Pre-printing effects in official statistics, an experimental study

Authors: Anders Holmberg[1] [2]

Keywords: Use of historical data, Experimental design, Measurement errors, Response variability

Abstract

In surveys where respondents are contacted repeatedly, information from previous data collections may be used in the following data collections. The responses then become dependent on the presentation and the quality of that information. Normally, the presented information is historical data concerning older reference periods and besides providing data for a current reference period, the respondents can verify and (if necessary) change this 'historical' data. The motives for presenting historical data can be that: It increases the efficiency of the data collection. It can correct previous errors. It reduces the response burden and sometimes it is believed that it reduces measurement errors and spurious response variability. The possible drawbacks are that it can conserve errors rather than correct them and it might lead to underreporting of changes from one period to another. Here, we focus on methodological issues of pre-printing 'historical' values on self-administered, (electronic or paper), questionnaires for business establishments. A planned experiment made in an ongoing large-scale survey indicates that on several aspects of data quality, questionnaires with pre-printed historical values outperform questionnaires without pre-printed data. We present the main results of the experiment, as well as a general discussion of pre-printing experiences at Statistics Sweden.

[1] Department of Research and Development, Statistics Sweden, SE-70189 Örebro, Sweden
[2] The author gratefully acknowledges the work from Lars Lindam-Olsson and Kerstin Fredriksson at Statistics Sweden and Marie Melkersson at the Employers Association of Property Owners. Without their contribution in the data collections, these studies had been impossible.

1 Introduction

When we collect data for official statistics, *pre-printing* (or its cousin dependent interviewing) is a frequently applied technique. The pre-printed information is historical data from older reference periods than the primarily requested reference period. Besides providing data for a new (up to date) reference period, respondents sometimes are asked to verify and (if necessary) change the pre-printed 'historical' data. The type of data that is pre-printed can be administrative variables such as addresses and names, but here we focus on the case when more or less complex survey variables are pre-printed. The typical situation is data collections repeated over time involving common objects (e.g. surveys with panel designs or repeated business surveys with take-all strata.)

When the requested information is repeated for the same object and when other survey circumstances are similar, most methodological papers focus on response variability (variation between repeated measurements) or response correlations. With few exceptions (e.g. O'Muirchetaigh (1986), Pafford (1986, 1988) and Stanley and Safer (1997)), these papers do not study the effects of using historical data, pre-printed on self-administered questionnaires or put forward in dependent interviews. However, the use of historical data in surveys is not new. Already, Hansen, Hurwitz and Pritzker (1964) discussed and questioned the effects of using historical data in surveys and since then various case studies have been done. (Mathiowetz and McGonagle (2000) give a review of case studies in household surveys; Corti and Campanelli (1992) summarize experiences of feeding forward earlier data in different panel surveys.)

In this paper we present the main results of a pre-printing experiment made within an on-going large-scale survey. The experiment was designed to study certain effects of pre-printed self-administered questionnaires and it was a part of a larger project undertaken at Statistics Sweden.

2 Motives for Using or not Using Historical Data

The motives for pre-printing historical data or not vary between the type of survey and the type of variables. From a review of the pre-printing practices of different statistical products at Statistics Sweden we can summarize these motives as follows (Holmberg (2002)):

- Respondent support: Pre-printing can support the respondent in a variety of ways, some examples are:
    - Reducing the response burden: Frequently asking the respondent to fill in the same (hardly ever changing) statistics can seem unnecessary from a respondent perspective. Since respondents simply can verify that the pre-printed values still are valid or update them if necessary, pre-printing available data avoid this.
    - Questionnaire guidance, memory support, anchoring: Pre-printing can help the respondent to recognise what information is requested, where to put it in the questionnaire, and whether the intended response is reasonable compared to the pre-printed information. Pre-printing can also help to keep track of events in time (e.g. if an event already has been reported or not.)
    - Feedback purposes: Sometimes pre-printing is an appreciated feedback to the respondents.
- Improved efficiency in the data collection: Depending on the shape of the statistical production process pre-printing can improve some factors, e.g. easier communication in call-back interviews, a decreased risk of coding errors and less required editing due to early stage involvement of respondents.
- Reduction of measurement errors and improved data quality: Pre-printing makes it possible for the respondent to check, react on and correct the previously collected information. The respondent support (above) also reduces measurement errors, since unwanted variation is

believed to decrease with pre-printing. The unwanted variation can come from spurious reports of changes or large inconsistencies in data over time.

