The accuracy of self-reported probabilities of giving recommendations

Jenni Romaniuk and Cathy Nguyen  
Ehrenberg-Bass Institute for Marketing Science, University of South Australia  
Robert East  
Kingston Business School

This paper shows that respondents are better at predicting when they won’t give a recommendation than when they will. The main reason for inaccuracy was an over-reliance on past circumstances (past receiving or giving of recommendations) in making future predictions of their own behaviour. Therefore, self-report probabilities are best used as measures of the potential or desire to give a recommendation, rather than predictions of future behaviour. The translation of this potential to behaviour will depend largely on the external environment, which is outside the respondent’s control. To improve the accuracy of aggregate-level predictions of how many people will give recommendations, we suggest reducing the number of those with a high self-reported probability to around 30% of survey estimates.

Introduction

Arndt (1967b) defined word of mouth (WOM) as any interpersonal communication about a product, service or brand occurring between a receiver and a communicator known personally to one another. Prior research has shown that WOM exerts substantial influence on consumer buying behaviour and choice (Arndt 1967a; Charlett et al. 1995; Keaveney 1995). Of all the diverse channels through which a consumer can acquire product-related information, WOM is perceived to have one of the highest levels of credibility and, therefore, impact on consumer behaviour (Katz & Lazarsfeld 1955; Fitzgerald Bone 1995; Mangold et al. 1999). This power has led to an increasing interest in WOM metrics that could potentially predict brand sales (East et al. 2010).
Given that the consumer generates recommendations, attention has turned to these consumers to provide estimates about whether they will give recommendations in the future. Self-generated estimates of giving recommendations are attractive for two reasons. The first reason is that they potentially give some indication of the state of mind of the consumer, as satisfaction with the brand is normally a basis for recommendation. The second reason is that self-generated estimates show the volume of recommendations that is likely to be produced by consumers as a whole, and so give an indication of the reach achieved by consumer recommendations. By encompassing retention and acquisition elements, this metric could potentially capture two important future revenue streams. One of the better-known self-report intentions to give recommendations is the Net Promoter Score (NPS), which is claimed to predict the future performance of a brand/firm (Reichheld 2003).

However, there is a considerable gap between what people say they will do and what they actually do (Theil & Kosobud 1968; Sheppard et al. 1988; Wright & MacRae 2007). It may be that self-reported predictions of recommending suffer from the same inaccuracy issues as purchase intentions.

This study examines the level of accuracy of self-report estimates for future recommending behaviour, and the reasons for inaccuracy between the self-estimate and actual behaviour. Further, we investigate the degree to which self-reports to give recommendations are possibly related to past experiences versus intended future actions. A review of the key literature and the ensuing research questions now follows.

**Literature review**

The ability to predict consumer behaviour has always been of interest to marketers. Early researchers believed that self-reported estimates could successfully forecast behaviour (e.g. Juster 1966; Ajzen 1985). However, the majority of studies within this body of literature have shown that self-reported estimates are only partially fulfilled. Pickering and Isherwood (1974) found that 61% of consumers who stated a 100% likelihood of purchase actually made a purchase. Similarly, eight separate studies by McQuarrie (1988) revealed that, on average, only 42% of respondents performed the behaviour they initially intended. Thus, the accuracy of self-reported intentions or probabilities is substantially lower than the 100% indicated by the respondent’s certainty. However, it seems that self-reported estimates are a good indicator of behaviour in cases where the
consumer’s initial response indicates that they are inclined not to act. For example, McQuarrie’s (1988) research found that 88% of respondents who stated that they were unlikely to purchase, did not purchase. East et al. (2008a) speculate that the asymmetry between intention and non-intention may be because it is easier for a person not to do something, which requires no expenditure of effort, than to do it. An intention to give WOM is more easily acted upon than fulfilling an intention to buy a brand. Being a verbal response, it requires little effort on behalf of the consumer, and no financial outlay.

As noted, one of the more popular measures of WOM is the NPS, which is designed to measure the impact of positive and negative WOM on sales. This metric requires respondents to specify their likelihood of recommending a firm/brand on an 11-point scale. The percentage scoring 0 to 6 (detractors) is subtracted from the percentage scoring 9 or 10 (promoters) to give the score. The passive group (scoring 7 or 8) is excluded.

