SEAM EFFECTS IN QUANTITATIVE RESPONSES

Frederick G. Conrad

University of Michigan

Lance J. Rips

Northwestern University

Scott S. Fricker

University of Maryland

Authors' addresses:

Frederick G. Conrad
Institute for Social Research
University of Michigan
426 Thompson Street
Ann Arbor, MI 48106-1248

Lance J. Rips
Psychology Department
Northwestern University
2029 Sheridan Road
Evanston, IL 60208

Scott S. Fricker
Joint Program in Survey Methodology
1218 Lefrak Hall
University of Maryland
College Park, MD 20742

Abstract: The *seam effect* is an artifact of panel surveys in which respondents are interviewed every few months (for example) and asked for data about each of the intervening months. The effect is defined by small changes in responses between adjacent months *within* a reference period (i.e., within an interview) relative to large changes between adjacent months *across* reference periods (i.e., across interviews). In the present studies, we use an experimental method to examine the seam effect for quantitative questions about dollar amounts. The studies produced robust seam effects, accompanied by forgetting of correct amounts and constant responding (no change between months) within the reference periods. The results also showed decreases in the seam effect when questions about the same topic appeared in different parts of the interview. In some conditions, dependent interviewing (providing respondents with their earlier answers) also reduced the size of the seam effect. Neither manipulation, however, improved the overall accuracy of responses. Dependent feedback may encourage respondents to make less drastic changes at the seam. But since the earlier answers may themselves contain error, dependent interviewing can also perpetuate the same incorrect response from one reference period to the next.

Keywords:

Dependent interviewing

Panel surveys

Longitudinal surveys

Response error

Seam effect

Acknowledgements: NSF Grant SES-9907414 supported the research reported here. We are grateful to Jami Barnett and Jennifer Behr for their help with these studies. We also thank Annette Jäckle for comments on an earlier version of this article, as well as audiences at the American Statistical Association, the Washington Statistical Society, an Academica Sinica conference on survey research methods, and an NSF conference on survey methods. This paper includes studies from the proceedings of the latter two conferences.

SEAM EFFECTS IN QUANTITATIVE RESPONSES

1. Introduction

By tracking change over time, social scientists have far more power to characterize behavior, attitudes, and demographics than if they take measurements at just one time point. Cross-sectional studies allow comparisons among individuals (or households or businesses) with different attributes on one occasion; panel surveys measure the same individuals (households, businesses) on multiple occasions and so make it possible to see how individuals change as their lives unfold.[1] Despite the analytic benefits that panel surveys bring to the research process, they do introduce certain costs. One such cost is response error arising from (or made visible by) repeated measurement of the same individuals both within and between interviews. One error of this sort is known as the seam effect, and survey researchers have observed it in panel surveys such as the Survey of Income and Program Participation (e.g., Jabine et al. 1990; Kalton and Miller 1991; Marquis and Moore 1989; Moore and Marquis 1989; Young 1989) and the Panel Study of Income Dynamics (Hill 1987). These surveys collect both qualitative (“yes” - “no”) and quantitative (e.g., salary) data. The current article concerns seam effects in quantitative data and whether these arise from the same error sources as seam effects in qualitative data.

The seam effect is a pattern of response change in panel surveys in which interviews are usually separated by several months and, in any one interview, respondents are asked about events in each of the months between the previous and current contact. An interview, therefore, produces reports about several months, and so, even though there is not an interview for every month, the survey yields data for every month. For example, the survey might interview respondents in April, during which they must report information about each of the months January, February, and March. The same respondents would be interviewed again in July for the data from April, May, and June. This sequence would continue for the rest of the respondents’ participation in the survey. The seam effect takes the form of more month-to-month change when reports for the comparison months come from two different interviews than from the same interview. It is hard to imagine that this pattern of change is veridical. Instead it almost certainly reflects error arising from how respondents answer the questions.

To make the seam effect concept more concrete, consider the pattern of response change in Figure 1, taken from SIPP data on reports of receiving food stamps and social security benefits (Jabine et al. 1990). Two characteristics of the seam effect can be seen in the figure. First, the amount of change is larger across the interviewing seam (where change is computed from reports provided in two different interviews) than within an interview. Second, the amount of change within an interview is not only small but quite uniform from month to month. This pattern could reflect inflated change across interviewing seams, depressed change within interviews, or both. Either small within-interview change or large cross-interview change can by itself produce seam effects; when both occur in the same study, seam effects can be particularly large.

In an earlier study, we (Rips et al. 2003) demonstrated that the seam effect for qualitative ("yes" and "no") responses was influenced by a combination of recall error and a response strategy known as "constant wave responding," in which respondents provide the same answer for all time periods queried in an interview (Martini 1989; Young 1989). We simulated the usual field data conditions with an experimental approach that compressed the time frame to weeks rather than months and replaced interviews with visits to the laboratory. This allowed us to determine the accuracy of responses.

Responses were in fact more accurate for the most recent than for the most distant weeks in the reference period, suggesting that respondents' recall deteriorated with the passage of time. Because change across the interviewing seam is based on the difference between a recent report in one interview and a distant report in the next, the relatively high levels of change at the seam is likely to reflect more accurate recall in the former case and more forgetting (misreporting) in the latter. In addition, respondents seemed to rely on the constant wave strategy within the interview. Even when the experiments were designed so that there was always some week-to-week change during the period covered by an interview, respondents answered identically for all weeks between 12 and 35 percent of the time (across studies).

The logic of constant wave responding is that it is easier than recalling actual events. If this is the case, then the strategy should be less attractive to respondents when it involves more effort. This was exactly what we observed. When the questions about a given topic in each time period were asked in succession (e.g., the first block of questions concerned purchases in the hardware store in weeks 1, 2, 3 and 4, the next block concerned transactions at the post office in weeks 1, 2, 3 and 4, the next set concerned doctor visits in weeks 1, 2, 3 and 4, and so on), it should be easy to repeat the answer to the first question for each subsequent question in a block. However when respondents were asked about all possible topics for one time period before being queried about the same topics in other time periods (e.g., a block of questions concerned purchases at the hardware store, transactions at the post office, visits to the doctor, and so on, in week 1, the next block concerned the same topics in week 2, the next block concerned the same topics in week 3, and so on), then more effort should be required to respond identically to all questions about a particular topic. In fact, to do so a respondent would need to recall how he or she answered previously – many items earlier – which may be more difficult than recalling the actual events. Moreover, thinking about all items from a given week might provide respondents with retrieval cues that facilitate recall, further reducing the attraction of constant wave responding, Just as one would expect, the proportion of constant wave responding was substantially reduced when the questions were blocked by time period (week) than by topic.

Although the seam effects in our earlier studies appear to be due to a mixture of recall error and constant wave responding, these findings could be quite specific to the qualitative data collected in those studies. It could be that questions requiring quantitative answers do not produce seam effects of the same magnitude, and to the extent that they do produce seam effects, they may do so for different reasons. It is possible, for example, that constant wave responding is less common and forgetting occurs more rapidly for quantitative than qualitative responses. Respondents' use of the constant wave strategy is particularly transparent when the same number is repeated for all time periods—four responses in a row of $29.17 may seem less plausible than four "yes" responses in a row. Because respondents may not want to be perceived as taking shortcuts, they may rely on the strategy less as their use of it is more likely to be exposed. Forgetting could occur more quickly for quantitative than qualitative answers because each instance of the event in question must be recalled to provide a numerical answer. For example, in order to report how much they spent on dairy products in January, respondents must recall each such purchase. But to report correctly whether they had purchased any dairy products in January, they need recall only a single instance. Even if respondents have forgotten the actual events in question, they can correctly guess a "yes" – "no" answer 50% of the time, but guessing will be accurate far less often for numerical figures.