The main motives for **not** pre-printing historical data are:

- Risk of bias due to underreporting of changes and conservation of errors: Edits, follow-ups and evaluations have detected phenomena such as frequent unmotivated pre-printing recurrences and respondent's unwillingness to deviate too much from previously reported pre-printed values. For variables whose values with certainty change between reference times, an error most certainly occurs when a response is an exact repetition of a pre-printed value (which normally refers to a previous reference time.) The respondent may find the pre-printed information fair or close enough and he is released of some response burden by using it. (Krosnick (1991) use the term 'satisficing' when a respondent choose a cognitive easier way to respond.) Unclear or skipped instructions are other reasons why pre-printed data incorrectly survives from one reference period to the next. The extent and total effects of these errors are often unknown, and especially for categorical variables that seldom change value over time they can be hard to detect.
- Loss in confidence and goodwill: If the pre-printed data has poor quality or if it is unrecognised by the respondent, there is a risk that the co-operation of the respondents is negatively affected.
- Disclosure risk: Even with good safety routines, there is always a risk that the pre-printed data is disclosed to outsiders.

In addition to the motives above economical considerations play an important role. Whether pre-printing is seen as more expensive or not depends on a lot of factors, e.g. the type of survey, its size, the environment and tools used in the survey process, the actual effects of the pre-printing and of course which quality aspects and cost functions that are used for evaluations. Economical aspects of pre-printing are not treated in this paper.

3. Background of the Pre-printing Experiment

The Survey of building rental units (in Swedish 'Lokalhyresundersökningen' (LHU)) collects information of total monthly rental incomes and rented area in square metres for commercial rentals of premises. Industries, shops and offices are the main categorization of the premises and the data is used to produce a Fisher price index for building rentals, which is used in the National Account statistics. The LHU-survey is conducted once a year and independent probability samples of building rental units are drawn every year. However, the number of building units that are common in two successive samples is considerable due to the sampling design, (stratified sampling with x-optimal allocation (Särndal, Swensson and Wretman (1991) p107).) For these common building units it is possible to pre-print the reported data from the previous year. In the survey of 2001 (henceforth named LHU 2000 after the reference year) a planned experiment was done to study the effects of pre-printing.

The data collected in one LHU-questionnaire stretches over a period of 14 months. The first two of these months are overlapping with the last two months of LHU-questionnaire from the year before. Hence, for the overlapping months there is a repeated measurement for the building units that are included in two successive years. Figure 1 gives a schematic idea of how this works for a building unit included in both the LHU 1999 and the LHU 2000 surveys.

FIG 1 in here

As illustrated in Figure 1, the data collection for the LHU 2000 survey is done in the beginning of 2001. The interesting reference months are December-99, January-00, December-00 and Januari-01, which are used for the Fisher index. If a building unit in the LHU 2000 sample reported to the LHU 1999 survey, data of the overlapping reference months (i.e. December-99 and January-00) are

available for pre-printing. If the unit then responds to the LHU 2000, we will have repeated observations over two surveys and it is possible to estimate the response variability. Clearly, if no measurement errors are present, data for December-99 and January-00 should be identical in LHU 1999 and LHU 2000. The first comparison in our study (indicated by the C1 arrow in figure 1) considers the response variability with or without pre-printing.

Earlier LHU data (the 1997 and 1998 data matched on building unit level) had indicated problems of high response variability *between* two LHU surveys for the overlapping reference months. Furthermore, there was a within questionnaire tendency that data for the two overlapping reference months was similar to the data of the latest two reference months, (e.g. a high frequency of building units reported no changes in rental price between January and January.) Since one of the objectives of the LHU is to measure levels of change this was unacceptable. We suspected that, at the time of the data collection, it is easier for the respondent to give accurate data for the latest two reference months and that the two overlapping (earliest) reference months are too far back in time. (Almost 1.5 years when the data collection is made.)

Because of this and a more general response burden argument, we considered pre-printing data from the LHU 1999 survey in the LHU 2000 questionnaires. With the LHU 1999 values printed nearby, the respondents were asked to fill in the corresponding reference months in the LHU 2000.[3] (Note that the values for December-99 and Januari-00 of the LHU 2000 survey is needed in the estimation of the Fisher index, and that old values not are available for the units not responding or not included in the LHU 1999 survey.)