Despite its wide acceptance by industry, the ability to predict company performance from the NPS has been challenged by several academics (e.g. Chandon et al. 2005; Morgan & Rego 2006; Keiningham et al. 2007; East 2008). The main criticisms, as suggested by East (2008), include those listed below.

- The method used for calculating NPS assumes that ‘one size fits all’. A score for a brand with a 10% share may have a different implication than the same score for a brand with 1% share. Uncles et al. (2010) have considered the way in which WOM volume relates to market share.
- Negative WOM is merely inferred by a low intention to recommend. There is no evidence to suggest that a low likelihood of recommending means that the individual will spread negative WOM. East et al. (2010) report that detractors may give more positive WOM than negative WOM.
- The NPS presumes that negative WOM carries more weight than positive WOM, as seven points are assigned to detractors while only two points are assigned to promoters. However, the empirical evidence is that positive WOM not only occurs more often, but also that it is slightly more influential on consumer intentions than negative WOM (Romaniuk 2007; East et al. 2008a).
- Past research shows that most NWOM is given by users of other brands (East et al. 2007), yet these people are excluded from the NPS sample.
NPS fails to measure the impact of WOM, as respondents cannot assess the impact that their advice has on another person. Impact can be measured by investigating the response to WOM that has been received.

Based on the existing literature, we suggest an additional weakness, which is the NPS’s reliance on self-reported estimates for predicting recommending behaviour. However, very few studies have examined the accuracy of self-prediction measures in the context of recommending behaviour, although measures of intention have been drawn on and are often treated as an equivalent to self-report probabilities. For example, Kumar et al. (2007) reveal that more than half of those who intended to give WOM did not do so. Their study examined WOM behaviour in the context of high-involvement service categories (telecommunications and financial services). We measure it in the context of television programmes.

The first area addressed by this research is to expand the body of literature reporting on actual accuracy levels. Specifically, we ask:

**RQ1**: How accurate are self-report probabilities of giving recommendations?

It is one thing to know something is inaccurate, but to improve the metric we need to understand why any inaccuracy is occurring. Currently, there is little knowledge about the reasons for the discrepancies between self-reported estimates of future behaviour and the actual behaviour undertaken by that person. With regard to buying behaviour, Bemmaor (1995) states that such disparities could result from: (1) the true probability of purchase being inaccurately measured by the scale point, and (2) in the intervening time, people altering their intention or becoming unable to fulfil it. The first inaccuracy has been addressed by researchers testing for scale reliability (Kassarjian & Nakanishi 1967), who recommend that 11-point scales are the most reliable. The second inaccuracy is harder to address. Research by Pickering (1975) revealed that the inconsistency between intentions and purchase arose when respondents changed their minds due to unforeseen circumstances. A similar study conducted by East (1993) found that respondents who failed to act on intended investment behaviour were equally divided between those who had a change of mind and inertia. To the best of our knowledge, no studies have focused on why people fail to act on intended recommending behaviour.
There are some issues inherent in giving a recommendation that make this behaviour distinct from purchase behaviour. First, while it does not require a financial outlay, it does require a degree of personal involvement – as described by Reichheld (2003, p. 48): ‘When consumers act as references, they do more than indicate that they have received good economic value from a company; they put their own reputation on the line.’ This involvement may lead the respondent to revise and tailor his or her own opinions to better suit the information needs of the recipient. A recommendation is also an interaction between two parties, and if no receiver is encountered there will be no WOM. This adds another variable that may interfere in the evolution of behaviour from intention. These factors may lie behind the inaccuracy in self-report probabilities.

This informs our second set of research questions:

**RQ2a:** Why do intenders fail to give recommendations?

**RQ2b:** Why do non-intenders give recommendations?

The final issue this paper addresses is the basis of the self-reported probabilities. When someone gives a ‘6 out of 10’ chance of giving a recommendation for a brand, what is the basis for this estimate? Satisfaction or dissatisfaction is associated with WOM (Anderson 1998), although a study by Mangold et al. (1999) found this was the main catalyst in only 12% of cases. More common catalysts were contextual, when the respondent’s environment produced conversations related to the product (18%) or when the giver of advice perceived that the receiver needed this advice (50%). These environments and receiver circumstances are not a direct part of an intention to give WOM measure. Predicting the likelihood of the appropriate situation arising may be a difficult task for respondents and may not be factored into estimates accurately. The information that will be salient, however, is what has occurred to them in the past. Therefore, the self-reported probabilities may be based on past experiences rather than being estimates of future behaviour. Past experiences include both giving and receiving recommendations, and social validation theory suggests that we will be more ready to give advice on a brand when we have heard others give advice on it (Festinger 1954).