We present two studies below that explore the factors contributing to the seam effect for quantitative answers. The studies document the proportion of constant wave responding and forgetting rates under these conditions. They also look at two factors that may mitigate the seam effect. Study 1 examines whether question grouping can alter the size of the seam effect for quantitative data in the same way as it does for qualitative data. Study 2 looks at dependent interviewing—providing respondents with information about what their answers had been in the previous interview—since this method is thought to reduce the seam effect and has been incorporated in SIPP (Moore et al. 2006), the U.S. Current Population Survey (Polivka and Rothgeb 1993), the Canadian Survey of Labour and Income Dynamics (Hale and Michaud 1995), and other panel surveys (see Mathiowetz and McGonagle 2000 for a review). We turn first to the general method which is very similar for the two studies.

2. Two Studies of the Seam Effect

To analyze the seam effect, it is useful to know the facts that respondents are trying to report. Unless we have access to the correct answers, it is difficult to determine whether the seam effect is due to exaggerated changes at the seam months, to suppressed changes at the nonseam months, or both. For this reason, we have designed a procedure that allows us to control the information that we later ask respondents to recall.

Figure 2 illustrates the basic schedule for these studies. We mailed a questionnaire to respondents each week for six consecutive weeks, as indicated by the vertical lines at the top of the figure. The questionnaires asked a series of 32 questions about the dollar amounts that the respondents had spent for goods and services during the preceding week. We asked the respondents to fill out each questionnaire and mail it back to us within 24 hours. The same respondents also came into the laboratory at the end of the third week and again at the end of the sixth week. During these test sessions, we asked them to report on the information they had seen *in the questionnaires* during each of the preceding three weeks. These two test sessions are our analogs to the survey interviews in SIPP, PSID, and other panel surveys, dividing the interval into two reference periods, as shown in the bottom part of the timeline in Figure 2. The questions that we asked during these sessions provide the week-to-week data that we need in order to study the seam differences. Changes in respondents' answers between weeks 3 and 4 are the seam changes, coming from two different test sessions. Changes in answers between the other pairs of neighboring weeks (1-2, 2-3, 4-5, and 5-6) are nonseam transitions, coming from the same test session. The time scale of the design is in weeks rather than months to allow us to study seam effects more efficiently.

The questionnaire items all had the form:

> *During the last week (A/B through C/D), did you (or someone in your household) spend more or less than $X on Y?*
>
> *Please circle either "More" or "Less" or "Did not purchase."*

For example, one item was: *During the last week (8/9 through 8/15), did you (or someone in your household) spend more or less than $2 on milk and cream from the grocery or convenience store?* Another item was: *During the last week (8/9 through 8/15), did you (or someone in your household) spend more or less than $17 on electricity for your home?* We based these questionnaire items on ones that appear in the Consumer Expenditure Survey. The range of amounts was $1 to $42 across items. The figures were typically whole-dollar amounts, as in these examples, though in a few cases the figure included 50 cents (e.g., $10.50). We asked an individual respondent about the same items (e.g., milk and cream, electricity, etc.) on each questionnaire. The specific amounts, however, varied for some items. For half the questionnaire items, we asked about the same dollar amount each week. For example, respondents might be asked on each questionnaire whether they spent more or less than $2 for milk and cream that week. For the remaining questionnaire items, the amount changed from week to week.[2] Items appeared in random order in each questionnaire.

During the test sessions, we asked respondents to recall the dollar amounts they had seen on the questionnaires. For example, one item in the first test session was: *On the third week's questionnaire, which you filled out on ..., when you were asked about milk and cream, what was the dollar amount you were asked about?* Respondents' answers to these questions provided the data that we analyze later in this article.

2.1 Study 1: Effects of Question Grouping

Our first study varied the way in which we grouped the questions during the test sessions. For qualitative items, we had found larger seam effects when respondents had to answer all questions about a given topic one after another (Rips et al. 2003), and we were interested here in the impact of this variable on quantitative items. Grouping by topic constitutes standard survey practice. For example, SIPP asks about the amount of food stamps received in the preceding month—say, March—then about the amount of food stamps received in February, and so on, before moving to questions about other sources of income. As we noted earlier, however, this practice can encourage constant wave responding because of the ease with which respondents can perseverate with the same response to each item in the series. In this study we compared this grouping by topic to a procedure in which questions are grouped by time period.

To examine the effect of grouping, we gave half the respondents the test questions in an order blocked by topic: In the first test session, for example, these respondents answered the question about milk-and-cream for week 3, week 2, and week 1; then they answered the question about electricity for weeks 3, 2, and 1; and so on. In the second test session, they answered the question about milk-and-cream for weeks 6, 5, and 4; then the question about electricity for weeks 6, 5, and 4; and so on through the full set of items. The remaining respondents answered the test questions in an order blocked by week: During the first test session, these respondents answered all the questions about week 3, then all the questions about week 2, then all the questions about week 1; in the second test session, they answered all the questions about week 6, then week 5, then week 4. Each respondent received a new random order of questions within a block.

We tested 55 adults, whom we had recruited through advertisements in local newspapers in the Washington, D. C. area. The respondents' mean age was 47 years (median 46 years). These respondents had a mean of 16 years of education, and there were 33 women and 22 men. Twenty-one of the respondents were black, two Asian, and thirty-two white.

2.2 Study 2: Effects of Dependent Interviewing

One possible way to reduce seam effects is to remind respondents of their answers from the previous interview before asking them similar questions about the present reference period. The earlier answers may provide respondents with a bridge from one reference period to the next, smoothing the abrupt seam transition. A respondent may be told at the beginning of her second interview, for example, that she had reported a salary of $5000 in the most recent month of the preceding reference period. She would then be asked about her salary during each month of the current reference period. This method of automatically providing respondents with their earlier answers is one ("proactive") form of dependent interviewing. Other forms of dependent interviewing are "reactive" in the sense that they trigger additional questions in response to discrepancies between the present interview and earlier ones (see Jäckle 2006 for a taxonomy of types of dependent interviewing). We will be concerned here only with the proactive method, as in the salary example.[3]

Survey methodologists who advocate dependent interviewing presumably believe that the dependent information jogs respondents' memory. The earlier answers may remind the respondents of facts about the prior reference period, and this may, in turn, help them recall information relevant to the current set of questions. The respondent who learns that she had earlier reported $5000 per month may now be able to remember her current salary correctly. But although dependent interviewing may boost memory, it may also have less desirable effects. Earlier answers may tempt respondents to repeat them, whether or not they are true for the current questions. The respondent may say she continued to receive

$5000, not because she now recalls this as the correct value, but because this figure allows her to avoid an effortful memory search. It is also possible that dependent interviewing provides an anchor for an incorrect estimate (Tversky and Kahneman 1974; Wilson et al. 1996). If the respondent is unsure of her present salary level, she may use $5000 as a starting point for an estimate and insufficiently adjust for subsequent changes. If so, the dependent information could lead to inaccurate answers if the salary level actually changed between periods (as Mathiowetz and McGonagle 2000 point out). Thus, dependent interviewing may exacerbate constant wave responding by extending it across interviews. The end result may be to reduce the size of the seam effect if respondents carry their answers over the seam. But there is no guarantee that reducing the seam effect will be accompanied by an increase in response accuracy.

Our experimental procedure is useful for assessing the effect of dependent interviewing, since we know the true values respondents should report. We can therefore test whether dependent interviewing reduces the seam effect and, independently, whether it achieves this benefit at the expense of correct responding. For this reason, Study 2 repeated the procedure of Study 1, but this time gave to one group of respondents feedback about their earlier answer and gave a second group no feedback.

