

Improving the Accuracy of Recall Data: A Test of Two Procedures

Mike Brennan, Joanna Chan, Dean Hini and Don Esslemont

Questions have been raised about the validity of data obtained from retrospective survey designs, since these are heavily dependent on the accuracy of respondent recall. One of the most serious problems is over-reporting, typically attributed to "forward telescoping", where respondents report events that occurred outside of the time period under consideration, thereby inflating the results. This study tested two procedures for reducing forward telescoping: the provision of a landmark event to clearly mark the beginning of the recall period; and a double question bounded recall procedure whereby respondents were first asked to recall events for a longer time period (previous eight weeks), then for the time period of interest (previous four weeks). This paper presents the results of this research and discusses the implications for studies dependent upon recall data.

Keywords: Juster Scale, purchase probability, telescoping, bounded recall

Introduction

Market researchers sometimes require information about respondents' past behaviour, and frequently rely on respondents' retrospective reports for this information. It is, however, well known that respondent recall is often inaccurate, consequently resulting in the over- or under-reporting of past events (Sudman & Bradburn 1973, 1974). Unfortunately, there are generally no other sources of the required information, and alternative procedures for obtaining the necessary information, such as diary panels, are often not feasible (Sudman & Bradburn 1973). Thus a researcher is faced with the dilemma of how to obtain recall data in a way that minimises potential bias.

While this issue is of importance to all researchers using retrospective reports, it is particularly important with regard to recent work on the development of the Juster Scale. The Juster Scale is an eleven point purchase probability scale, developed in the 1960's by Thomas Juster (Juster 1966). The scale has been developed in recent years for use in both face-to-face, self-completion and telephone interviews, to estimate demand for both branded and unbranded goods, and services (Day, Gan, Gendall & Esslemont 1991; Gendall, Esslemont & Day 1991; Hamilton-Gibbs, Esslemont & McGuinness 1992; Brennan & Esslemont 1994; Seymour, Brennan & Gendall 1994; Brennan, Esslemont & Hini 1995a; Brennan, Esslemont & Hini 1995b; Brennan, Esslemont & U 1995). In all of these studies, the accuracy of the purchase predictions has been determined by contacting respondents at the end of the time-period for which the purchase predictions were made, and asking them what they actually bought. In other words, the evaluation of the Juster Scale has depended on recall data. Thus, if the recall data is inaccurate, the conclusions drawn from these studies may be erroneous.

There are three major sources of error in recall data: omission, rounding, and "telescoping". Omission can be simply the result of forgetting, which is most likely to occur with low-involvement products or services, and infrequent events or purchases. It can also be due to the intentional withholding of information, which more usually occurs when that information relates to socially undesirable or unacceptable behaviour. Omission also occurs when a

respondent is not aware of an event, such as a purchase by another family member (Sudman & Bradburn 1973; Cook 1987).

The second source of recall error, rounding, reflects a tendency for people to recall events or behaviour in terms of commonly used increments or multiples of time or quantity. For example, time may be reported in terms of hours or half-hours rather than minutes; a month may be reported as 30 days, or vice versa; days may be reported in terms of weeks, or months; quantities may be reported as so many dozen, or in terms on multiples of 5 or 10 or 100, and so on. The use of these incremental values as approximations for the true values can lead to significant under- or over-reporting.

A related phenomenon is averaging, which can arise when people respond to questions about irregular events. Averaging occurs when, for example, the retrospective report is based on a figure calculated by multiplying the estimated "average" number of purchases in a time period by the number of time periods under consideration. In effect, the irregular event has been "regularised", and the reported figure is likely to grossly overstate the true situation (Cook 1987; Huttenlocher, Hedges & Bradburn 1990).

Finally, telescoping, or more correctly, "forward" telescoping, is the displacing of events in time towards the present. In other words, events occurring outside of the specified time-frame are included in the time-frame, leading to over-reporting of events. This effect, also known as "border bias" (Mahalanobis 1946), has been widely reported, and appears to be a much more serious problem than "backward" telescoping, where events are seen as being further in the past than they really were (Sudman & Bradburn 1973).

Of the three main sources of recall error, forward telescoping is considered to be the major cause of problems when respondents are asked to recall behaviour during a short recent time period (Sudman & Bradburn 1974; Sudman, Finn & Lannom 1984). However, several procedures have been reported that appear to effectively reduce forward telescoping.

One of these procedures involves the use of bounded recall. Developed by Neter & Waksberg (1964), the procedure involves two separate interviews with the first interview serving as a bound for the second interview. Unfortunately, the procedure used by Neter and Waksberg requires a panel and two interviews. Thus it is expensive. It is also wasteful, as the data from the first interview is not used in the analysis, and the use of two interviews almost inevitably results in a reduced sample size due to attrition.