This leads to the following research questions about the drivers of the self-reported probability of giving recommendations:
RQ3a: Are self-reported probabilities more closely related to recommending behaviour in the past (given or received) or recommending behaviour in the future?

RQ3b: Does the prior receipt of a recommendation increase the probability that someone will say they will give a recommendation in the future?

In the following section, we present the method used to answer our research questions.

Research method

Television was the category chosen for this study, as TV shows are known to generate a high level of WOM (Eubank & Fay 2008). Further, few studies of WOM have been in lower-risk and common choice categories (Romaniuk 2007). Six new programmes were chosen to maximise the likelihood that the TV programmes would be discussed. Respondents were randomly recruited via telephone, and were screened to meet the criterion of being aged 16–55 years and to have watched at least one night of television in the past week. There were two stages to the data collection. In the first, 235 people were interviewed about their television habits, past WOM behaviour towards the programmes and likelihood of recommending up to six programmes (depending on which they were aware of) in the subsequent week. Respondents were asked whether they could be re-contacted, but were not told when the re-contact would be; 90% gave permission for this. One week later, 138 respondents were reached – a re-contact rate of 59%, which indicates a low likelihood of non-response error (Gendall 2000). Fewer than 5% of respondents refused to complete the re-contact interview, with the remainder simply unavailable for re-interview.

In this follow-up interview, among other general TV viewing behaviour questions, respondents were asked about their actual recommending behaviour in the past week. The short, one-week, time-frame was chosen so that the memory of the recommending instances would still be fresh, and because discrepancies between attitude and behaviour may increase with the time period that elapses between attitude and behaviour measurement. In total, the re-contact sample comprised 36% males and 64% females. Respondents were from a range of age groups: 16–25 (21%); 26–35 (18%); 36–45 (29%); 46–55 (33%).
The key metrics used for this research are presented in the following sections.

**Self-reported estimates of giving a recommendation in the future**

Respondents were asked to indicate, on an 11-point probability scale, the chance that they would recommend each TV show they were aware of in the subsequent week. To ensure we had sufficient sample for testing between group differences, we split respondents into ‘intender’ and ‘non-intender’ groups, based on their responses. Those who gave a high probability (7–10) were classified as ‘intenders’ and those who gave a low probability (0–4) were classified as ‘non-intenders’. Respondents who indicated a score of 5 or 6 were excluded from the analysis to ensure we had a clear discrimination between groups. The verbally referenced scale used here has been employed in similar studies examining customer intention and behaviour (Day et al. 1991; Wright & MacRae 2007).

**Past recommendation experiences**

Prior to the initial interview, respondents were asked whether or not they had spoken positively about or had recommended each show they were aware of. The respondents’ responses were coded as given PWOM or not. A similar question and procedure was conducted for received PWOM.

**Subsequent recommending behaviour**

In the follow-up interview one week later, respondents were asked whether or not they had actively recommended or said something positive about each show they were aware of during the intervening week between interviews. We included general WOM as well as explicit recommendations as we hoped to compare the wider positive behaviour. The respondents’ responses were coded according to WOM action or non-action.

**Reason for inaccuracy between stated probability and behaviour**

Where the behaviour did not match the self-report probability from the prior week, respondents were asked to specify the reason for this, which was recorded by the interviewers. The verbatim answers were then organised according to common themes. The results, combined with a discussion of the results, are presented in the following section.
Results and discussion

RQ1 was ‘How accurate are self-report probabilities of giving recommendations?’ To address this, we identified and then compared the proportion of ‘intenders’ (those who stated a probability of 7–10 of giving a recommendation) who gave a recommendation, and the proportion of ‘intenders’ who did not recommend. The same was done for ‘non-intenders’ (i.e. those who stated a probability of 0–4). All six TV programmes were included in the initial analysis, though three of these were later removed when we analysed the results at the individual programme level because of low sample size.