Respondents again received one questionnaire per week for six consecutive weeks, with each questionnaire asking a series of questions about whether they had spent more or less than a specified amount for goods and services (see Figure 2). The procedure for the test sessions was also similar to Study 1 but with two exceptions. First, we grouped all test questions by topic. For example, respondents answered all milk-and-cream items before moving on to other topics, as in typical surveys. Second, during the second test session, we told respondents in the feedback group what their answers had been for the last week in the earlier test. Suppose, for example, that a respondent had said during the first test session that she had been asked whether she had spent more or less than $2 for milk and cream on week 3's questionnaire. In the second test session, we told the respondent, *during the last session, you said that when you were asked about milk and cream during week 3, the amount you were asked about was $2*. We then asked the respondent about the corresponding dollar amounts for weeks 6, 5, and 4 (e.g., *When you were asked about milk and cream, during week 6, what was the dollar amount you were* *asked about?, When you were asked about milk and cream, during week 5, what was the dollar amount you were asked about?,* etc.). We decided to feed back to respondents the amount they had reported (rather than the correct answer) and to use only the amount from week 3, since this procedure is closest to survey practice. Respondents in the no-feedback condition answered exactly the same questions, but of course they received no information about their previous answers.

We tested 35 adults in this study, 20 in the feedback and 15 in the no-feedback condition. The mean age of respondents was 47 years (median 42 years), and they had a mean of 15 years of education. There were 22 women and 13 men in this sample; 24 were black and 11 white. We recruited the respondents in the same way as in the first study, but no respondent took part in both.

3. Seam Effects as a Function of Question Grouping and Dependent Interviewing

The questions about quantitative information produced clear seam effects in both studies. Figure 3 displays the effect in terms of the absolute change in respondents' answers from one week to the next. For example, if a respondent said that the questionnaire for week 1 had asked whether s/he had spent $1 for milk and cream and that the questionnaire for week 2 had asked whether s/he had spent $5 for milk and cream, then the change for this item would be $| 1 - 5 | = 4$. Observed values from the two studies appear as solid lines, and the true absolute change as the dashed line. In these terms, the results from both studies show that the observed change at the seam weeks (from week 3 to 4) is more than double that of the change at the nonseam weeks. The patterns are, in fact, quite similar for the two studies, as we would expect from the fact that the same items appeared in each.

A repeated-measures analysis of variance for these data confirms significant week-to-week differences ($F(4,197) = 27.96$, $MS_e = 7.76$ in Study 1 and $F(4,122) = 11.15$, $MS_e = 15.61$ in Study 2, $p < .0001$ in both cases). All nonseam transitions produced less change than the seam transition—the critical value of Dunnett's (1955) test at the .05 level is $2.17 in Study 1 and $2.18 in Study 2. We have included all items in this analysis and in Figure 3 (the same will be true, unless explicitly noted, in the analyses to follow). However, as we mentioned earlier (see sec. 2 and footnote 2), some of the items remained constant from week to week, while others actually varied in the questionnaires. Some of the reported changes may therefore correctly reflect true changes. But even if we confine attention to those items with constant values, we find similar seam effects. In Study 1 the mean absolute value of the change is $4.03 at the seam but only $1.74 at the nonseam weeks for the constant items. In Study 2 the corresponding values are $5.05 and $2.14.

Figure 3 also reveals evidence for constant wave responding. The observed change at nonseam weeks is considerably smaller than the true change (dashed line); however, the observed change at the seam weeks approaches the correct value. The seam effect for these data is therefore the result of too few changes across the nonseam weeks rather than too many changes at the seam weeks. Planned comparisons (based on the analysis of variance described above) show that the changes for nonseam weeks are all significantly less than the true value (all t's > 8.5, $p < .001$). There is no significant difference between the observed change at the seam week and the true value in Study 2 ($t(146) = 1.01$, $p > .10$), though the corresponding difference is just significant in Study 1 ($t(233) = 2.08$, $p < .05$).

Clearly, then, respondents were not varying their responses sufficiently within the reference periods, but we can also ask whether the responses were truly constant: How often did respondents give *exactly* the same answer for each week in the reference period? (These are cases, e.g., in which a respondent said that s/he was asked about spending $2 for milk and cream in week 1, $2 in week 2, and $2 in week 3). In examining this issue, we consider only those cases in which the respondent produced a numerical answer for all three weeks within a reference period (week 1 through week 3 or week 4 through week 6); that is, we exclude cases in which an answer is missing for one or more weeks of the period. We also look at just those items that actually varied during the reference period, so that, ideally, no constant responses should appear for these questions. Despite this true variation, however, responses were perfectly constant on 36% of items in Study 1 and 44% in Study 2. These figures are slightly higher than rates for constant responding in our earlier studies of qualitative questions (Rips et al. 2003).

We conjectured that the seam effect depends on respondents forgetting the correct amounts they were to recall, substituting instead estimates based on response strategies like constant wave responding.

To check for forgetting, we can look (within each reference period) at the percentage of items to which respondents gave the correct response. Forgetting should produce a decrease in correct answers as a function of elapsed time between the test session and the relevant questionnaire (see Figure 2). Thus, correct responding should be highest for the most recent weeks in the reference period (weeks 3 and 6), and lowest for the earliest weeks in the reference period (weeks 1 and 4). The data show a modest decrease of this sort in both studies. In Study 1, accuracy was 14.2% for the most recent week, 12.8% for the middle week, and 9.3% for the earliest week. Comparable figures from Study 2 were 10.1%, 10.0%, and 6.8%. The differences between the most recent and earliest weeks are somewhat smaller than those in our earlier studies (6.8-11.9 percentage points), perhaps because of floor effects. To assess this effect, we calculated the percentage of correct answers for each respondent and then transformed the data using logits (see Mosteller and Tukey 1977, ch. 5). The resulting analysis yielded a significant effect of week ($F(2,102) = 11.63$, $MS_e = 1.84$, $p < .0001$ in Study 1 and $F(2,66) = 6.14$, $MS_e = 1.38$, $p = .004$ in Study 2). Overall accuracy is relatively low in both experiments, but this is at least partly due to our strict criterion: Respondents had to recall the exact dollar amount from the questionnaire in order to be scored correct on an item. We report further evidence for forgetting in sec. 5, based on a less stringent criterion.

3.1 Effects of Question Grouping

The aim of Study 1 was to see whether separating questions with the same topic would reduce the size of the seam effect. We expected larger seam effects from respondents who answered questions grouped by topic (e.g., all questions about milk and cream together) than from respondents who answered them grouped by week (e.g., all questions about week 3 together) because the former should encourage constant wave responding. Figure 4 plots the relevant data in the same format as Figure 3 and exhibits the predicted difference between conditions. Blocking questions by week decreases the seam effect both by raising the number of changes for the nonseam weeks and lowering changes for the seam week. As a result, the interaction between type of blocking and week is significant in the analysis of variance described earlier ($F(4,197) = 10.84$, $MS_e = 7.76$, $p < .0001$).

We can use our measures of constant wave responding and forgetting to get a clearer picture of this effect of question grouping. As expected, constant wave responding was much more frequent when questions were blocked by topic than when they were blocked by week. Confining attention again to those items that actually varied during the reference period, we find that respondents gave constant answers to 48.7% of questions when blocked by topic, but to only 19.4% when blocked by week. Overall accuracy, however, was not greatly different for these two conditions. Respondents whose questions were blocked by topic gave 13.1% correct answers, and respondents whose questions were blocked by week gave 10.9% correct answers. This difference was marginal in the above analysis ($F(1,51) = 2.97$, $MS_e = 2.98$, $p = .09$).

