Testing the accuracy of Verbal Probability Scale for predicting short -term brand choice

Jenni Romaniuk

In this research the predictive accuracy of the Verbal Probability Scale (VPS) in predicting four-week brand choice was compared to benchmarks from prior research and past behaviour. The results show that the measure of past behaviour provided the most accurate predictions at both brand and individual level. The exception to this was for the smallest brand in the market, where the VPS was more accurate. This suggests that, at least in this context, past behaviour is a better measure of future behaviour, in that it is the most accurate and the simplest. The findings also highlight that past behaviour should also be an integral part of any study looking to test the predictive accuracy of any measure.

Keywords: Verbal Probability Scale, Juster Scale, past behaviour, brand purchase, predictive accuracy

Introduction

The Verbal Probability Scale (VPS) was developed as a version of the Juster Scale that can be administered via telephone (Brennan, Esslemont et al. 1994). While probabilistic scales have been shown to be more accurate than intentions measures (Day, Gan, Gendall & Esslemont. 1991), the majority of research has been undertaken for product category purchases and/or for purchases over a 12 month or more time period. There has only been limited research that has examined the prediction of shorter-term purchases in three months or less (eg, Clawson 1971) and even fewer studies that have attempted to predict purchases of specific brands, rather than just the product category. In this research, the accuracy of customer’s responses to VPS questions about brand purchases in a short time frame is examined.

For brand managers, prediction of brand choice is arguably more important than the prediction of product category purchases. This brand choice prediction becomes even more difficult in markets where a single brand offers multiple products. Prediction of brand purchase requires going through all of the possible product categories individually, with all of the complexity and confounding errors that would be associated. This is particularly relevant in retail environments where the brand is the destination and the specific product purchases can be influenced by sales staff once customers are in store. In these markets, the ability to predict short-term brand choice in itself is important to provide indications of likely store traffic, which can be used to plan aspects such as staffing levels and opening hours. Likewise retail purchases are often relatively frequent in nature, and as such predictions in shorter time frames would be of greater interest.

The purpose of this research is therefore to test the accuracy of the VPS in predicting short-term brand choice, without a prior product category reference. To provide context to the accuracy figures two benchmarks are used. The first is the error margins found in other similar Juster/VPS research and the second is a self reported measure of past behaviour. This ensures some objective assessment of the quality of the results.

Background

The Juster scale was developed in the 1960's as a technique for predicting the purchase of consumer durables, a key indicator of future economic activity for the US government. It is a carefully worded, 11-point scale that asks people to assign probabilities to the likelihood of them undertaking a specific behaviour in the future (Juster & Wachtel 1972). The initial development, undertaken in the car industry (referred to as the 'Detroit Experiment' and published by Byrnes (1964)), showed how a substantial number of those classified as non-intenders[1] actually reported a non-zero probability of purchase. This was then further developed by Thomas Juster (Juster 1966) to create the scale and verbal anchors as illustrated in Figure 1. This scale has been reported to be more accurate than measures of purchase intentions or other attitude based measures (eg, Day, Gan et al. 1991).
The Juster scale, in written form, is as follows:

Figure 1. Juster Probability scale

10 Certain, practically certain (99 chances in 100)
9 Almost sure (9 chances in 10)
8 Very probable (8 chances in 10)
7 Probable (7 chances in 10)
6 Good possibility (6 chances in 10)
5 Fairly good possibility (5 chances in 10)
4 Fair possibility (4 chances in 10)
3 Some possibility (3 chances in 10)
2 Slight possibility (2 chances in 10)
1 Very slight possibility (1 chance in 10)
0 No chance, almost no chance (1 chance in 100)

The Juster scale has been subjected to years of use and testing of its psychometric properties in its written form (Juster 1960; Juster 1966; Juster & Wachtel 1972; Seymour, Brennan & Esslemont 1994) and also as modified versions for application over the telephone (Gendall, Esslemont & Day 1991; Brennan, Esslemont & Hini 1994; Brennan 1995). The VPS, developed by researchers at Massey University, provides a version of similar accuracy to the written form that can be read out over the telephone. This version, taken from Brennan, Esslemont & Hini (1994), is as follows:

Now I would like you to think about your chances of buying from <insert reference>. If you are certain or practically certain that <insert time frame> you will purchase <insert reference> you will choose the answer 10. If you think that there is no chance, or almost no chance the best answer would be zero. If you are uncertain about the chances choose another answer as close to zero or 10 as you think it should be.