Overall, only 30% (n = 16) of respondents who stated a high probability to recommend a TV show ended up doing so (see Table 1). Fewer than half of respondents who stated they were highly likely to give a recommendation for any of the programmes did so, with the highest being Masterchef, where 43% (n = 9) of ‘intenders’ gave a recommendation. We then tested whether this level of inaccuracy was consistent at each probability group (i.e. 7, 8, 9 and 10). Across all six programmes, we found that as the probability of recommending increased (from 7 to 10), the proportion of respondents who recommended in the subsequent week also increased (i.e. only 12% of respondents who stated a 7/10 probability of recommending did, whereas a higher 53% of respondents who stated a 10/10 probability of recommending did). However, these differences between scale points failed conventional statistical significance tests, due to small sample sizes. This pattern, however, reflects positively on the validity of the measure.

A greater degree of accuracy was observed for low probabilities. The majority of respondents (95%, n = 175) who had little or no chance of giving a recommendation did not end up giving a recommendation, while only 5% did give a recommendation (p < 0.05). This proportion did not vary significantly across the three TV shows tested (93% for Masterchef,

Table 1  Respondents who acted on their intention – by programme (%)

<table>
<thead>
<tr>
<th></th>
<th>‘Intenders’</th>
<th></th>
<th>‘Non-intenders’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>% who recommended</td>
<td>Overall</td>
<td>% who did not recommend</td>
</tr>
<tr>
<td>Masterchef (n = 21)</td>
<td>43*</td>
<td>Masterchef (n = 83)</td>
<td>93*</td>
</tr>
<tr>
<td>Generation (n = 20)</td>
<td>30*</td>
<td>Generation (n = 58)</td>
<td>91*</td>
</tr>
<tr>
<td>Castle (n = 6)</td>
<td>17*</td>
<td>Castle (n = 50)</td>
<td>98*</td>
</tr>
<tr>
<td>Unweighted average</td>
<td>30</td>
<td>Unweighted average</td>
<td>95</td>
</tr>
</tbody>
</table>

*p < 0.05: significant difference between the proportion of (non)intenders who gave WOM and the proportion of (non)intenders who did not give WOM.
91% for Talking about your Generation and 98% for Castle). Using Z-tests, we examined the difference in accuracy between the percentages of ‘intenders’ who gave a recommendation and the percentages of ‘non-intenders’ who did not give a recommendation. The difference was significant at \( p < 0.05 \) overall for two programmes. Therefore, those who have a low probability of recommending are more likely to fulfil their self-predictions than those who express a high probability of recommending. Thus, answering RQ1, we find that the intention measure is not very accurate but that it does seem to discriminate, with those giving higher scores being more likely to recommend.

We next investigated two types of inconsistency. The first centred on those with a high probability who did not report giving a recommendation about the programme. This was on average 70% of respondents in this group. The second inconsistency was where those with a low probability did give a recommendation in the following week. This was, on average, 5% of respondents in this group. We analysed the verbatim responses provided by these people when asked about the reasons for the inconsistency between their estimates and their behaviour. All six TV programmes were included in this analysis.

RQ2a was ‘Why do intenders fail to give recommendations?’ In total, 44 comments were available for this analysis. There were two dominant reasons for not recommending a TV show despite initially high estimates. The first was due to unforeseen circumstances, which comprised about half of responses (e.g. ‘I was too busy last week’ or ‘Did not end up watching the show, so couldn’t talk about it’). The second is due to a lack of perceived opportunity, which made up about one-third of responses (e.g. ‘Haven’t seen anyone to talk about it’ or ‘No one was interested’). The remainder reported that they simply forgot to recommend.

These findings are consistent with past research into the (in)accuracy between purchase intent and behaviour (Pickering 1975). This result suggests that, in order to generate a recommendation, a consumer may need to be in the market for a ‘conversation’, such that they perceive that there is a reason to talk about the brand. If this circumstance does not arise, then the recommendation will not be given, despite the best of intent. This partially explains the error in WOM estimates.

