 1 &amp; 3 - 15 outlets each</td>
</tr>
<tr>
<td>Raters 2 &amp; 3 - 15 outlets each</td>
</tr>
<tr>
<td>Raters 2 &amp; 4 - 15 outlets each</td>
</tr>
</tbody>
</table>

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 Cool 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 Tables III and IV.
TABLE III INTERRATER RELIABILITY VALUES

<table>
<thead>
<tr>
<th>Item</th>
<th>ICC value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q. 4</td>
<td>0.90</td>
</tr>
<tr>
<td>Q. 6</td>
<td>0.90</td>
</tr>
<tr>
<td>Q. 7</td>
<td>0.86</td>
</tr>
<tr>
<td>Q. 11</td>
<td>0.76</td>
</tr>
<tr>
<td>Q. 12</td>
<td>0.55</td>
</tr>
<tr>
<td>Q. 15</td>
<td>0.81</td>
</tr>
</tbody>
</table>

TABLE IV RELIABILITY - CATEGORICAL QUESTIONS

<table>
<thead>
<tr>
<th>Item</th>
<th>Average Proportion Agreement</th>
<th># categories</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q. 1</td>
<td>0.87</td>
<td>7</td>
<td>0.90</td>
</tr>
<tr>
<td>Q. 2</td>
<td>0.83</td>
<td>3</td>
<td>0.85</td>
</tr>
<tr>
<td>Q. 3</td>
<td>0.67</td>
<td>5</td>
<td>0.75</td>
</tr>
<tr>
<td>Q. 5</td>
<td>0.87</td>
<td>3</td>
<td>0.89</td>
</tr>
<tr>
<td>Q. 8</td>
<td>0.98</td>
<td>2</td>
<td>0.99</td>
</tr>
<tr>
<td>Q. 9</td>
<td>1.0</td>
<td>3</td>
<td>1.0</td>
</tr>
<tr>
<td>Q. 10</td>
<td>0.95</td>
<td>3</td>
<td>0.98</td>
</tr>
<tr>
<td>Q. 13</td>
<td>0.10</td>
<td>4</td>
<td>1.0</td>
</tr>
</tbody>
</table>

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. Tables 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 V. They show that eleven out of the thirteen variables had a strong association with the global assessment of service quality.

TABLE V MEASURES OF ASSOCIATION BETWEEN SERVICE QUALITY COMPONENTS AND GLOBAL SCORE

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type</th>
<th>Association with overall service quality assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendliness</td>
<td>metric</td>
<td>R=0.73, p&lt;0.001</td>
</tr>
<tr>
<td>Interest</td>
<td>metric</td>
<td>R=0.78, p&lt;0.001</td>
</tr>
<tr>
<td>Enjoyable encounter</td>
<td>metric</td>
<td>R=0.82, p&lt;0.001</td>
</tr>
<tr>
<td>Confidence in answering a question</td>
<td>metric</td>
<td>R=0.52, p&lt;0.001</td>
</tr>
<tr>
<td>Positive reflection on the principal</td>
<td>metric</td>
<td>R=0.38, p&lt;0.001</td>
</tr>
<tr>
<td>Cleanliness and tidiness of the outlet</td>
<td>metric</td>
<td>R=0.40, p&lt;0.001</td>
</tr>
<tr>
<td>Immediate service or wait</td>
<td>categorical</td>
<td>χ²=57, df 12, p&lt;0.0001</td>
</tr>
<tr>
<td>Form of greeting</td>
<td>categorical</td>
<td>χ²=283, df 4, p&lt;0.0001</td>
</tr>
<tr>
<td>Friendly verbal exchange</td>
<td>categorical</td>
<td>χ²=303, df 8, p&lt;0.0001</td>
</tr>
<tr>
<td>Seller ending exchange with well wishing</td>
<td>categorical</td>
<td>χ²=312, df 4, p&lt;0.0001</td>
</tr>
<tr>
<td>Mentioning of other products by the seller</td>
<td>categorical</td>
<td>χ²=23, df 2, p&lt;0.0001</td>
</tr>
<tr>
<td>Prompting for purchase/cross selling</td>
<td>categorical</td>
<td>χ²=6, df 2, p=0.21</td>
</tr>
<tr>
<td>Product knowledge</td>
<td>categorical</td>
<td>χ²=182, df 4, p&lt;0.0001</td>
</tr>
<tr>
<td>Technical problems</td>
<td>categorical</td>
<td>χ²=2, df 2, p=0.46</td>
</tr>
</tbody>
</table>