---

[1] Non-intenders were those who said they had no plans or intentions to buy a car

You can think of the numbers as chances out of 10, for example 3 would mean 3 chances out of 10 that you will purchase <insert reference> during <insert time frame>, while a 7 would mean 7 chances out of 10 and so on.
What are the chances that you personally will purchase from <insert reference> in the <insert time frame>?

The majority of the initial testing of the Juster scale was conducted in 'durable' product categories such as cars or whitegoods and over time periods of 12 months or greater. However, a small stream of researchers has also explored the ability of purchase probabilities to predict purchases within a shorter time period (generally time periods of 3 months) in markets where goods are frequently purchased. Day, Gan et al. (1991) presented a comparison of the accuracy of the Juster scale across studies and markets conducted over a 3 month time period and found that 43% (9 out of 21) of the studies they reported had predictive errors of less than 20%. Hamilton-Gibbs, Esselmont and McGuiness (1992) examined two methods of predicting frequently purchased items. The first was a multiple question method, where households were asked to estimate their probability of buying one item, then two items and so on. The second was a constant sum method, which forced respondents to give a total sum of purchase probabilities of 1. The mean predictive error for each of the techniques was 15% for the multiple question method and 6% for the constant sum method. In a further comparison of these methods, conducted by Seymour, Brennan and Esselmont (1994), the unweighted predictive error was about 26% for all methods. The error margin was reduced down to 12-16% once the results were weighted for probability of purchasing from the category.

Clawson (1971) examined predictive power of purchase probabilities over a 90 day period over 9 product categories, ranging from moving house to buying a TV to riding the local bus. Only two of the product categories are what could be considered as (relatively) frequently purchased products. These were riding in a local bus and attending the movies. These two product categories had a percentage error margin of 4% and 5% respectively.

There have been few explorations into the ability to predict brand purchases. Attempts to predict purchases of 3-4 brands in each of two product categories of yoghurt and soup achieved a comparative unweighted predictive error of 83% and a weighted error of 16% (Brennan & Esselmont, 1994). Using these derived shares to predict the proportion purchasing each brand, the predictive error is reduced to 10%[2]. In the latter case, the ability to predict the brand purchase accurately was dependent on the initial product category purchase probabilities. This initial step was possible in this instance as the brands were very narrowly defined in a single category.

Other research into the accuracy of purchase probabilities at brand level has explored the ability of responses to the Juster/VPS to predict defection from brands in subscription markets such as telecommunications, banking and insurance (Danenberg 1998). Subscription markets are those where customers typically ascribe 100% of their share of category requirements to a single brand. In this context the use of another brand is defection in that the customer ceases (or significantly downgrades) use of the first brand (Sharp, Wright et al. 2002). Danenberg (1998) reported that the mean percentage error was relatively high at 60% but this was noted to be due to the low proportions of people switching. The Mean Absolute Deviation (MAD) of 2.7 was in line with the error margins in other purchase probability studies. In these

[2] These percentage errors were calculated from the percentage point errors contained in the paper. In addition one brand, no frills, which had percentage error of over 290% was removed as inclusion resulted in misrepresentative averages.

studies ambiguity with the category for the behaviour was avoided by using the idea of switching from 'main brand' (with that brand predetermined and then specified), or a sub category was specified (eg, switching from brand X for long distance calls). There was no direct comparison between more and less specific product category delineation.

Across all of the studies reported here (summarised in Table 1), the average proportional error of the most accurate approaches across all studies of short term or brand purchases (with extreme outliers removed) is about 10%. This is one of the benchmarks that will be used to objectively assess the accuracy obtained in this study.