[3] Among the respondents who got pre-printed questionnaires there were surprisingly few who gave blank answers in the LHU 2000. These are handled as non-response in this study. However, in the editing work of the survey most of them turned out as verifiers of the pre-printed values, and some respondents thought they did not have to fill in new values since they agreed with the pre-printed values.

The effects we hoped to achieve with pre-printing was; (I) to decrease the response burden by simplifying the questionnaire and (II) to decrease the response variability of the overlapping reference months and thereby decrease the measurement errors and increase the between questionnaire consistency of the units taking part in two consecutive LHU surveys. However, since the pre-printing of previously reported values also may affect the data quality of the latest reference months we designed the experiment to also study pre-printing effects on; (III) the size of and frequency of outlier values before editing, (IV) the number of pre-printing recurrences and changed values between reference months and (V) estimates of yearly changes. To investigate these aspects we used the following experimental plan.

3.1. Experimental Plan and Data Collection

The plan to obtain necessary data was as follows: From the LHU 1999, 1590 building units were identified as common with the sample drawn for the LHU 2000 data collection. 1115 of these 1590 units had reported both rental price and rental area and for at least one of the categories shops, offices or industrial premises. The number of building unit owners (often but not always identical to the respondent) is less than this since several building units have the same owner. For policy reasons it was inappropriate to let one and the same unit owner get both pre-printed and not pre-printed questionnaires. To handle this, and to control possible variation due to owner size, three blocks were constructed. Block A was building units owned by owners with a large number of units and a high total rental income, block B was a middle sized group and block C contained building units owned by owners with only one building unit. In order not to get a too unbalanced design the units inside each block were ordered into three relatively homogenous groups with respect to owner size. Three different treatments were then randomly assigned among the units within each block. The three treatments were, (1) no pre-printing, (2) pre-printing of rental price and rental area for December-99 only and (3) pre-printing of rental price and rental area for both December-99 and

Januari-00. In questionnaires with treatments (2) and (3), data was pre-printed for all types of premises although only shops, offices and industrial premises were of interest in this study. For sake of space this paper concentrates on the results for office premises. Table 1 shows the number of building units in each treatment cell for offices, were n is the number in the design and n^* is the number that was available for the analysis at the data collection deadline.

Table 1: Number of building units with office premises by block and treatment. n = Number of units in the experimental design plan (chosen from LHU-1999.) n^* = Number of responding units matching between LHU 1999 and LHU 2000.

	Treatment								
	(1) No pre-printing			(2) Dec-99 pre-printed			(3) Dec-99 jan-00 pre-printed		
	n^*	n	%	n^*	n	%	n^*	n	%
Block A	31	48	65	41	52	79	49	54	91
Block B	43	52	83	33	39	85	28	36	78
Block C	126	170	74	142	194	73	146	187	78
Total	**200**	**270**	**74**	**216**	**285**	**76**	**223**	**277**	**81**

As seen in table 1, the number of available units varies between 65 % and 91 % of the units in the design plan. The main reason for missing data is unit non-response but changes in type premises (from offices in LHU 99 to some other type of premise in LHU 2000) are another common reason. (We did not observe any clear results that pre-printing affected the unit non-response.)

3.2. Study Variables and Statistical Models

The main variables in the LHU are total monthly rental income and rented area in square metres. For every type of premises we use the ratio of total monthly income and total rented area as our basic variable, i.e. let Z_{tq_k} be the total monthly income per square meter for building unit k; t is a reference month indicator, where $t=1$ if the reference month is Januari-00 and $t=0$ if the reference month is December-99; q is an indicator of the questionnaire (the LHU survey from

which data origins) $q = 1$ if data come from LHU 2000 and $q = 0$ when data come from the LHU 1999 survey.