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 R²'s 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:

\[ Y_t = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4 + B_5 X_5 + B_6 Y_{t-1} + \text{error} \]  

(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 VI.
TABLE VI STANDARDISED REGRESSION (BETA) COEFFICIENTS
(Standard errors in parentheses)

<table>
<thead>
<tr>
<th></th>
<th>Wave 1</th>
<th>Wave 2</th>
<th>Wave 3</th>
<th>Wave 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q4 (friendliness)</td>
<td>.32</td>
<td>.24</td>
<td>.30</td>
<td>.24</td>
</tr>
<tr>
<td>Q6 (interest)</td>
<td>.29</td>
<td>.35</td>
<td>.37</td>
<td>.45</td>
</tr>
<tr>
<td>Q11 (confidence in answering query)</td>
<td>.19</td>
<td>.20</td>
<td>.13</td>
<td>.18</td>
</tr>
<tr>
<td>Q15 (clean and tidy area)</td>
<td>.22</td>
<td>.17</td>
<td>.15</td>
<td>.15</td>
</tr>
<tr>
<td>Q5 (end exchange with well wishing - yes/no)</td>
<td>.05</td>
<td>.08</td>
<td>.07</td>
<td>.10</td>
</tr>
<tr>
<td>Q16 LAGGED</td>
<td>.28</td>
<td>.15</td>
<td>.17</td>
<td>.10</td>
</tr>
<tr>
<td>Adj. R^2 with autoregression term</td>
<td>.80</td>
<td>.73</td>
<td>.77</td>
<td>.76</td>
</tr>
<tr>
<td>Adj. R^2 without autoregression term</td>
<td>.72</td>
<td>.72</td>
<td>.74</td>
<td>.75</td>
</tr>
</tbody>
</table>

Note: all coefficients significant at P < 0.01.

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 retains 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 (i.e., calculated from wave 1) has the ability to predict overall service quality assessments in later data sets (i.e. 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.

**FIGURE 1 BETA VALUES FOR THE FOUR SURVEY ROUNDS (AVERAGE OF 3 ENCOUNTERS EA. ROUND)**

![Beta Values Graph]

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

FIGURE II VARIANCE IN SCORES FROM WAVE TO WAVE - SIX RANDOMLY SELECTED OUTLETS
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:

$$\mu \pm 1.96 \frac{\sigma x} {\sqrt{n}}$$

(2)

Where $\mu$ 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, $\sigma x$ 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 VII.

<table>
<thead>
<tr>
<th>Waves</th>
<th>Number of outlets displaying differences significant at P&lt;0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 &amp; 2</td>
<td>5</td>
</tr>
<tr>
<td>2 &amp; 3</td>
<td>7</td>
</tr>
<tr>
<td>3 &amp; 4</td>
<td>1</td>
</tr>
</tbody>
</table>
This number of cases is under what would be expected by chance. At a 95% confidence level we would expect 228 x 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 VIII.

<table>
<thead>
<tr>
<th>TABLE VIII SIGNIFICANT DIFFERENCES - AGGREGATING SURVEY WAVES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison between aggregated scores for waves 1&amp;2 compared to aggregated scores for 3&amp;4</td>
</tr>
</tbody>
</table>

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 IX. 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.
TABLE IX VARIATION IN SCORES FOR OUTLETS IN ANY ONE SURVEY

<table>
<thead>
<tr>
<th>Range</th>
<th>Percentage of outlets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to less than 5 points</td>
<td>3</td>
</tr>
<tr>
<td>5 to less than 10 points</td>
<td>10</td>
</tr>
<tr>
<td>10 to less than 20 points</td>
<td>28</td>
</tr>
<tr>
<td>20 to less than 30 points</td>
<td>28</td>
</tr>
<tr>
<td>30 to less than 40 points</td>
<td>18</td>
</tr>
<tr>
<td>over 40 points</td>
<td>13</td>
</tr>
</tbody>
</table>

Table IX 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


Morris, Ruby Turner and Claire Sekulski Bronson (1969), "The Chaos of