The bounded recall procedure of Neter & Waksberg (1964) was adapted by Sudman et al. (1984) for use in a single interview. Sudman et al. employed a simple two-step procedure. Respondents were first asked about their behaviour in a previous period (e.g. last month), then about their behaviour in a more recent period (e.g., the current month). From the results of two studies, one involving the recall of personal medical events and the other the recall of snack purchases, the modified procedure was reportedly as effective in reducing telescoping as the panel procedure of Neter & Waksberg (1964). However, while the level of recall was certainly reduced by the procedure, no validation was undertaken to determine whether these lower figures were in fact more accurate.

A slightly different but essentially similar procedure was used by Loftus, Klinger, Smith & Fiedler (1990) to examine effects on recall of recent personal medical events. They first asked respondents to recall their behaviour during a longer reference period (last six months),

then in the reference period of interest (last two months). That is, the second period was a subset of the first. They also tested the effect of reversing the time periods, that is, the more recent time period was used before the time period of interest (e.g., one month, then two months). They concluded, using validated data, that both two-time period procedures produced more accurate results and less forward telescoping than when a single time period was used.

A second procedure for reducing forward telescoping is to use a landmark event to bound the recall period. This landmark event may be a unique event, such as the eruption of Mt. St. Helens; a personal landmark, such as a birthday or personal event; or a public landmark event, such as a public holiday like Christmas or Easter. Loftus & Marburger (1983) conducted five studies to investigate the effect on forward telescoping of all three types of landmark event, and the specification of a calendar date. The behaviours or events examined included: being attacked, calling the police, not reporting a crime to the police, having a birthday in the past six months, eating lobster in the past six months, recalling a failed attempt to release hostages. Loftus and Marburger concluded that all three types of landmark event did reduce telescoping, that is, substantially reduced reports of the activities. It should be noted, however, that the accuracy of the recall was not verified, although post-interview probing was used to try to identify telescoping. Provision of a specific calendar date also reduced telescoping, but was not as effective as the landmark events.

The studies reviewed above suggest that two simple procedures can be used to substantially reduce forward telescoping in retrospective reports. This raises the question of which of the two is the most effective, and whether the two procedures could be used in combination to increase their effectiveness. A counter to criticism of recall data in previous Juster studies has been that the first interview, when respondents were asked about their prospects of purchasing, was in fact itself a landmark event (Brennan, Esslemont & U 1995). Thus the purpose of the study reported here was twofold: First, to determine whether an initial interview involving a purchase probability task does in fact act as a landmark event, thereby reducing the effects of over-reporting in the subsequent recall interview. Secondly, to compare the effectiveness of three techniques for reducing over-reporting: two-step bounding, a landmark event, and the two procedures in tandem¹.

The study was run in conjunction with a short survey for a local video shop, which was conducted to estimate demand for a new service. The respondent behaviour examined in the experimental study was respondents' video hire during the previous four and eight week period. The recall study was run in conjunction with a purchase probability study, but was unaffected by it, so only the procedure and results pertaining to the recall study are reported here.

Method

Sample

A sample of approximately 900 customers of a local video store was systematically assigned to one of four treatment groups (see Table 1). These groups were balanced, as far as possible, on the basis of number of video tapes respondents had hired in the past two months. This procedure ensured that all respondents were current customers of the video shop. The sample excluded respondents with a birthday falling within the study time-frame, as this personal landmark event may have affected both video hiring behaviour and recall. The groups were

also balanced with respect to respondents' month of birth, to balance any effects associated with the recency of a personal landmark event (birthday). The treatment groups were also balanced across interviewers. For each interviewer the respondents in each group were sorted alphabetically to produce a random order for the interviews.

Of the 908 respondents in the initial sample, 296 could not be contacted, thus the usable sample had 612 respondents; 499 interviews were completed, representing a response rate of 82% (see Table 2).

Table 1. Research Design

Group 1	CONTROL:	Recall only; no initial interview
Group 2	LANDMARK EVENT:	Verbal Purchase Probability Scale used in initial interview; followup recall (standard procedure)
Group 3	TWO-STEP BOUNDING	Two-step recall; no initial interview
Group 4	COMBINATION	Verbal Probability Scale used in initial Interview: Two-step follow-up recall

Table 2. Participation Rates

	Group 1	Group 2	Group 3	Group 4	Total
Initial sample	227	227	227	227	908
Non-contact	71	68	70	87	296
%	(31.3)	(30.0)	(30.8)	(38.3)	(32.6)
Usable sample	156	159	157	140	612
Refusals	25	35	22	31	113
%	(16.0)	(22.0)	(14.2)	(22.1)	(18.5)
Final sample	131	124	135	109	499
Participation Rate	84.0	78.0	86.0	78.0	81.5

Note: Participation rate = (final sample/usable sample)*100

Procedure

Interviews were conducted, by telephone, by five trained interviewers. Each interviewer interviewed respondents from all four treatment groups. Interviews were rotated across treatments, to balance any effects due to day of week and time of day. Up to six calls were made to each respondent in an attempt to complete an interview.