It is easy to understand why blocking by week increased the amount of change for nonseam points. Placing the items about milk and cream, for example, in different parts of the interview makes it difficult for respondents to give the same answer to all of them. But why should blocking by week produce *less* change at the seam than does blocking by topic? One possibility is that both groups of respondents retain approximately the same sense of how much change occurs in the questionnaires from week to week but distribute their changed responses differently. When items with the same topic appear together, as they do in the blocked-by-topic group, constant wave responding may tempt respondents to displace any changes to the seam. By contrast, respondents in the blocked-by-week group may be more willing to place the changes between both seam and nonseam weeks.

3.2 Effects of Dependent Interviewing

Study 2 was devoted to the issue of whether dependent interviewing moderates the seam effect. It is reasonable to suppose that dependent interviewing would reduce the size of the effect by reminding respondents of their responses for the earlier reference period. Respondents in the feedback condition were told during the second test session what their responses had been in the earlier test session for week 3. They might learn, for example, that they had answered “$2” to the milk-and-cream question for week 3 before being asked about milk-and-cream for week 4. This reminder might lead them to provide a similar response for these two weeks by anchoring their estimates, thus reducing the seam effect. Respondents in the no-feedback condition, of course, received no such information about their earlier responses and should produce a seam effect of the usual size.

As we've already noticed, the seam effect was robust in this study (see Figure 3), and dependent feedback had a somewhat limited ability to decrease it. In the feedback condition, mean absolute change was \$4.22 at the seam and was \$1.86 elsewhere. In the no-feedback control, mean absolute change was \$5.09 at the seam and \$2.23 elsewhere. Although the seam-nonseam difference is in the expected direction—slightly larger in the no-feedback condition—the interaction between week and condition was not significant in the analysis described earlier ($F(4,122) < 1$). Neither was a planned comparison between the seam versus nonseam weeks in the two conditions ($F(1,122) < 1$). Effects of dependent feedback did manifest themselves, however, for those items that had constant values. As Figure 5 shows, the seam effect was much larger for the constant items in the no-feedback than in the feedback condition. At most of the nonseam transitions, the two groups of respondents performed similarly, but the no-feedback group produced a larger mean change at the seam than did the feedback group. A contrast like the one just described was significant for the constant items ($F(1,438) = 12.11$, $MS_e = 7.71$, $p < .01$). These results suggest that the actual changes in amounts for the variable items partially masked the difference between conditions, and in accord with this, the triple interaction between item type (constant vs. variable), condition (feedback vs. no-feedback), and week was marginally significant ($F(16,438) = 1.52$, $MS_e = 7.71$, $p = .09$).[4] The finding that dependent interviewing reduces but does not eliminate the seam effect is consistent with recent results from both experimental surveys (Jäckle 2006) and actual surveys (Hill 1994; Moore et al. 2006).

As we would expect from the somewhat similar performance of the two groups at nonseam weeks, constant wave responding was about equally frequent for those in the feedback and in the no-feedback conditions. Respondents in the feedback group produced 46.7% constant wave answers to the constant items and 42.0% constant wave answers to the variable items. For the no-feedback group, these figures were 45.7% and 47.8%, respectively. Similarly, there were no significant differences in accuracy between the two groups. Respondents in the feedback condition provided correct answers to 9.1% of items, and those in the no-feedback condition provided correct answers to 8.8%. This difference was not significant in our logit analysis, and there were no significant interactions between groups and other variables. ($F < 1$ for interactions between group and type of item (constant or variable), group and reference period, and group and week within the reference period. Higher-order interactions were also nonsignificant). The low accuracy for both groups again reflects the relative difficulty of correctly guessing exact numbers as opposed to qualitative answers like “yes” and “no.”

This last result may seem surprising, given the feedback group’s smaller seam effect in Figure 5. But recall that the dependent facts we provided to respondents were the answers they had produced in the earlier test session, and this information was not always correct. If respondents adjusted their current answers to conform to the dependent information, they would not necessarily be converging on a correct response. To put this is in a slightly different way, the feedback condition exhibits correlated error across the seam, whereas the no-feedback condition exhibits uncorrelated error. The likelihood of producing an error and the size of the errors are approximately the same in the two groups, but the feedback condition yokes the errors across the seam because of the presence of the dependent information. (Despite propagating error from earlier responses, dependent interviewing can in principle improve measures of change even if the exact levels are not correct. For example, respondents may know that they spent more money in the most recent period than in the period from which their response was fed back; thus they can formulate their current response by incrementing their previous answer, accurately reflecting change, despite error in the anchor.)

4. A Model for Quantitative Seam Effects

One way of explaining seam effects is to assume that respondents begin by attempting in good faith to remember the queried information for the most recent part of the reference period. If they are successful in doing so, they report this information; but if they are unsuccessful in remembering relevant facts, they make use of constant wave responding or other heuristics. This idea informed our earlier account of the seam effect for yes/no questions (in Rips et al. 2003), and it is of interest to see whether it can also be used to explain seam effects for quantitative data.

A crucial issue in applying this model is what answer respondents give when they fail to remember any relevant facts. In the present studies, all questions were about amounts, and the actual values varied within a fairly narrow range (from $1 to $42). It seems possible, under these conditions, that respondents would retain some estimate of the items' central tendency—a subjective value of the usual amount they encountered. (In what follows, we refer to this estimate as a "subjective average," but we do not mean to imply that respondents necessarily calculate this value explicitly nor that it corresponds exactly to a statistical measure of central tendency, such as a mean or median.) Barring nonresponses, respondents could then use this subjective average if no further information came to mind about the quantity in question. Repeated use of this same estimate within a reference period would produce the constant wave responding that we observed. It is also possible that a respondent's subjective average could differ from one test session to the next, contributing to the size of the seam effect. For example, if the subjective average for the first reference period is $9 and the subjective average for the second is $7, then the $2 difference could contribute to the change at the seam.

Respondents who are successful at remembering the value from the most recent week (week 3 or 6) could, of course, produce that value as their response. It might also seem reasonable to expect these same successful respondents to attempt to retrieve correct values from earlier weeks as well (weeks 1-2 and 4-5). Preliminary modeling suggests, however, that respondents make little use of earlier correct amounts. We can give a better account of the results by supposing that respondents provide the most recent week's value if they can remember it, either persisting with this same value for earlier weeks or reverting to the subjective average.

We can assume, in particular, that respondents give the value of the most recent week with a probability, p_i, that decreases with elapsed time since the queried week, $week_i$ (1, 2, or 3), within the reference period:

$$p_i = \exp(-b(t - week_i)). \quad [1]$$

Here, *b* and *t* are parameters of the exponential decay function. Exponential decay of this type has previously been used successfully in a survey context to model effects of forgetting (Sudman and Bradburn 1973; see Tourangeau, Rips, and Rasinski 2000, ch. 3, for a review). With probability $1 - p_i$, respondents will produce the subjective average as their response, rather than the most recent value. If we let A_j be this subjective average for reference period *j* and F_{jk} the true value for the final week of reference period *j* for item *k*, then Equation [2] gives the predicted value of the response:

$$R_{ijk} = p_i F_{jk} + (1 - p_i) A_j \qquad [2]$$

The model predicts that the respondent's answer will be a weighted average of the value for the most recent week of the reference period (F_{jk}) and the subjective central value (A_j), where the latter will tend to dominate for early weeks of the reference period.