Table 1: Published accuracy levels for short term and/or brand purchases

Study	Market	Percent-age point error	% error
Clawson (Clawson 1971)	Movies	3	5
Clawson (Clawson 1971)	Ride the local bus	1	4
Day (1987)	Meal out	-2.7	-4
Day (1987)	Movie	2	5
Day (1987)	Paperback book	2	5
Gan et al (1985)	LP record to Tape	11	31
Gan et al (1985)	Pair of shoes	-3	-5
Gan et al (1985)	Hard cover book	7	19
Hamilton-Gibbs et al (1992)	Cheese	Na	3
Hamilton-Gibbs et al (1992)	Margarine	Na	4
Hamilton-Gibbs et al (1992)	Butter	Na	1
Hamilton-Gibbs et al (1992)	Eggs	Na	9
Hamilton-Gibbs et al (1992)	Toothpaste	Na	9
Hamilton-Gibbs et al (1992)	Ice cream	Na	9
Hamilton-Gibbs et al (1992)	Spaghetti	Na	5
Brennan and Esslemont (1994)	Watties (brand)	0	2
Brennan and Esslemont (1994)	Campbell (brand)	1	30
Brennan and Esslemont (1994)	Heinz (brand)	1	17
Brennan and Esslemont (1994)	Fresh and fruity (brand)	-3	-12
Brennan and Esslemont (1994)	Yoplait (brand)	-1	-6
Brennan and Esslemont (1994)	Ski (brand)	-3	-23
Seymour, et al (1994)	Margarine	-9	-14
Seymour, et al (1994)	Butter	-1	-1
Seymour, et al (1994)	Eggs	1	2
Seymour, et al (1994)	Toothpaste	3	13
Seymour, et al (1994)	Ice cream	2	17
Seymour, et al (1994)	Spagetti	-19	-22
Danenberg (Danenberg 1998)	MFI (brand)	2	37
Danenberg (Danenberg 1998).	Toll call (brand)	-1	-15
Mean absolute deviation (MAD)		4	11

Product category specification

The aim of this study is to assess the accuracy of brand choice probabilities, as measured by the VPS, without any product category specification. In the few studies that have previously attempted brand predictions, error margins were dramatically improved with the additional of specification from the category prior to getting brand choice probabilities (eg, the probability of buying yogurt, then probability of buying Ski). Therefore it might be expected that not specifying a category will result in less accurate predictions. However in the retail context of this research, the brand is actually the destination of the buyer, as buyers have to enter a specific store (and therefore "choose" the store), prior to making any specific product purchases (eg, walking into McDonald's then looking at the menu board to decide what to eat). Understanding what specific products will be bought is important for inventory levels, however this can be influenced by in store staff (eg, McDonald's' ploy of asking 'would you like fries with that). The ability to predict store traffic is important for service and accessibility retail issues, such as staffing levels. Therefore it is useful to understand if it is possible to use the VPS to accurately predict retail brand choice without having to specify a product category.

Method

The research was conducted in the retail fast food market. This is appropriate as it is one where brands offer a wide range of products, and a considerable amount of on-selling effort takes place once a customer is at the destination (e.g., the bundling of several complementary products to create 'value meals'). Therefore the final products bought may not be related to the products the customers intended to buy upon entrance to the store, if indeed they had any specific intention prior to looking at the in-store menu options. Thus the need to predict brand choice, without the category purchase qualifier, is relevant in this context.

This study is the time frame of four weeks, versus an average of three months for other research. This shortened timeframe should have a positive rather than negative effect on accuracy, as the time for prediction is closer to the time when the prediction is made (Ajzen and Fishbein 1980).

The longitudinal data was collected in two waves. The first was an Australia wide 12.5 minute telephone survey with 623 randomly recruited respondents. As part of the survey, respondents were read out the VPS and asked to assign a number indicating the chance they would buy from each of a series of five brands in the subsequent four-week period. The respondents were questioned about each brand individually and to avoid any order effects, the brands were randomised. As part of the VPS, respondents were asked to think of the full range of products offered by the brand to ensure respondents thought of all of the possible items they could buy when responding. In the same interview respondents were also asked from which brands they had purchased in the 4 weeks prior to the first wave interview as a measure of recent past behaviour.

At the end of the interview, respondents were asked permission to be recontacted. A total of 92% of respondents gave this permission. Four weeks after the first wave interview, re-contact attempts were made over a three-night period. A total of 342 (60%) respondents completed the follow up interview. Tests comparing the characteristics of those who were reinterviewed versus those who could not be recontacted revealed no significant differences, suggesting low risk of non-response bias. In the follow up study, as part of a series of questions on the same topic, respondents were asked if they had purchased from each of the brands in question in the four weeks between interviews.

Results

The results are presented in two stages. The first is the aggregate level prediction for each of the brands. This is calculated by multiplying the mean VPS probability for each brand by a factor of 10 (Day, Gan et al. 1991). In addition to the VPS predictions, Table 2 also shows the predictions based on past behaviour (% who bought the brand in the four weeks prior to the first wave interview), and the claimed buying behaviour captured in Wave 2.