RQ2b was ‘Why do non-intenders give recommendations?’ There were 14 usable responses for this analysis. Three main reasons emerged to explain why ‘non-intenders’ ended up giving WOM, each reflecting about one-third of responses. The first was due to an unexpected change in preference (e.g. ‘I’m starting to like this show more’). This suggests a change in one
of the factors that underpinned their initial estimates, as suggested by Bemmaor (1995). The second reason for recommendations being given by ‘non-intenders’ was when the subject or genre of the TV show arose during general discussion (e.g. ‘Was cooking with somebody and we started talking about Masterchef’). This finding is in keeping with past literature on WOM drivers, which found that ‘coincidental conversation’ often triggers the giving of WOM (Mangold et al. 1999; Nguyen 2009). The third reason was a request for information by others or interaction with a person whom the giver knew was interested in the show or genre of the show (e.g. ‘My wife asked me about it’). This is consistent with prior research that found ‘concern for others’ or ‘altruism’ to be a motivation behind the giving of WOM (Mangold et al. 1999; Mazzarol et al. 2007). All three of these reasons are difficult to predict by a consumer at the time of giving the estimate, thereby explaining why recommendation may be given, despite an initially low probability. It should be noted that this occurrence was relatively rare, which suggests people more often anticipate circumstances that might occur and then don’t arise, than have circumstances arise that were not anticipated. To understand this more, we now address RQ3a and b, which are about the influence of past WOM experiences versus future recommendation behaviour as drivers of the self-report estimates.

To address these research questions, we conducted multiple linear regression using the self-reported probability of giving a recommendation as the dependent variable, and past given recommendations, past-received recommendations and future given recommendations as the independent variables. This allows us to determine which of the independent variables has the strongest relationship with the self-reported probability of recommending. The temporal data collection, as well as the one-week, time-bounded nature of both the estimates and the behaviour collection means that we give the respondent the best opportunity to provide accurate responses. This analysis was conducted across all programmes.

The results show that all three variables were statistically significantly related to the self-reported probability (adj. $R^2 = 0.47, p < 0.001$). Past given recommendations had the strongest relationship ($\beta = 0.48$); followed by past received recommendations ($\beta = 0.26$) and then future given recommendations ($\beta = 0.19$). Therefore, current self-reported probabilities of recommending were more strongly related to past experiences with recommendations (given or received) than to the respondent’s actual future WOM behaviour. This suggests that, when giving estimates of their future recommendations, respondents are factoring in their own past giving of WOM and whether someone else has recommended or spoken positively
about the brand to them. This focus on the past inflates respondents’ estimates, as past experience may not reflect the probability of such circumstances reoccurring. Supporting this idea is that error in prediction was greater for people who thought they would give a recommendation and then did not. These people thought that a past circumstance would recur but, as suggested by the follow-up verbatims, it did not and so they did not give their (expected) recommendation. The reverse case, where a circumstance suddenly arose to give a recommendation, was less common. This suggests that there may be people ‘primed and ready’ to give WOM, which is what this metric can identify, but that this readiness converts into action only if the right circumstances arise.

Conclusions and implications

We find in this study that self-report probabilities to give recommendations about brands are inaccurate. Those who claim to have a higher probability of doing something often do not do it; to a much lesser extent, some of those who did not plan to recommend, do. This is in line with what was found by East (1993), Pickering and Isherwood (1974), McQuarrie (1988) and Kumar et al. (2007) in the context of intentions to buy/invest. Low probabilities were shown to accurately predict subsequent behaviour, which supports the assertion that it is easier for a person not to do something than to do it. When recommendation is predicted, the inaccuracy between self-report probabilities and behaviour can be explained by the heavy reliance on past behaviour to give estimates. This reliance on past behaviour means that consumers do not take into account unforeseen changes in circumstances and lack of future opportunity. Consistent with much of prior research on WOM antecedents (Mangold et al. 1999; Mazzarol et al. 2007), ‘non-intenders’ gave a recommendation when their preference for the TV show unexpectedly changed, when the topic arose during general discussion, and when they perceived that it was of interest to the information receiver. The idea that this reassessment can happen in a very short time, in this case one week, should be of concern to those using recommendation self-report estimates (such as the NPS) for long-range predictions. Samson (2006) proposes that the NPS is a useful indicator among current customers, in that promoters could be more likely to repurchase or recommend the brand than detractors. Therefore, growth in this metric may be an indicator of improved company performance. Our findings do support this in that a stronger intention to give WOM did indeed come from
those that would be classed as promoters. However, our results also imply that it may not be as effective in predicting actual WOM behaviour, which means that the metric will not be a good predictor of acquisition through current customers recruiting new customers.