As an example of the applicability of this model, we can use it to predict some of the results we have just reported. For these purposes, we chose the data from those respondents in Study 1 whose questions were grouped by topic. As we've noted, grouping by topic is the usual method in surveys, and because a larger number of respondents took part in this condition in Study 1 than in the comparable (no feedback) condition in Study 2, the resulting data are more stable. We averaged responses across questions, separately for the two reference periods and for those items that actually increased, decreased, and stayed constant during those periods.[5] The resulting means appear as points in Figure 6, with upward pointing triangles for the increasing items, downward pointing triangles for the decreasing items, and circles for the constant items. Several trends are worth noting in these data. First, the means for the increasing and decreasing items vary in the correct direction: Responses to increasing items increase and responses to decreasing items decrease within the reference periods. Respondents were therefore sensitive, at least to some extent, to the actual values of what they were to report. Second, however, the reported changes are much smaller than the changes to the actual amounts. For the increasing items, for example, the mean actual value was \$4.14 for the first week, \$8.25 for the second week, and \$16.50 for the third week of the reference periods. The means for the decreasing items are the reverse of these. The observed responses never approach the low or high value, but tend to diverge during the reference period from an intermediate position. Third, the mean responses are slightly offset for the two reference periods. For example, responses to the constant items are somewhat higher for the first reference period than for the second, despite the fact that the actual values do not differ.

Predictions from the model appear as lines in Figure 6, and they accord with the trends just described. To derive these predictions, we employed Equations [1] and [2], estimating the parameters of the exponential, b and t, and the subjective average values, A_1 and A_2. Because these equations are nonlinear, we employed the Gauss-Newton estimation method, as implemented in SAS's PROC NLIN. This model provides a good fit to the data, $F(4, 14) = 5013.28$, $p < .0001$, with a mean squared error of 0.054, and it accounts for 93% of the variance among the means in Figure 6. The obtained values of the parameters are: $b = 0.59$, $t = 5.65$, $A_1 = 7.60$, and $A_2 = 7.25$. The model yields the increasing and decreasing trends because earlier weeks in the reference period are heavily influenced by the subjective average values (A_1 or A_2), whereas recent weeks depend to a greater extent on the retrieved value from the final week (F). For example, consider the increasing items (upward pointing triangles) in the second reference period. As just noted, the true average value for these items in week 6 was $16.50. According to Equation [2], the observed value for this week should be a compromise between this true value and the subjective average, $7.25, where the two values are weighted by p_3 and $1 - p_3$ (since this is the third week of the reference period). By Equation [1], $p_3 = .21$, given the obtained parameter values; so the predicted value is .21*$16.50 + .79*$7.25 = $9.19. For week 4, however, the values are weighted by p_1 and $1 - p_1$ (since this is the first week of the reference period), and the value of p_1 is .06. Hence, the predicted value is .06*$16.50 + .94*$7.25 = $7.80.

Shrinkage of the observed responses with respect to the true values is also due to respondent's reliance on the subjective average. According to the values of the parameters just given, respondents are using the subjective average on 94% of the items for the first week of the reference period (i.e., $1 - p_1 = .94$), 88% of the items for the second week, and 79% for the third week. Thus, even for the most recent week, respondents usually give an averaged response. Finally, the offset between the reference periods is the result of the difference between the subjective averages, A_1 and A_2.

The results of the model fitting should, of course, be viewed cautiously. Our assumption about decreased use of actual values as a function of time was based on our earlier work, but the assumptions about subjective averages as estimates were partly the result of the trends we observed in the present experiments. It is well worth determining whether these latter assumptions extend to other data sets. Likewise, it may be useful in further studies to ask respondents to think aloud during the test session to observe whether respondents report strategies like constant wave responding or use of a subjective average. Respondents may, of course, employ these strategies in an automatic and unconscious way. But it is also possible that they have cognitive access to these short-cuts, and if so, their think-aloud protocols may shed light on the model's assumptions.

In addition, inspection of the deviations between predicted and observed values in Figure 6 suggests that the present model fits the data less well in the first week of reference period 1 than of reference period 2. In particular, the model misses the fact that responses tend to converge more closely during the first week of reference period 1 (week 1) than in the first week of reference period 2 (week 4). This difference may be due to the fact that respondents did not know prior to the first test session that they would be tested on the questionnaire values and so had relatively poor memory for the questionnaire items. This would have led them to rely rather heavily on subjective averages for those items, more heavily than assumed in the model. By the time of the second test session, however, they knew what to expect, and this may have made them less likely to rely on subjective averages as responses in the second test. These deviations could be handled within the model's framework by estimating the exponential decay parameters separately for the two reference periods, but at a cost of additional free parameters.

5. Summary and Implications

Seam effects appear nearly universally in panel surveys that collect data both within and between reference periods. Similarly, all the experimental studies we have conducted to date have obtained seam effects—larger changes in responses when the data come from two different reference periods than from the same reference period. Seam effects appear both for questions that demand recall (e.g., reports about the content of an event) and for those that demand recognition (e.g., yes/no responses; see Rips et al. 2003). They occur under different ways of grouping or ordering questions, although some of these manipulations affect the size of the effect. In most of our studies, including the present ones, the key questions that produced the seam effect concerned information that we had supplied. However, we have also reported similar findings for naturalistic events, like those used in actual surveys (Rips et al. 2003, Study 2). The present studies extend our results to questions about quantitative data (amounts paid for goods, in this case), and they show that the seam effect appears for these items as well. There is little reason to doubt the stability of the effect.

The results of these studies implicate the same factors that contributed to seam effects for qualitative responses. Both studies turned up evidence of constant wave responding. Across all items, responses underestimated the true week-to-week changes, except for seam weeks (see Figure 3). As we reported earlier, respondents in Study 1 gave perfectly constant responses to 36% of items that actually varied within the weeks of a reference period. This percentage increases slightly (to 38%) if we adopt a less stringent criterion, counting as constant those responses that varied less than 10% from the respondent's mean for a particular item. And it increases to 57% if we count as constant those responses within 20% of the mean. Similarly, both studies obtained evidence for forgetting of the original amount. Although respondents did not often give exactly the correct amount, their accuracy decreased significantly as a function of the time between the questionnaire and the test session. Much the same trend is apparent if we relax our criterion for a correct answer. If we score responses as correct if they are within 20% of the correct amount, then respondents produce correct responses for 23% of items for the most recent week of the reference period and 17% for the earliest week. According to the theory developed in the preceding section, these results are a consequence of respondents using the amounts for the most recent week of the reference period when they are able to remember them, but regressing toward an average amount when no pertinent information comes to mind.

An important departure from earlier results, however, is that the present studies found little evidence for overestimates of change at the seam. Seam effects in both Studies 1 and 2 are largely due to underreporting of true change at nonseam transitions, whereas both earlier experimental studies (Rips et al. 2003) and record-checks of actual surveys (Marquis and Moore 1990; Moore and Marquis 1989) show both underreporting at nonseam points and overreporting at the seam. This difference may be attributable to the properties of quantitative versus qualitative questions, since all the earlier studies just cited concern qualitative items (reports of program participation in the case of the Marquis-Moore reports). It is possible that quantitative questions produce less change in general, tipping the balance of errors toward underreporting. An exact figure (e.g., \$8.50) may be relatively difficult to recall, and respondents may therefore be tempted to rely heavily on responses based on the subjective average for most intervals, in line with the model-fitting results of sec. 4. It is true that, according to the model, the estimated subjective averages differ somewhat ($A_1 – A_2$ = 35¢) between reference periods, and this contributes to the change at the seam. Nevertheless, this change is dwarfed by the actual week-to-week change for our items—approximately \$5, as shown by the dashed line in Figure 3. Relying on the subjective average will therefore tend to suppress change relative to the true differences at all intervals.