The aggregate level results show that overall there is considerable consistency across all three measures in terms of rank order. However, the predictions based on past behaviour were more accurate at predicting the % who will buy in the next time period, with consistently lower deviations. Both measures over-predicted, but the VPS deviations were about 2.5 times greater. In comparison with other purchase probability studies, the absolute percentage error (42%) is over twice the average percentage error of 10% previously reported. This figure of 42% is substantively influenced by the over-prediction of the smallest brand (error of 104%). Removal of this figure provides an average of percentage error of 26%, which is 2.6 times the benchmark of 10%. In contrast the error for past behaviour (11%) is of the same magnitude of the benchmark. Percentage point error for past behaviour is also half that evident in Table 1. The results therefore suggest that at brand level, past behaviour of the market is a better predictor of future behaviour than self reported purchase probabilities.

Table 2. Aggregate level brand predictions

		VPS			Past behaviour		
	Actual % behaviour	Predicted %	Percentage point error	% error	Predicted %	Percentage point error	% error
Brand 1	57	39	-19	-32%	52	-5	-8%
Brand 2	20	22	2	11%	23	3	16%
Brand 3	19	18	-1	-7%	19	0	0%
Brand 4	11	18	7	58%	11	0	-3%
Brand 5	6	11	6	104%	4	-2	-27%
MAD	23	22	7	27%*	22	2	11%

*Without Brand 5

In a stationary market, past behaviour might provide more accurate predictions at aggregate level, however self-reported purchase probabilities may better detect who will actually purchase from the brand. Logistic regression was used to determine the accuracy in helping classify both buyers and non-buyers. Models were calculated for VPS and Past behaviour individually (see Table 3), and then in combination (seeTable 4). All results were statistically significant at $p<0.001$ level. Overall, past behaviour was substantially better at classifying buyers (51% compared to 35% accuracy) and had a similar accuracy in classifying non-buyers. The exception is for Brand 5 where VPS was better at identifying those who will become buyers. To test for combined effects, and to assess if the VPS added to overall accuracy when combined with past behaviour, both past behaviour and VPS were analysed together. The results (shown in Table 4) show an improvement in the accuracy of classifying buyers for Brands 1 and 5. The other three brands had increases in the correct classification of non-buyers, but substantive decreases in the proportion correctly classified for buyers, thereby not improving predictive accuracy. Therefore, given the added complexity of collecting VPS information, it is difficult see any benefit of including VPS in addition to past behaviour for these brands in this context.

Table 3. Individual model logistic regression results

	Past behaviour			VPS		
	Exp (b)	% correct class as buying	% correct class as not buying	Exp (b)	% correct class as buying	% correct class as not buying
Brand 1	3.6	75	78	1.5	77	80
Brand 2	3.2	62	87	1.6	32	96
Brand 3	4.1	51	94	1.4	23	97
Brand 4	4.8	67	90	1.5	30	97
Brand 5	7.6	0	100	1.5	11	98
Average	**4.7**	**51**	**90**	**1.5**	**35%**	**94**

Note: All results sig at P<0.001 level

Table 4. Combined model logistic regression results

	Past behaviour Exp (b)	VPS Exp(b)	% correct class as buying	% correct class as not buying
Brand 1	2.2	1.3	81	78
Brand 2	2.0	1.4	52	95
Brand 3	2.8	1.3	36	97
Brand 4	3.4	1.3	56	93
Brand 5	3.9	1.3	49	99
Average	**2.9**	**1.3**	**55**	**92**

Note: All results sig at P<0.001 level

Discussion

This research was undertaken in a specific context of brand purchases (with no specific category specifier) in a short time frame of 4 weeks. Using longitudinal data, the results suggest that past behaviour is a better predictor of future brand buying behaviour than self reported probabilities. This was evident both at aggregate (brand) and individual level. Ascertaining past behaviour during the time period is substantively shorter in time and effort. As such it appears to be a more useful measure than self-reported future probabilities when the aim is to predict short-term brand purchase in frequently bought situation and the nature of specific category purchases is difficult to identify.

These results are likely explained in part by the habitual nature of most brand purchase, in that people buy from a steady repertoire of brands over time (Ehrenberg 2000). Therefore the brands chosen in the past are likely to be strong indicator of what people will buy in the future. The exception was for the smallest brand in the market, where the VPS did produce more accurate results than past behaviour. This is most likely because being a small brand, it is not bought by many people, nor very frequently by those people (the Double Jeopardy Law as per McPhee (1963); Ehrenberg, Uncles et al. (2004)). Therefore the chance of the same person purchasing twice within an 8-week period is much lower than for other brands. In this instance the VPS is a better choice of measure than past behaviour.