Phase 1, which involved only treatment Groups 2 and 4, began in week 5. Each interview began in the same way. That is, respondents were asked if they had a videoplayer in their household, if they personally had ever rented a video tape, and which local video stores they belonged to. They were then informed that the remaining questions related to a particular

video store, and asked about the probability of using the new service if this was available at the store. This information was gathered using the Verbal Purchase Probability Scale, based on the Juster Scale, developed for use in telephone interviews by Brennan, Hini & Esslemont (1995a). Still using the Verbal Purchase Probability Scale, information was then gathered on the respondent's prospects of hiring videos during the next four weeks.

Respondents were also asked if they shared their video store card with anyone, and if they did, were asked to answer the same questions on behalf of these people. At the end of the interview, the interviewer requested permission to re-interview the respondent at a later date.

In Phase 2, recall data were obtained from respondents in all four groups. Those who had not been previously contacted in Phase 1 (Groups 1 and 3) received the introduction used in Phase 1. A question about a new store service was followed by the recall questions, then by purchase probability questions for another study.

The groups who had been interviewed previously (Groups 2 and 4) were then asked the recall questions, to determine what they personally had hired, and what they thought anyone else who had used their card may have hired. For Groups 1 and 2, the recall period was the previous weeks, while for Groups 3 and 4, respondents were asked about video hire during the previous weeks, then the previous weeks. Respondents were then asked these questions with regard to other people who had used their card.

For the duration of the study, beginning four weeks before the first interviews, records of actual video hire were extracted, for each participant, from the store computer records.

Results and Discussion

The main results of the study are shown in Table 3, which reports the mean number of tapes respondents recalled hiring and the mean number of tapes actually hired, as well as the recall error, for each of the treatment groups.

Table 3. Effects of procedures on accuracy of recall: Total sample

Group	N	Recall ¹	Actual ²	Difference	Error ³
1	131	2.36	1.48	.88	59.5
2	124	1.78	1.26	.52	41.3
3	135	1.15	1.14	.01	.9
4	109	1.80	1.47	.33	22.5

Note:

1. Mean number of tapes recalled.

2. Mean number of tapes hired.

3. Error = (recall-actual/actual)*100

Three main points emerge from the results shown in Table 3. First, it is clear that over-reporting has occurred, as the number of tapes recalled was greater than the number of tapes actually hired, for all four groups. While this may be due to forward telescoping, it is not possible to confirm this, and there may be other possible explanations for this result.

However, knowing the cause of the over-reporting is not crucial to this study. The key question is whether or not the techniques originally designed to reduce forward telescoping do effectively reduce over-reporting.

Second, the procedures did reduce over-reporting. The experimental procedures (Groups 2, 3 and 4) all lead to lower reported tape hire than the control (Group 1). This finding is consistent with the findings of Sudman et al. (1984) and Loftus et al. (1990).

Thirdly, the two experimental procedures, alone and in combination, lead to more accurate recall than the control procedure. The most accurate results were obtained from the two-step bounding procedure (Group 3), followed by the combined procedures (Group 4). Both of these procedures produced significantly more accurate recall than the simple recall procedure used in the control.

The procedure involving the landmark event on its own (i.e., the purchase probability interview: Group 2) also produced more accurate recall than the control, but the effect was not great and not statistically significant. This is contrary to expectations, as it has been assumed in the past that the initial interview in Juster studies would serve as a landmark event and significantly reduce over-reporting in the follow-up recall interviews. Of course, one explanation might be that the over-reporting is due to forward telescoping, and co-incidentally is not reduced by a landmark event procedure known to reduce forward telescoping. Alternatively, the initial interview may simply not be an effective landmark event.

These reservations about the effectiveness of the first interview as a landmark event, or more precisely, about the effectiveness of the first interview for reducing over-reporting, appears to be justified when a more detailed analysis is undertaken. A certain proportion of the video card users share their cards, and there was no way of identifying and eliminating these people from the sample before the study began. However, respondents were asked if they shared their card, and if so, were asked about its use by these other people. Clearly, the answers from the card sharers is likely to be less accurate than those from sole users of the cards. An analysis of the accuracy of recall for sole users and shared users is provided in Table 4.

As one would expect, the mean rate of actual video hire is greater for the shared card users than for the sole card users, with the exception of Group 4, whose results appear to be an anomaly. However, the rate of recalled card use is much higher for the shared card users. This indicates that over-reporting rather than under-reporting of the card usage by others has occurred. This has resulted in a much higher recall error for the shared card users group, for all four treatments.

