Item Nonresponse Analysis for a Mixed-Mode Survey

Boris Lorenc

The literature is equivocal on the existence of a systematic difference in item completion rates between the mail (postal) mode and web modes, two self-administered data collection modes. This exploratory study addresses this issue in the context of a concurrent mixed-mode contact strategy with self-selection of the data collection mode. In order to balance the differences in demographic properties of the respondents, who were free to choose the mode for their participation, a propensity score weighting was used; and in order to take into account various properties of the items comprising the questionnaire, a logistic regression was applied. The results show a small but significant difference in the base item completion rate in favour of the web mode, not attributable to any of the investigated item properties. Additionally, several item level factors were found to differentially influence item completion rates in the two modes. Some of the results are directly translatable into improvements in questionnaire design while others await further theoretical development in order to understand their causes better.

Key words: Mail surveys; web surveys; propensity scores; logistic regression.

1. Introduction

With an ambition to achieve an increase in data quality while reducing costs, Statistics Sweden started offering – as of approximately the year 2000 – the web mode for collecting data parallel to the mail mode in an increasing number of social surveys. A potential for data quality improvement arises from, for instance, better routing through the questionnaire, basic data checking performed already during data collection, elimination of the scanning step and so on, while cost reduction can be achieved by preparing, sending out and post-processing a significantly fewer number of mail questionnaires. However, adoption of the web mode by persons sampled into these surveys has remained low in spite of a dominant proportion of the Swedish population having access to the Internet. Since the use of the mixed-mode web and mail data collection started, these surveys have usually had a proportion of responses by web in the range of 10–15% of all the responses. This was deemed too low, leaving the majority of the collected data without the benefit of an improvement in data quality, while increasing the costs by setting up a very little used web option.

Acknowledgments: The approach taken in this paper is that of the author, not necessarily endorsed by Statistics Sweden. The author is grateful to Johan Hallqvist and Cecilia Magnusson of the Stockholm Public Health Cohort project for the opportunity to analyse the project’s 2007 data. The author acknowledges valuable discussions with Kristoffer Olsson and Andreas Persson in the course of working on this study, and a reliance on the expertise of Kristina Halldin, Tania Hayden, and Michael Nilsson at Statistics Sweden in performing the study.
In 2007 a large-scale experiment was conducted at Statistics Sweden regarding contact strategies, that is, alternative ways of “mixing” mail (postal) and web modes in a mixed-mode data collection setup. The primary aim was to explore the ways to increase proportions of web responses in the standard contact strategy of Statistics Sweden for concurrent mixed-mode mail and web surveys. The experiment identified alternative contact strategies which succeeded in increasing the proportion of web responses to between 45% and 65% of all the responses while achieving a cost reduction of up to 20% (Holmberg, Lorenc, and Werner 2009).

However, prior to Statistics Sweden committing itself to a new standard contact strategy, it can be of relevance to study other properties of the data collected using different modes and the mixes of the modes. In particular, little has hitherto been known about item nonresponse in the two modes, mail and web. The goal of the present study was to provide such knowledge – in particular, to ascertain whether there are differences between the two modes with respect to item nonresponse. There is in the literature not a simple and straightforward answer to this question.

It is, in general, well documented that interviewer-assisted data collection modes result in higher item completion rates, in addition to higher unit response rates, than self-administered modes (cf. the excellent review paper by de Leeuw, Hox, and Huisman 2003). Different mechanisms have been offered as explanations of this difference: cognitive – for instance, probing by the interviewer leading to a higher likelihood of recall; social – explicit compliance with the interview request and consistency of behaviour within a survey, and situational – for instance, multitasking in a self-administered mode (Tourangeau, Rips, and Rasinski 2000; Groves, Cialdini, and Couper 1992; Heerwegh and Loosveldt 2009).