Most of the past research into the predictive accuracy of the Juster scale/VPS has been very positive in its conclusions. This research is important as it suggests there may be boundary conditions for when these purchase probabilities are less accurate, and alternatives better options. This does not negate the benefits of the VPS in situations involving a change of behaviour (eg, brand defection) were past behaviour is not possible to include. Also it does not contradict those studies that looked at predictions over longer time periods and/or incorporating the product category. However it should be noted that even those studies looking at longer term purchase of the category did not seem to take into account past buying behaviour. Comparing the results achieved with past behaviour benchmarks is an important area for future research. The results from this research suggest that past behaviour should, where possible, be an integral test of any measure purporting to accurately predict future behaviour to provide an objective comparison to assess accuracy.

Limitations

The first key limitations of this study are that it dealt with a stationary market, in that over the two 4 week purchase periods there was little change in the aggregate level purchase of brands. This may be the reason why past behaviour provided the more accurate predictions. Purchase probabilities, where people can express future changes in behaviour may provide more accurate predictions when brands/markets are growing or declining. The second key limitation is that it relied on claimed purchase behaviour for both past and subsequent behaviour measures. This many have introduced some bias in that the respondents who had a higher propensity to have claimed purchase in the past 4 weeks may have had a higher propensity to claim purchase in the subsequent 4 weeks, regardless of actual behaviour. Indeed there may have been some overlap whereby people were ‘telescoping’ when reporting the purchases in the second 4 week period by re-including the purchase (s) mentioned in the first 4 week period. These two factors would increase the apparent accuracy of the past behaviour measure. Ideally the behavioural measure should be collected independently (eg, via a diary or sales data). Further exploring these two limitations should form the basis for future research. Other avenues of future research should compare purchase probabilities against more traditional measures of attitudes and intentions. Despite a number of claims, there still have been few studies that directly compare predictive ability of these two types of measures. More is needed to assess the effectiveness of measures for predicting future behaviour. However such tests should include past behaviour to check that any measures improve on this benchmark.

References

Ajzen I & Fishbein M (1980). *Understanding attitudes and predicting social behaviour.* Englewood Cliffs, NJ, Prentice-Hall.

Brennan M (1995). Constructing demand curves from purchase probability data: An application of the Juster Scale. *Marketing Bulletin (6)*, 51-58.

Brennan M & Esslemont D (1994). The accuracy of the Juster Scale for predicting purchase rates of branded, fast-moving consumer goods. *Marketing Bulletin (5)*, 47-53.

Brennan M, Esslemont D & Hini (1994). *Obtaining purchase predictions via telephone interviews*. Australian Marketing Conference, Adelaide, Marketing Science Centre.

Byrnes JC (1964). Consumer intentions to buy. *Journal of Advertising Research **4***(3), 49-51.

Clawson CJ (1971). How useful are 90-day purchase probabilities? *Journal of Marketing **35***(October), 43-47.

Danenberg NJ (1998). *Predicting Customer Loyalty: A Probabilistic Approach. School of Marketing*. Adelaide, University of South Australia**:** 205.

Day D, Gan B, Gendall P & Esslemont D (1991). Predicting purchase behaviour. *Marketing Bulletin **2***, 18-30.

Ehrenberg ASC (2000). Repeat buying - facts, theory and applications. *Journal of Empirical Generalisations in Marketing Science **5,*** 392-770.

Ehrenberg ASC, Uncles, MD et al (2004). Understanding brand performance measures: using Dirichlet benchmarks. *Journal of Business Research (*forthcoming).

Gendall P, Esslemont D & Day D (1991). A comparison of two versions of the Juster Scale using self-completion questionnaires. *Journal of the Market Research Society **33***(No. 3, July), 257-263.

Juster FT (1960). Prediction and consumer buying intentions. *American Economic Review, 50,*: 604 -622.

Juster FT (1966). *Consumer Buying Intentions and Purchase Probability: An Experiment in Survey Design,* National Bureau of Economic Research, Columbia University Press.

Juster FT & Wachtel P (1972). Anticipatory and objective models of durable goods demand. *American Economic Review **72***: 564-579.

McPhee WN (1963). *Formal Theories of Mass Behaviour*. New York, The Free Press of Glencoe.

Seymour P, Brennan M & Esslemont D (1994). Predicting purchase quantities: Further investigation of the Juster Scale. *Marketing Bulletin(5):* 21-36.

Sharp B, Wright M et al (2002). Purchase loyalty is polarised into either repertoire or subscription patterns. *Australasian Marketing Journal 10*(3): 7.

Jenni Romaniuk is a Senior Research Fellow, Marketing Science Centre, University of South Australia.