The most important findings in Table 4, however, are the recall error rates for the four treatments among the sole card users. Group 3, which experienced the two-step bounding procedure, again produced the most accurate recall, closely followed by Group 4, which experienced the combined procedures. However, the accuracy of recall for Group 2, which used the first interview as a landmark event, was less accurate than the control.

Conclusions

The results of this study confirm previous findings (Sudman et al. 1984; Loftus et al. 1990) that a two-step bounding procedure can effectively reduce over-reporting in recall tasks. In

Table 4. Effects of procedures on accuracy of recall

Group	N	Recall ¹	Actual ²	Difference	Error ³
Sole Card User					
1	101	1.69	1.25	.45	36.0
2	84	1.61	1.06	.55	51.9
3	110	.94	.99	-.05	-5.0
4	82	1.48	1.65	.17	10.3
Shared Card User					
1	30	4.60	2.27	2.33	102.6
2	40	2.15	1.67	.48	28.7
3	25	2.08	1.80	.28	15.6
4	27	2.78	.93	1.85	198.9

Note: 1. Mean number of tapes recalled.
 2. Mean number of tapes hired.
 3. Error = (recall-actual/actual)*100

contrast, a prior interview, even when it involves a purchase probability task, does not effectively reduce over-reporting. Fortunately, the two procedures can be used in tandem, although the effects are not as great as when the two-step procedure is used on its own. For studies testing the Juster Scale or other purchase probability scales, a prior interview is unavoidable, so there is no option. In this case, the two-step bounded recall procedure should be used in the follow-up recall interviews.

Of course, these conclusions are based on the results of only a single activity, video tape hire. Since previous studies of the Juster Scale have found wide variation in the error of predictions, based on recall, for different types of products and services, the conclusions drawn here require validation. Even so, the two-step bounding procedure produced results consistent with those reported elsewhere, so there is little reason to suspect that the procedure would not also be effective with other products and services. And, because it is simple to use, it could easily be incorporated into future surveys gathering retrospective data.

References

- Brennan M & Esslemont D (1994). The accuracy of the Juster Scale for predicting purchase rates of branded, fast-moving consumer goods. *Marketing Bulletin*, 5, 47-52.
- Brennan M; Esslemont D & Hini D (1995a). Obtaining purchase predictions via telephone interviews. *Journal of the Market Research Society*, 37 (3), 241-250.
- Brennan M; Esslemont D & Hini D (1995b). *A test of three methods for estimating levels of purchase*. In: Proceedings of the Seventh Bi-annual World Marketing Congress, Melbourne, July 6-10.
- Brennan M; Esslemont D & U C (1995). Using the Juster Scale to estimate the price- demand relationship. *Asia-Australia Marketing Journal*, 3, in press.
- Cook WA (1987). Telescoping and memory's other tricks. *Journal of Advertising Research*, 27, RC5-RC8.

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

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 (3), 257-263.

Hamilton-Gibbs D; Esslemont D & McGuinness D (1992). Predicting demand for frequently purchased items. *Marketing Bulletin*, 3, 18-23.

Huttenlocher J; Hedges LV & Bradburn NM (1990). Reports of elapsed time: Bounding and rounding processes in estimation. *Journal of Experimental Psychology*, 16 (2), 196-213.

Juster FT (1966). Consumer buying intentions and purchase probability. Occasional Paper 99, *National bureau of Economic Research*, Columbia University press.

Loftus EF; Klinger MR; Smith KD & Fiedler J (1990). A tale of two questions: benefits of asking more than one question. *Public Opinion Quarterly*, 54, 330-345.

Loftus EF & Marburger W (1983). Since the eruptions of Mt. St. Helens, has anyone beaten you up? Improving the accuracy of retrospective reports with landmark events. *Memory and Cognition*, 11 (2), 114-120.

Mahalanobis PC (1946). Recent experiments in statistical sampling in the Indian Statistical Institute. *Journal of the Royal Statistical Society*, 109, 326-370.

Neter J & Waksberg J (1964). A study of response errors in expenditures data from household interview. *Journal of the American Statistical Association*, 59, 18-55.

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

Sudman S & Bradburn NM (1973). Effects of time and memory factors on response in surveys. *Journal of the American Statistical Society*, 68 (344), 805 - 815.

Sudman S & Bradburn NM (1974). *Response Effects in Surveys*. Chicago: Aldine publishing Company.

Sudman S; Finn A & Lannom L (1984). The use of bounded recall procedures in single interviews. *Public Opinion Quarterly*, 48, 520-524.

Mike Brennan and Don Esslemont are Senior Lecturers, and Joanna Chan and Dean Hini were graduate students, in the Department of Marketing, at Massey University.

Footnote.

1. In the literature, over-reporting is often referred to as forward telescoping, even though forward telescoping is actually the assumed cause of the over-reporting. When the cause cannot be confirmed, it is probably more prudent to refer to the effect (over-reporting) than the assumed cause (forward telescoping).