The present studies also point to two ways to reduce, though not eliminate, the seam effect for quantitative items. Separating questions about the same topic (e.g., expenditures for electricity in three different months) reduces the likelihood that respondents will give the same answer to each of these questions. Dividing up the items increases nonseam changes and reduces seam changes, thus making the seam effect less pronounced (see Figure 4). As a kind of converse, dependent interviewing reminds respondents of their answers from the previous reference period, decreasing seam changes, at least for items that have actually constant values over the reference period (see Figure 5).

But although these procedures decrease the seam effect, the results are not so clear about their effects on accuracy. Because we have access to the correct values for respondents' answers in these studies, we can determine whether reducing the seam effect also improves the overall quality of the responses. Separating questions about the same topic in Study 1—grouping by week—resulted in a marginal decrease in accuracy, while providing dependent information increased it by a nonsignificant amount. Although the benefits may outweigh the costs of these procedures, we need to be cautious about the possibility that errors may be shifted from one interval to another without being eliminated. It may seem disappointing that the size of the seam effect isn't a clear indicator of overall data quality. However, the seam effect depends on the variability of responses from month to month, and we shouldn't expect variability (or stability) to correlate perfectly with the responses' accuracy.

Of course, these results should not be taken to mean that dependent interviewing (or other measures) cannot improve data quality. There are many forms of dependent interviewing, and the individual forms may have differential effects on particular kinds of questions (Jäckle 2006). For example, surveys that probe whether respondents participate in each of many programs or whether they derive income from each of many sources may benefit from dependent interviewing. Feedback about the respondents' earlier status may remind them of their present one, overcoming memory loss for less important items. Most reports of successful dependent interviewing concern data of this type (Hill 1994; Lynn et al. 2006; Moore et al. 2006). Likewise, reactive dependent interviewing may help resolve inconsistencies from neighboring interviews. Our own questions asked about dollar amounts that changed in ways that respondents were probably not able to predict, and the respondents were faced with the task of determining an exact magnitude. Feedback about an earlier magnitude may be less helpful in this setting, since it may not have furnished respondents with useful information that they did not already possess. However, our results do provide a caveat about dependent interviewing, suggesting that researchers shouldn't be too quick to take reduction of the seam effect as evidence of the success of a dependent procedure.

We can probably find new ways to reduce seam effects by counteracting biases (e.g., the constant wave tendency) in respondents' answers. One promising technique may be the use of event-history calendars (see Callegaro 2007 and Kominski 1990 for efforts in this direction), which provide forceful memory aids for queried information. There may be limits, however, to what such techniques can do for respondents who are faced with the difficult task of recalling amounts. Data from this study suggest that respondents' memory was quite poor for such information, and while we might hope for better performance for more important amounts (e.g., salaries), many quantitative facts may prove quite hard to recover. Furthermore, in implementing these techniques we need to be careful that we don't also introduce further sources of error. Our hope is that the experimental method we are developing in this and our earlier studies can serve as a useful way of determining whether proposed techniques will impact other aspects of data quality. As such, methodologists could use the method alongside other procedures, such as cognitive interviewing, for anticipating problems in actual surveys.

References

Burkhead, D., and Coder, J. (1985). Gross Changes in Income Recipiency from the Survey of Income and Program Participation. Proceedings of the Social Statistics Section. Washington, D.C.: American Statistical Association.

Callegaro, M. (2007). Changes in Seam Effects Magnitude Due to Changes in Question Wording and Data Collection Strategies: An Analysis of Labor Force Transitions in PSID. Paper to be presented at the Annual Conference of the American Association for Public Opinion Research.

Dunnett, C. W. (1955). A Multiple Comparison Procedure for Comparing Several Treatments with a Control. Journal of the American Statistical Association, 50, 1096-1121.

Hale, A., and Michaud, S. (1995). Dependent Interviewing: Impact on Recall and on Labour Market Transitions. SLID Research Paper 95-06.

Hill, D. H. (1987). Response Errors around the Seam: Analysis of Change in a Panel with Overlapping Reference Periods." Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Hill, D. H. (1994). The Relative Empirical Validity of Dependent and Independent Data Collection in a Panel Survey. Journal of Official Statistics, 10, 359-380.

Jabine, T. B., King, K. E., and Petroni, R. J. (1990). Quality Profile, Survey of Income and Program Participation. Washington, D. C.: Bureau of the Census.

Jäckle, A. (2006). Dependent Interviewing: A Framework and Application to Current Research Working Paper of Institute for Social and Economic Research, paper 2006-32. Colchester: University of Essex.

Kalton, G., and Miller, M. E. (1991). The Seam Effect with Social Security Income in the Survey of Income and Program Participation. Journal of Official Statistics, 7, 235-245.

Kominski, R. (1990). The SIPP Event History Calendar: Aiding Respondents in the Dating of Longitudinal Events. Proceedings of the Section of Survey Research Methods (pp. 553-558). Washington, D. C.: American Statistical Association.

Lynn, P., Jäckle, A., Jenkins, S. P., and Sala, E. (2006). The Effect of Dependent Interviewing on Responses to Questions on Income Sources. Journal of Official Statistics, 22, 357-384.

Marquis, K. H., and Moore, J. C. (1989). Response Errors in SIPP: Preliminary Results. Proceedings, 1989 Annual Research Conference, U. S. Bureau of the Census. (Arlington, VA, March 19-22, 1989). Washington D.C.: U. S. Bureau of the Census.

Marquis, K. H., and Moore, J. C. (1990). Measurement Errors in the Survey of Income and Program Participation (SIPP) Program Reports. Proceedings, 1990 Annual Research Conference, Washington, DC: U.S. Bureau of the Census.

Martini, A. (1989). Seam Effect, Recall Bias, and the Estimation of Labor Force Transition Rates from SIPP. Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

Mathiowetz, N. A., and McGonagle, K. A. (2000). An Assessment of the Current State of Dependent Interviewing in Household Surveys. Journal of Official Statistics, 16, 401-418.

Moore, J., Bates, N., Pascale, J., Griffiths, J. K., and Okon, A. (2006). Use of Dependent Interviewing Procedures to Improve Data Quality in the Measurement of Change. Research Report, Survey Methodology #2006-02. Washington D. C.: Census Bureau.

Moore, J. C., and Marquis, K. H. (1989). Using Administrative Record Data to Evaluate the Quality of Survey Estimates. Survey Methodology, 15, 129-143.

Mosteller, F., and Hoaglin, D. C. (1991). Preliminary examination of data. In Fundamentals of Exploratory Analysis of Variance, ed. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, New York: Wiley.

Mosteller, F., and Tukey, J. W. (1977). Data Analysis and Regression. Reading, MA: Addison-Wesley.

Polivka, A. E., and Rothgeb, J. M. (1993). Redesigning the CPS Questionnaire. Monthly Labor Review, 116, 10-28.

Rips, L. J., Conrad, F. G., and Fricker, S. S. (2003). Straightening the Seam Effect in Panel Surveys. Public Opinion Quarterly, 67, 522-554.

Sudman, S., and Bradburn, N. (1973). Effects of Time and Memory Factors on Response in Surveys. Journal of the American Statistical Association, 68, 805-815.

Tourangeau, R., Rips, L. J., and Rasinski, K. (2000). The Psychology of Survey Response. Cambridge, UK: Cambridge University Press.

Tversky, A., and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. Science, 185, 1124-1131.

Wilson, T. D., Houston, D. E., Etling, K. M., and Brekke, N. (1996). A New Look at Anchoring Effects: Basic Anchoring and Its Antecedents. Journal of Experimental Psychology: General, 125, 387-402.

Young, N. (1989). Wave-seam Effects in the SIPP. Proceedings of the Section on Survey Research Methods, American Statistical Association. Alexandria, VA: American Statistical Association.