Further, it has been demonstrated that item nonresponse is related to demographic and other properties of the respondent. For instance, in the context of a diary survey (Tucker 1992), higher item completion rate was associated with both higher age and higher educational level; and in the context of a health survey (Elliott et al. 2005), age, race/ethnicity and some substantive health variables and medical treatment variables were predictive of item nonresponse.

With respect to the goal of the present study, comparisons between different self-administered modes (e.g., mail and web) are of special interest. Results here seem, however, to be inconclusive. On the one hand, for instance Kwak and Radler (2002) find, in a student population, a significantly higher item completion rate in the web mode than in the mail mode, both overall and for close-ended questions only. On the other hand, again in a student population, Kaplowitz, Hadlock and Levine (2004) find no significant differences in item completion rates between the mail mode and the web mode. De Leeuw (2008), in her review of computer-assisted data collection methods, finds a pattern of inconsistent results regarding the difference between the two modes.

Thus, it is of relevance for Statistics Sweden as a statistics producer to determine what levels of item nonresponse rates in the two modes in its “standard” mixed-mode contact strategy do exist and to understand their causes.

Several methodological issues need to be taken into account before drawing any inference about the relationship between item nonresponse and mode of survey participation. First, the kind of survey model used. Commonly, the distinction is made...
between concurrent and sequential mixed-mode models (Cobben, Schouten, and Bethlehem 2006). As the standard contact strategy of Statistics Sweden for mixed-mode mail and web surveys is a concurrent mixed-mode model, will this paper be limited to that kind of model.

Second, within concurrent models a distinction is made between those where assignment to a specific mode is under the control of the researcher and those where respondents are themselves selecting between the modes offered for participation (i.e., self-selection). In general, a way of treating differences in demographic and other auxiliary variables that are predictive of item nonresponse is to include them in a regression model together with an indicator of mode, and then analyse the significance of the mode difference by the statistical significance of the corresponding regression coefficient (Kwak and Radler 2002; Heerwegh and Loosveldt 2008). However, when there is self-selection of the response mode, the approach is invalid if there is association between the selection mechanism and the study variable. In such a case, it is necessary to make assumptions about the mechanism behind the self-selection. One approach consists of building complete models of the kind met within econometrics (Heckman 1979), another involves relying on an assumption known as ignorability of mode assignment (Rosenbaum and Rubin 1983), that is, that all variables causally relevant for mode choice and for occurrence of item nonresponse have been observed and that mode choice and item nonresponse are conditionally independent given these other variables, in which case propensity score weighting or matching may be used.

Finally, items in a questionnaire commonly vary in their properties (e.g., close-ended vs open-ended questions, short or long questions, simple or complex questions), and these properties can be associated with different rates of item nonresponse. Particularly relevant for the present study are situations where an interaction occurs between item properties and the mode of response, contributing to a mode-specific item nonresponse rate on a particular item.

The structure of this paper is as follows. Section 2 gives a brief summary of the results of the 2007 experiment and establishes the rationale for the present study. Section 3 presents the questionnaire of the survey on which the 2007 experiment was conducted and discusses aspects of item nonresponse in it. Section 4 treats methodology for the item nonresponse rate analysis, presenting a way of both taking care of respondents’ differences on auxiliary variables and including item properties in the analysis. The results themselves, indicating a small but significant difference in item nonresponse rate in favour of the web mode, are presented in Section 5. A summary and discussion in Section 6 close the paper.

2. The 2007 Experiment

A characteristic of the standard mixed-mode mail and web strategy as conducted by Statistics Sweden is that sampled persons have the two modes to choose from throughout the field period of the survey. In that sense, the standard strategy’s concurrent mixed-mode model differs from that in for instance Cobben et al. (2006), where probabilities of allocation to mode, \( \eta(.) \) in their model, are under control of the surveyor.

By altering the time of introduction of one or the other of the two modes, Holmberg et al. (2009) succeeded in increasing the proportion of responses by web to between 45% and
65% of all the responses. The alternative contact strategies led to a slight decrease in the overall response rate – a reduction of between 2% and 4% with respect to the standard strategy’s overall response rate of 76%. The authors note that an important consequence of the higher proportion of web responses is the potential for cost reduction by up to 20% of the total data collection costs.