To measure discrepancies between repeated measurements and response variability (see arrow C1 in figure 1) we compute the ratio $Y_{t_k} = Z_{t1_k} / Z_{t0_k}$ and the logarithm $Y^*_{t_k} = \ln Y_{t_k}$, for every unit *k* and the two overlapping reference months. By means of an ANOVA model (given in equation 1) we can investigate if pre-printing has any effect on response variability.

$$Y^*_{ijk} = \mu + B_i + T_j + BT_{ij} + \varepsilon_{k(i,j)} \qquad (1)$$

where,

Y^*_{ijk} = Deviation in rental income/m^2 between LHU 2000 and LHU 1999 for the k:th building unit in the i:th block and the j:th treatment. (log-scale)

B_i = Effect of the i:th block. $(i = A, B, C)$ (fixed effect)

T_j = Effect of the j:th treatment $(j = 1, 2, 3)$ (fixed effect)

BT_{ij} = Interaction effect of the j:th treatment and the i:th block.

$\varepsilon_{k(i,j)}$ = Random effect in the (i,j):th cell

The randomised block model in equation 1 is a basic model that was used for all types of premises and for both the December-99 and January-00 reference months separately. Note that treatment 2 is a pre-printing treatment when the response variable refers to December-99, but treatment 2 is not a pre-printing treatment when the response variable refers to Januari-00. However, in the latter case one might guess that the pre-printed values of December-99 is helpful for the respondent when he fills in the values of January-00.

To study pre-printing effects on the properties of outlier values, pre-printing recurrences and estimates of yearly change we used traditional t-test and chi-2 tests.

4. Statistical Analysis and Results

4.1. Data Considerations

To fulfil the objectives of our study we chose to do all analysis on unedited data. A drawback then, is that some model assumptions in our analysis may become less valid. Preliminary screenings and model diagnostics revealed the presence of outliers and non-normality. The severity of these violations of assumptions (with respect to changing the general test conclusions) was checked by also performing alternative non-parametric analysis. Since the conclusions from non-parametric and parametric analysis coincided, we believe that the results are fairly robust against these violations.

4.2. Pre-printing effects on response variability

Table 2 show the ANOVA results (based on sum of squares of type III in SAS 8.1) when we fit model (1) to the data for December 1999 and January 2000 respectively.

Table 2: ANOVA results for office premises: December -99 and January-00.

Source	Degrees of freedom	Reference month	
		December-99	January-00
		p-values	
Model	8	0.02*	0.07
Block	2	0.55	0.74
Treatment	2	0.03*	0.01*
Interaction	4	0.06	0.15
Error	630		

* Significant on $\alpha = .05$ level

As seen, we did not benefit from any block effect but there is a treatment effect for both December and January.[4] The comparison of treatment means showed that the biggest differences exist between treatment 1 and treatment 3, (p-values of mean difference tests were 0.016 for December and 0.009

[4] The non-significant block effect means that our blocking did not fulfil its purpose of reducing systematic variation due to owner size. For the analysis however, the effect is only that we lose degrees of freedom in our model test and that the precision of our treatment means is unimproved. In sections 4.3 and 4.4, the blocking is disregarded in the presentation.

for January.) For December data there was also a large difference between the means for treatment 1 and treatment 2 (p-value 0.069). For both December and January the means of the pre-printing treatments were lower than the means of treatment 1.

Without pre-printing the data quality is in fact very poor. The (geometric) mean of treatment 1 was significantly >1 both for December and January, and 95% confidence limits were (1.02;1.09) for December and (1.02;1.11) for January. This indicates that the responses on average were 2%-11% higher in LHU 2000 than in LHU 1999. Obviously, the response variation between repeated measurements is very high without pre-printing, and this is unacceptable in a survey whose aim is to monitor changes of a few percent. The response variability was lower for treatments 2 and 3. All confidence intervals of the treatment means covered 1.0 (although the mean of treatment 2 were suspiciously high for January data.)

4.3. Pre-printing effects on the presence and size of outliers

Smaller response variability is one indicator that pre-printing decreases the measurement errors. Another indicator is the presence and properties of outliers. If we compare data from our treatments, we note that pre-printing give us both fewer and smaller extreme values. Treatment by treatment, table 3 shows the minimum deviation, the maximum deviation and the proportion of units where the response in LHU 2000 deviates by more than 25 % from the response in LHU 1999.