Our findings shed light on the contributors of such self-reports and how they should be used. People are more likely to base these self-estimates on their behaviour in the past, and the recommendations they have experienced from others. Therefore, it captures potential or motive. But the desire to give a recommendation is only part of what stimulates actual WOM, and opportunity is also required. Unless the measure is modified to take into account the probability of being in a situation where a recommendation is warranted, it will provide a poor prediction of behaviour. That this measure indicates desire to give WOM explains why NPS is highly correlated with other brand attitude measures such as satisfaction (East et al. 2010). One way of doing this is to measure propensity, given certain circumstances and separately the perceived likelihood of those circumstances. For example, ‘If the topic of xxxx came up in your conversations, would you recommend watching it’ and ‘How likely is it that the topic of xxxx will come up in your conversations in the next week?’ The second possible approach would be to make separate estimates of the probabilities that each of the key environmental triggers will arise in a specified time-frame – for example, ‘How likely do you think it is that someone else will ask you to recommend a TV show in the next week?’ These probabilities can then be used to weight the responses given, either at aggregate or individual level. The danger is that this approach may improve accuracy at the expense of simplicity. The simplicity of measures like the NPS is a major source of their appeal. Which of these is the best approach is an important topic for future research. At this stage, as a minimum modification to improve accuracy, we suggest multiplying the estimates from those with a high intention to give WOM by 0.3 to obtain more accurate estimates of WOM penetration.

WOM involves two parties: the giver and the receiver. The giver needs to give the WOM for it to have an impact on the receiver. Therefore, the role of metrics such as the NPS depends greatly on whether it is important for the giver to actually give the WOM, or whether it is simply enough for them to want to, or be willing to, should circumstances arise. If the metric is simply an indicator of likely customer retention, then intention to give is probably enough. However, then metrics like the NPS compete directly with metrics like satisfaction as measures of customer base sentiment.
Limitations and future research

The key limitations of this research are that it is a single study with a limited number of TV programmes being investigated. The act of asking respondents to report their probability of giving WOM during the prior week may have also influenced the likelihood of them engaging in the intended behaviour (as per Chandon et al. 2005). However, given that our research points more to inaccuracy rather than accuracy, this increases the validity of our conclusions about inaccuracy.

An avenue for further investigation would be to examine how the results from this study would differ over an extended time period. It may be that an intended recommendation might actually be delivered in a future time period. Further, it could be possible that asking consumers about negative WOM may be more indicative of how satisfied they are and their willingness to be brand advocates, as proposed by Samson (2006). As such, an extension of this study would be to examine negative WOM in the same manner, as its scarcity compared to positive WOM may lead to more accurate intentions (Romaniuk 2007).

References


**About the authors**

Jenni Romaniuk is an Associate Research Professor of Brand Equity and Associate Director (International) at the Ehrenberg-Bass Institute, University of South Australia. Her research interests are Brand equity metrics, Brand salience, Word-of-mouth and Advertising effectiveness. In addition to her work with corporations such as Unilever, Colgate Palmolive, ITV, Mars and The Edrington Group, she has been awarded funding from the Australian Research Council to research the influence of word-of-mouth and advertising on how viewers learn about new television programmes.

Cathy Nguyen is a Research Associate at the Ehrenberg-Bass Institute for Marketing Science, based at the University of South Australia. Her research focuses on word-of-mouth communication, specifically in areas relating to the content of word of mouth and how its behaviour and effects differ in online and offline environments. She is also interested in the role that word of mouth plays in the adoption of entertainment products, particularly films and television programmes.


Address correspondence to: Assoc. Prof Jenni Romaniuk, Ehrenberg-Bass Institute for Marketing Science, University of South Australia.
Email: Jenni.Romaniuk@MarketingScience.info