Footnotes

[1] This distinction is blurred by so-called retrospective panel surveys in which researchers conduct just one interview with each respondent but collect reports about multiple time periods, in effect, building up a longitudinal database from a single data collection session. Our concern here, however, will be solely with prospective panel surveys, which have the structure described above.

[2] The amounts for these variable items changed according to four patterns. One group of items increased in amount for weeks 1-3 and increased again for weeks 4-6; a second group increased for weeks 1-3 and then decreased for week 4-6; a third group decreased, then increased; and a fourth group decreased then decreased. For example, if the item about milk and cream was in the increase-increase group for a specific respondent, that respondent was asked on the first questionnaire, *Last week, did you spend more or less than $1 for milk or cream...?*, on the second questionnaire, *Last week, did you spend more or less than $2 for milk or cream...?,* and on the third questionnaire, *Last week, did you spend more or less than $4 for milk or cream...?*. This same sequence was then repeated for weeks 4-6. A respondent saw an equal number of items in each of the four groups. Individual items were rotated through the item groups across respondents in order to control for content.

[3] It is more usual in production to apply dependent interviewing to qualitative rather than quantitative responses, perhaps because of the fear that dependent interviewing will increase constant wave responding for quantitative values. One of the aims of the present study is to see whether these fears are well-grounded. It is good to bear in mind, however, that the effects of dependent interviewing on quantitative responses that we found in this study may differ from those on qualitative responses.

[4] We calculated the error term for the last two tests from the interaction between item type (constant + four types of variable items), week (1-2, 2-3, 3-4. 4-5. 5-6), condition (feedback vs. no-

feedback), and respondent within condition. Hence, the larger number of degrees of freedom than in the earlier tests, which did not involve item type.

[5] Because the data were highly skewed, we first took logs of the respondents' answers and then removed outliers. Let Q_1 be the first quartile and Q_3 the third quartile of the distribution of log values within each condition. Then responses were omitted if their logs were greater than $Q_3 + 1.5*(Q_3 - Q_1)$ or less than $Q_1 - 1.5*(Q_3 - Q_1)$; see Mosteller and Hoaglin (1991). Anti-logs of the means of the resulting data appear in Figure 6 and were used to fit the model.

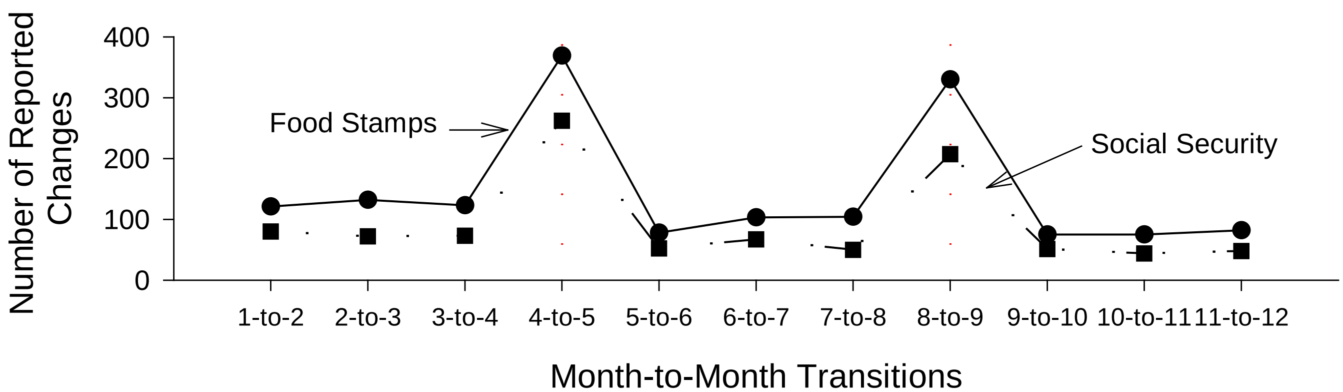

Figure 1. Month to month changes in reports of receiving food stamps (circles) and social security benefits (squares) in SIPP (from Burkhead and Coder 1985). Dashed vertical lines mark seams, where change is computed with data from two different interviews.

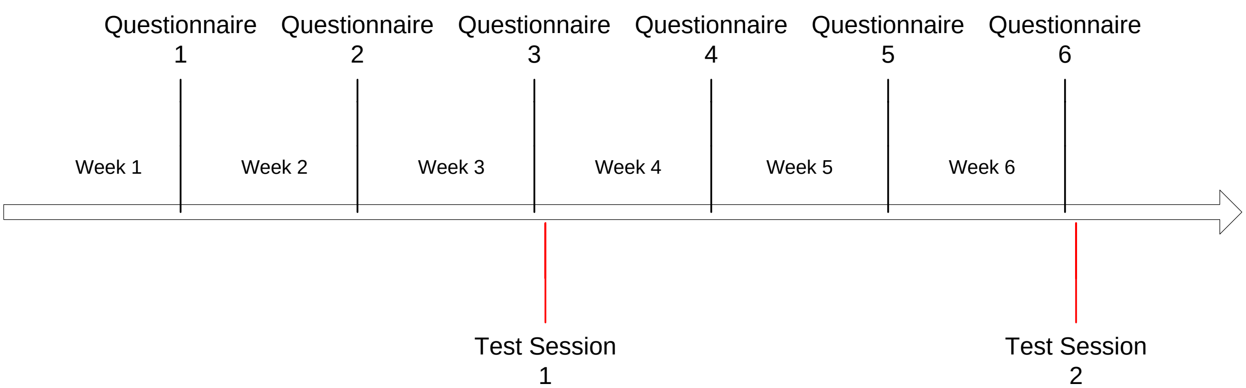

Figure 2. Timeline of events in Studies 1 and 2. Vertical lines at top show approximate time at which questionnaires were mailed to respondents. Vertical lines at the bottom show approximate time of the two test sessions.

Figure 3. Mean week to week change in amounts produced during test sessions for Study 1 (filled circles) and Study 2 (open circles). Error bars represent ±1 standard error of the mean.

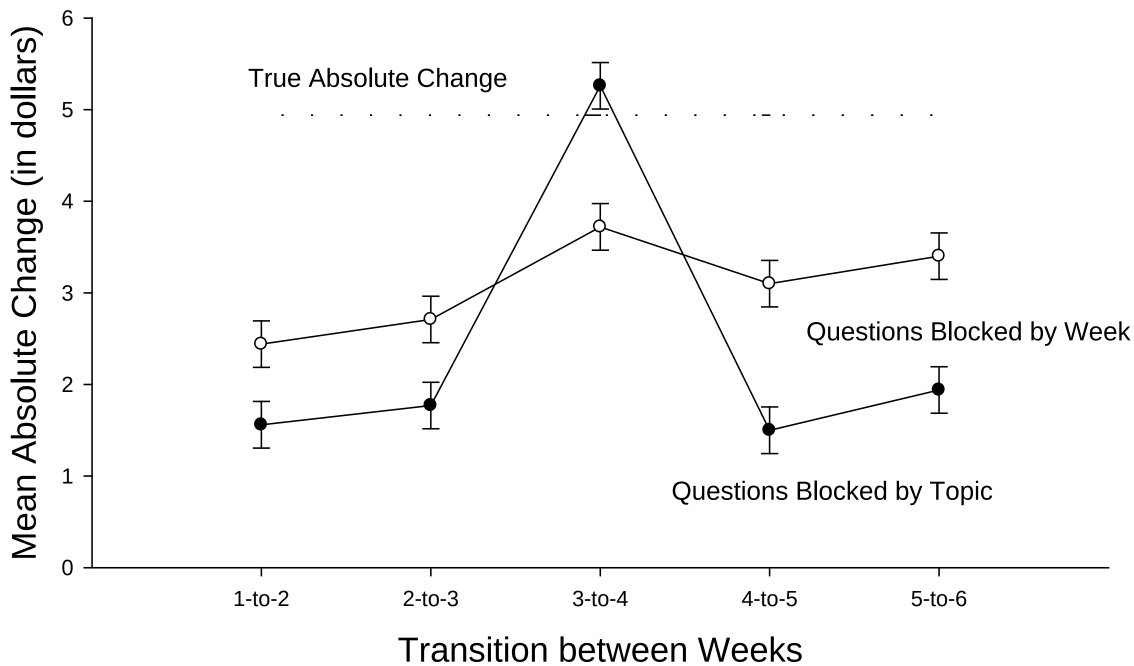

Figure 4. Mean week to week change in amount in Study 1 as a function of whether test session items were blocked by week (empty circles) or blocked by topic (filled circles). Error bars represent ±1 standard error of the mean.