Table 3: Max. and Min. values for $100(Y_{(t=0)_k} - 1)$, and the proportion of building units with $\left|100(Y_{(t=0)_k} - 1)\right| > 25$, by treatment and type of premises. (December -99 values)

		Treatment		
Type of Premises		**1**	**2**	**3**
Shops	Minimum	-87	-29	-22
	Maximum	497	77	292
	Proportion of units with a deviating value > 25%.	14.3	4.2	2.0
Industries	Minimum	-92	-67	-45
	Maximum	1016	31	223
	Proportion of units with a deviating value > 25%.	10.6	4.8	9.4
Offices	Minimum	-48	-75	-71
	Maximum	783	119	275
	Proportion of units with a deviating value > 25%.	14.0	5.5	6.7

The difference between pre-printing and not pre-printing is clear in table 3. (Note that it contains December values i.e. treatment 2 is a pre-printing treatment.) For all types of premises, treatment 1 always has the largest absolute extreme values. The worst was observed for an industry unit, with a 1016 % higher value in LHU 2000 compared to the value in LHU 1999. From a cognitive viewpoint the presence of outliers should perhaps be interpreted a bit different for the treatments. When an outlier value turns up for the pre-printed treatments, it is put there although the respondent is aware of the previously reported value. It can therefore be a correction of the previous value. For the not pre-printing treatment the respondents can be unaware that the value he reports deviates markedly from reported value of the previous year.

The proportion of units with highly deviating values is also higher for treatment 1, (over 10 % of the units for all the premises.) A similar table as table 3 for January data had the same pattern.

4.4. Pre-printing recurrences and effects on other reference months.

One concern using pre-printing was a high frequency of recurrences and that changes would be underestimated. In this application it is actually the other way around. The pre-printing treatments yield more changed values from one reference period to the next. In the LHU 2000, we compared the responses of rental income for the reference months covering a 12-month period. (Refer to arrow C2 in figure 1.) Table 4 illustrates the number of units that responded lower, the same or higher values of rental income in December-00 compared to December-99.

Table 4: Number units with office premises reporting different values of rental income in December-00 compared to December-99, by treatment and direction of change. (%)

Direction of change	Treatment			Total number of units
	1	2	3	
Decrease lower value reported for Dec-00	38 (19.2)	62 (28.7)	35 (15.8)	198
Unchanged Same value reported	60 (30.3)	38 (17.6)	51 (23.1)	132
Increase higher value reported for Dec-00	100 (50.5)	116 (53.7)	135 (61.1)	368
Total	198 (100)	216 (100)	221 (100)	635

The difference between the treatments is statistically significant (Chi-2 statistic is 18.5 with four d.f.) Treatment 1 has the largest percentage units reporting the same value for December-99 and December-00. This does not support the concern that pre-printing would mean more unchanged values. On the contrary there are more reporting of the same values without pre-printing. However, we guess that the large number of 'same-value reporting' without pre-printing is due to respondent's practical problems of giving correct information for the oldest reference period.

The comparison above just counts the number of units giving different values for two different reference periods. It does not tell us whether pre-printing would affect our estimates of yearly change. Therefore, with the null hypothesis of no differences between treatments, we tested the differences in the mean yearly rate of change with the following statistic:

$$\hat{D} = \left(\frac{1}{n_j^*} \sum_{k=1}^{n_j^*} \frac{Z_{t'_{1jk}}}{Z_{t_{1jk}}} - \frac{1}{n_{j'}^*} \sum_{k=1}^{n_{j'}^*} \frac{Z_{t'_{1j'k}}}{Z_{t_{1j'k}}} \right) \qquad (2)$$

Where t and t' are either reference months December-99 and December-00 or reference months January-00 and January-01, and j and j' are pre-printed and not pre-printed treatments respectively. Hence, positive values of $\hat{D}$ indicate a higher mean rate of change for pre-printing treatments.

None of the tests were statistically significant. However, the sign of $\hat{D}$ was positive for every pair wise comparison. This indicates that pre-printing would lead to higher estimates of yearly change than no pre-printing would. Together with results from analysis such as the one in table 4, this observation point to that the risk of underestimation of change because of pre-printing is low in this survey. At least compared to the alternative of not using pre-printing.