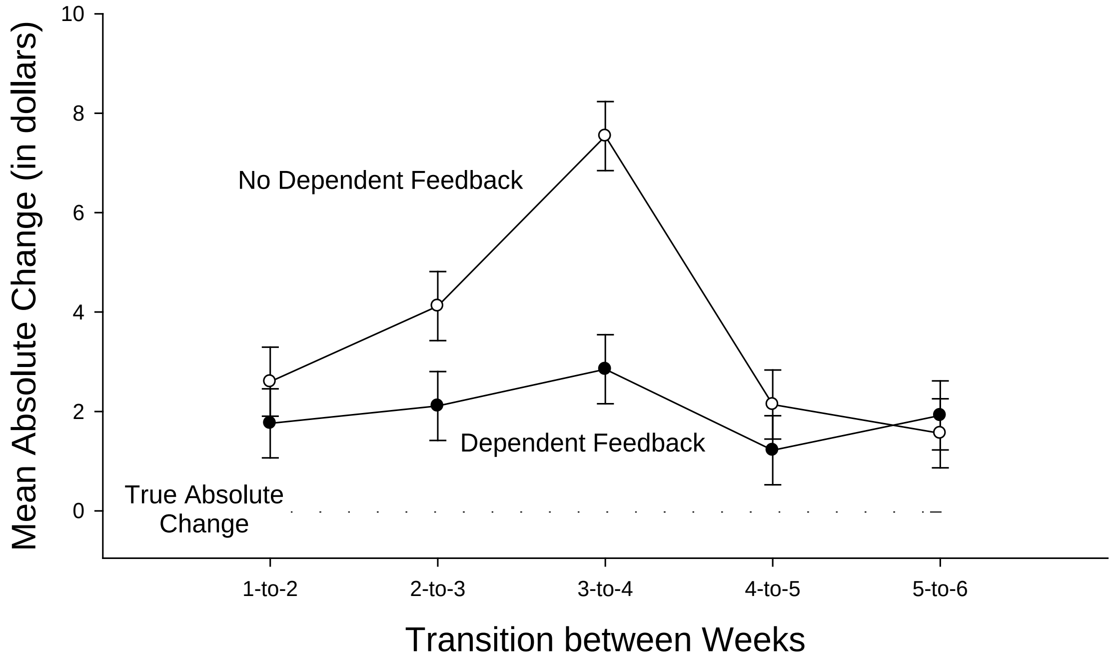

Figure 5. Mean week to week changes in amounts in Study 2 for those items whose true values were constant. Filled circles are results from the group receiving dependent interviewing; empty circles are results from the group receiving independent interviewing. Error bars represent ±1 standard error of the mean.

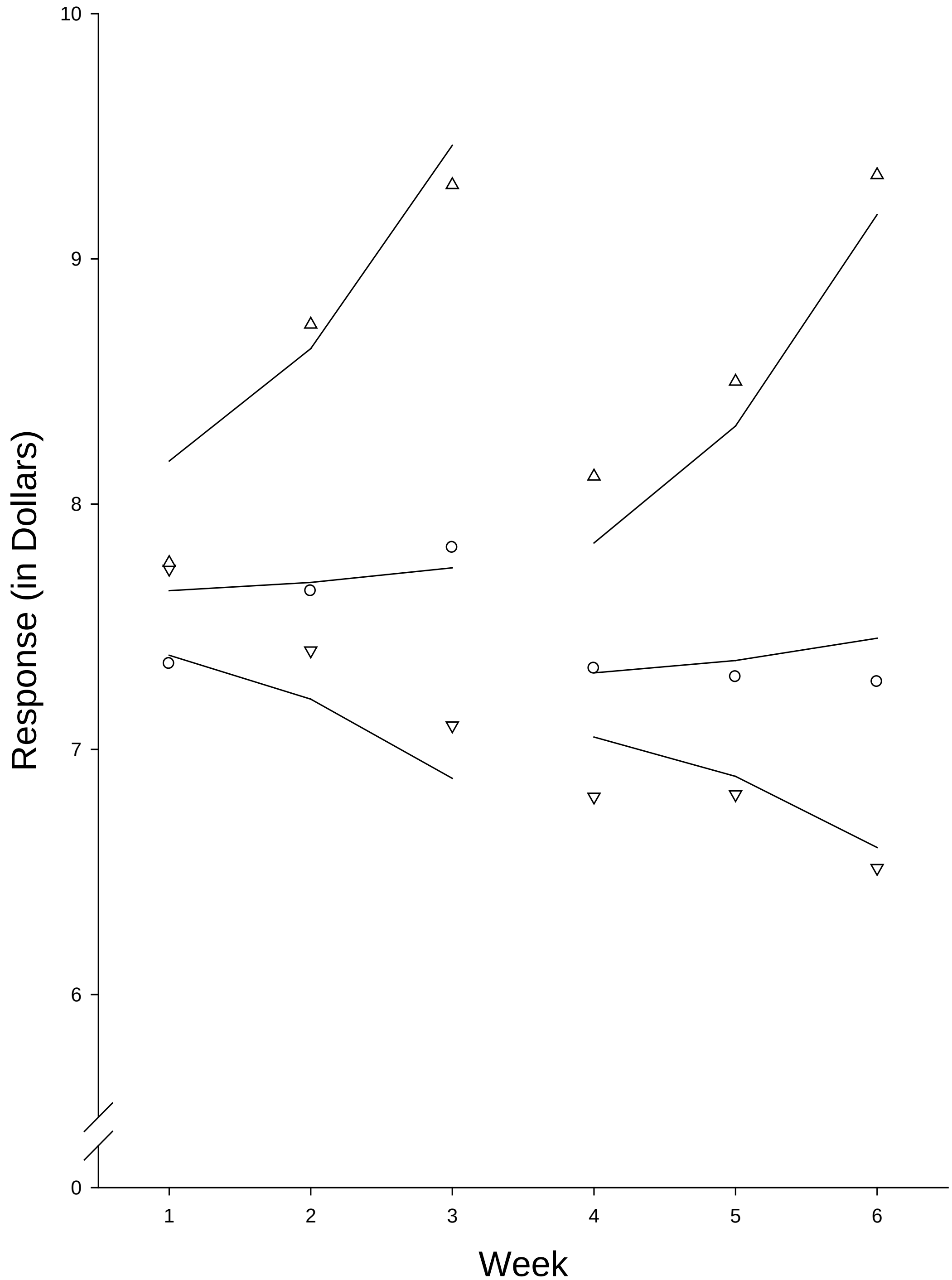

Figure 6. Mean responses from Study 1 for respondents whose items were blocked by topic. Upward-pointing triangles indicate items whose values actually increased during the reference period, down-ward pointing triangles indicate items whose values decreased, and circles indicate constant items. Lines are fitted values from the model described in the text.