In practise, values from the foregoing survey are not used in the estimation in the LHU. The survey needs values for the overlapping months for all units (not only the units whose questionnaires can be pre-printed.) Nevertheless, from a pre-printing perspective comparing the treatment means of yearly rate of change across questionnaires can make another interesting comparison. (See arrow C3 in figure 1)

If we in the denominators of $\hat{D}$ use the values from the LHU 1999 questionnaire, i.e. exchange $Z_{t_{1jk}}$ and $Z_{t_{1j'k}}$ by $Z_{t_{0jk}}$ and $Z_{t_{0j'k}}$ in equation (2). Then the signs of the $\hat{D}$ statistic become negative.

Hence, if we measure the first reference month with values taken from the previous survey, it seems as if pre-printing has a slight dampening effect on the mean of the yearly rate of change.

In summary: When we compute the yearly rate of change within the LHU 2000 questionnaire (C2), the pre-printing treatments give higher values on average, but when we compute across questionnaires (C3, using the LHU 1999 values for the overlapping months) then the pre-printing treatments give smaller estimates. These patterns were the same across all types of premises and regardless whether we study December to December changes or January to January changes. A possible explanation to this observation is that pre-printing affects both the overlapping (pre-printed) reference months and the latest reference months. The effects may be in the same direction but with different strength. Without pre-printing, the measurement of the overlapping months tend to be biased towards the values of the latest reference months (which probably are easier for the respondents to give.) With pre-printing, this bias is smaller since the respondents are helped by pre-printed values from the previous survey. (In previous year's survey those values were probably easier to give (see data collection time in figure 1.)) However, the rate of change comparisons using LHU 1999 data suggest that pre-printing also may affect the measurements of the latest reference months towards the pre-printed values. Whether the latter is good or bad is hard to tell since we do not have the true values and it depends on the quality of the pre-print. Careful evaluation studies have to be performed to compare the magnitudes of these two possible effects.

4.5 Further comments and results

There were no resources for doing a systematic follow up among the respondents. Therefore the only knowledge we have of the respondent's reaction to pre-printing comes from communication during the data collection and editing process. In that communication there were only positive

opinions to the pre-printed questionnaires, some respondents mentioned their satisfaction with the possibility to correct their own previously reported pre-printed values.

Treatment 2 (pre-printing December-99 only) was included in the experiment for two reasons: Firstly, since December and January represent two different calendar years and changes in rents and contracts might often occur between those two months. Secondly, by pre-printing only one of the overlapping reference months it was possible to make additional comparisons of the pre-printing properties.

In our study, the major differences between treatments were between treatment 1 and treatment 3. Treatment 2 behaved as treatment 3 for December data. For January data, when it is a non pre-printing treatment it performed better than Treatment 1, i.e. with less response variability and fewer and smaller outlier values. There were indications that the respondents to some extent used the pre-printed December values to fill in the January values. However, they did not just copy the December values on to January. The correlation of rental income/m^2 for December-99 and January-00 was 0.81 for treatment 2, whereas the corresponding correlation for treatment 1 and treatment 3 were higher. However, since there were no clear evidence that treatment 2 was better than treatment 3, and since treatment 3 was more practical to implement. The LHU now pre-prints both December and January values for the units where it is possible.

5. Discussion

The conclusions from the experiment were in favour of pre-printing. There are clear indications that pre-printing improves the data quality. The response variation of the overlapping reference month is significantly decreased with pre-printing. There are indications of smaller measurement errors and less editing work will be required with pre-printing. (In a sense, pre-printing moves some editing

work from the data collector to the respondents and they are in many cases more suitable to do it.) None of the possible drawbacks were discovered. The same value for two different months was actually more common without pre-printing. The fear that pre-printing led to underestimation of change was not fulfilled, and last but not least pre-printing help the respondent to fill in the questionnaire. Remember that the overlapping reference months are almost 1.5 years back at the time of the data collection. If that data is hard to get, seeing the previously reported values (collected and reported closer to the reference month) should be helpful.

However, the measurement situation in the LHU is not the common situation for pre-printing practises. The LHU includes both a repeated and a new measurement (the overlapping months and the December-00 and Januari-01 reference months respectively). Usually when pre-printing is applied it is only the new measurement that is interesting. Nevertheless, if a repeated measurement is important, as in the LHU, and if a previous measurement was done at a more favourable point of time, it seems natural to use the first measurement to assure good quality in the next. Therefore it is not surprising that pre-printing work well for repeated measurements. Our results concerning pre-printing effects on the quality of the new measurements are more vague. We did not find any statistically significant effects other than it was more common that same values were reported *without* pre-printing. We also noted a tendency with lower rates of change when we compared against data from LHU 1999 (C3) and higher rates of change when we made comparisons within the LHU 2000 (C2). None of these results can with certainty be interpreted as positive or negative for pre-printing, which of course also is interesting since it does not rule out pre-printing as a bad practise.

In Sweden it is more common to pre-print categorical variables. The study variables in the LHU experiment are continuous and the discovered positive properties of pre-printing may be related to that. With continuous variables it easier to practically study response variation and analyse effects.

For a hardly ever-changing categorical variable this can be more difficult. Often there is a methodological problem in analysis of pre-printed data. Without pre-printing we do not want too many changed values over time, since that is an indication that many of the changes are spurious. With pre-printing, on the other hand, we do not want too many unchanged values, since that may be a sign of respondents choosing a simple way to respond or not noticing that the values should be changed.

The study in the LHU survey was done as a part of a bigger review project at Statistics Sweden. The study is rare since it is seldom possible to do large-scale experiments within an ongoing (traditional) survey. Although pre-printing is frequently used in data collections, there are few methodological studies of its properties. Nevertheless, within the project we also performed an experiment on categorical variables. This was done in cooperation with The Employers Association of Property Owners and made on their annual wage survey, which uses an electronic questionnaire (see Holmberg (2002).) That experiment was designed to study the effects of ending the pre-printing for some categorical variables. It was thought that the pre-printed variables was not updated frequently enough and thereby led to a lot of recalculations and editing, since they were connected to other wage related variables. We made an analysis of the change and update frequencies, which showed that pre-printed questionnaires were updated just as often as those without pre-printing. Hence, that study did not support the idea that pre-printing conserves outdated values in a larger extent.

The Swedish experience is that the response burden argument often prevails when pre-printing is decided upon. These experimental studies show that a data quality argument for pre-printing also can be strong. Especially in cases where there are problems with a high response variation such as in the LHU. Moreover, in our review we did not find any cases where the possible negative properties of pre-printing had any significant effect on macro level statistics. If it is feared that pre-printing will influence the data in such a manner that estimates get biased, we still suggest that it is

worth looking at. Studies indicate that unwanted, spurious response variability is reduced with pre-printing. With carefully designed experiments it is possible to estimate and monitor the eventual bias due to pre-printing. With such an estimate we can then correct the statistics and at the same time keep the advantages with increased measurement accuracy. Not to mention the advantages of a reduced response burden, especially in business surveys.

6. Summary

In this paper we have discussed the possible advantages and disadvantages with pre-printing historical data on questionnaires. These matters were studied in a large-scale experiment, and in that study, treatments with pre-printed questionnaires outperformed treatments without pre-printing. Besides a decrease in response burden, pre-printing led to better data quality in the sense of less spurious response variability and fewer and smaller outliers. Moreover, the study revealed no indications of the possible drawbacks, such as tendencies to underestimate change or frequent recurrences of the pre-printed values. Pre-printing is clearly recommended for the survey where we did the experiment. We also discussed Swedish experiences of pre-printing. In general these are positive. In a recent review no really bad examples of pre-printing were found. Another experiment with categorical variables in an electronic questionnaire also indicated advantages with pre-printing. We suggest that pre-printing is considered when there are problems with high response variability and measurement errors. With a proper testing of its effects and a proper redesign of questionnaires, we believe that pre-printing is a useful method to improve data quality, and a method that has advantages for the respondents as well as in the practical editing work.

References

Corti, L. and Campanelli, P. (1992) The Utility of Feeding Forward Earlier Wave Data for Panel Surveys, *Survey and Statistical Computing* 109-118.

Hansen, M. H. Hurwitz, W. N., and Pritzker, L. (1964) The Estimation and Interpretation of Gross Differences and the Simple Response Variance. In C.R. Rao ed. Contributions to Statistics Presented to P.C. Mahalanobis on the Occassion of his 70:th birthday, Pergamon press Ltd, Calcutta., 111-136.

Holmberg, A., (2002) Rapport om förtryckningens fördelar och nackdelar: En sammanställning av erfarenheter på SCB samt resultat från två experimentella studier. Statistiska Centralbyrån. (in Swedish).

Krosnick, J. (1991) Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* **5** 213-236.

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.

Montgomery, D. C., (1984) *Design and Analysis of Experiments* 2 edition Wiley, New York.

O'Muircheartaigh, C., A., (1986) Correlates of Reinterviewer Response Inconsistency in the Current Population Survey, *Proceedings of the second Annual Research Conference,* U.S. Bureau of the Census march 23-26, Reston, Virginia.

Pafford, B. V., (1986) 'Studies of response errors in NASS surveys: The effect of using previous survey data. *Proceedings of the Section on Survey Research Methods. American Statistical Association.* 574-579

Pafford, B. V., (1988) The Influence of Using Previous Data in the 1986 April ISO Grain Stock Survey. Washington D.C.: National Agricultural Statistical Service.

Särndal, C-E., Swensson, B., Wretman, J., (1991) *Model Assisted Survey Sampling*, Springer, New York.

Stanley, J. and Safer, M., (1997) Last time you had 78, How Many Do You Have Now? The effect of providing Reports on Current Reports of Cattle Inventories. *Proceedings of the Section on Survey Research Methods. American Statistical Association,* 875-879.

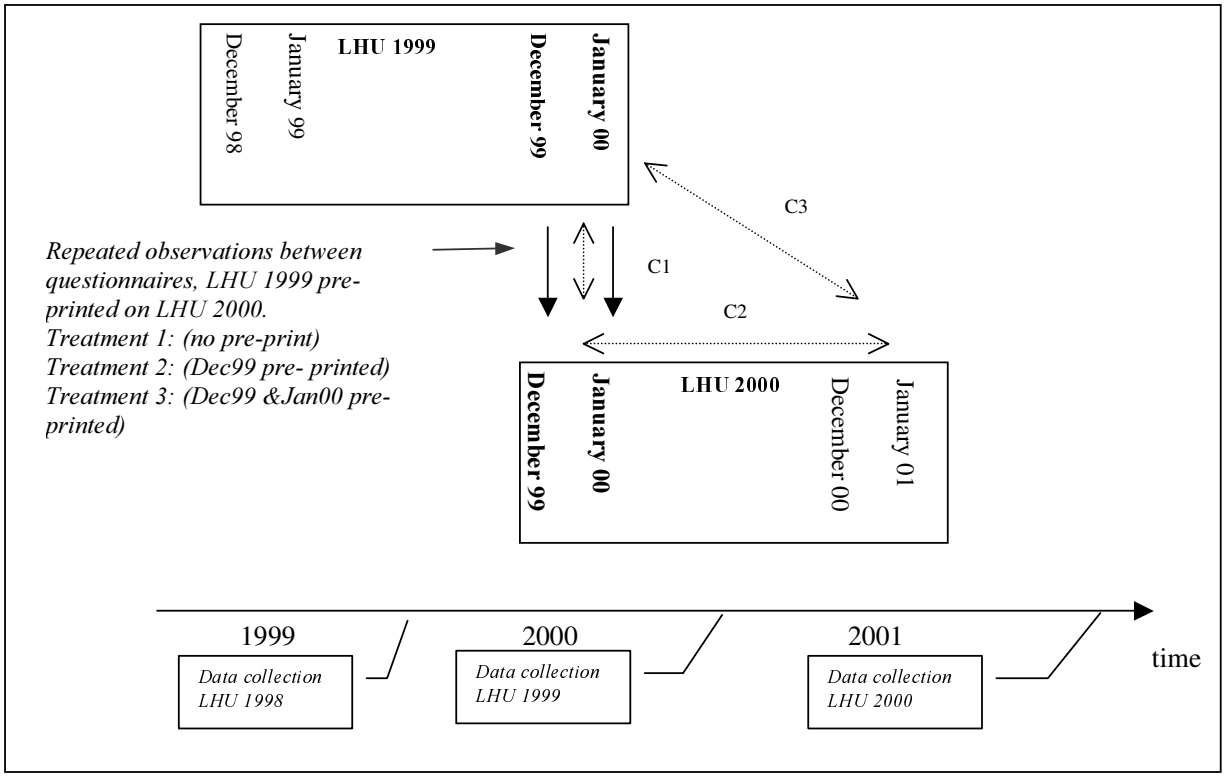

Figure 1 Two successive LHU surveys