However, other data quality properties associated with the two modes, in addition to cost, can be of relevance when considering whether a statistics producer is to switch to a new “standard” contact strategy for mixed-mode mail and web surveys. One of them is item nonresponse, to which the present study is dedicated.

3. Survey Topic and Study Variables

The survey within which the 2007 experiment was embedded was the 2007 follow-up of the 2002 Stockholm County Council Public Health Survey (HS02, for short). The target population of the 2002 survey was persons living in the County of Stockholm aged 18 to 84 years. The sample size was 50,000, of which about 31,200 persons (62%) responded. The 2007 follow-up (HS07; more about that survey can be found in Magnusson 2008) had as its sample all respondents in HS02. With the aim to boost the response rate, it was decided that data collection would be a mixed-mode mail and web strategy, in contrast to HS02 which was a mail only survey. For subject matter reasons, partially different sets of questions were given to those up to the age of 64 years inclusive and those 65 years and above. The current analysis was carried out on the data collected from the former age group.

That part of the HS07 questionnaire that provided data for the analysis in the present study consisted nominally of 50 items. In fact, as several of the items were divided into subitems, and sometimes even these divided into their own subitems, there were actually 182 variables collected with the part of the questionnaire that was used in the current study. With reasonable grouping of variables, it was possible to reduce this number to 133 items, representing the totality of items considered in the present study. Of these, 104 items were on the “main path” of the survey (i.e., these items ought to have been answered by all survey respondents) while the other 29 were branched items to be answered conditional on specific preceding responses, not applicable otherwise.

Item nonresponse, the study variable for the present analysis, was determined as follows. A variable indicating item nonresponse was introduced for each of the 133 items. For each item on the main path, for each respondent, the indicator was set to 1, indicating item nonresponse, unconditionally whenever there was no response whatsoever on that item by that respondent. For items off the main path, the indicator was set to “not applicable” (thus, that data point was not included in further analysis) if a response on the item did not meet the condition for the item to be applicable.

Item 53, requesting an open-ended designation of own current occupation, was not used as it carried specific quality concerns.

2 The final items 56–62 of the HS07 questionnaire were intended for only those respondents that were employed in the year preceding the survey. However, there was no explicit question collecting this specific piece of information. Consequently, in that part of the questionnaire it was not possible to distinguish between missing response due to item not being applicable and proper item nonresponse. For that reason, the final part of the questionnaire had to be left out of the analysis. Items 1–4 (e.g., height and weight) were not used at all as they would increase the disclosure risk; age was used only in a categorised form. Finally, Item 53, requesting an open-ended designation of own current occupation, was not used as it carried specific quality concerns.
item which determined whether the branch would be entered or not was such that the branch was not entered. Otherwise, if there was no response on an item on the branch, the item’s item nonresponse indicator was set for that respondent.

The variables of interest were indicators, for each item $i, i = \{1, 2, \ldots, N\}$ (as stated before, $N = 133$), of item nonresponse, $I_i^{(m)}$, for both modes $m = \{1-mail, 2-web\}$ and all respondents $k, k = \{1, 2, \ldots, K^{(m)}\}$, where $K^{(1)} = 9,342$ and $K^{(2)} = 1,602$. Note that these variables were observed for all respondents and all items, except when “not applicable” for items on branches.

In a study comparing modes, it is essential to identify and, if possible, eliminate any confounding factors. Specifically, the web version of the questionnaire could have had – but, with two exceptions, did not have – controls forcing a certain format or value range. The exceptions were: (i) open-ended questions asking for numeric values did not accept other values than numeric, and (ii) questions of the type “choose one [response alternative]” did not accept more than one response alternative to be selected. The mail version naturally did not implement any of these constraints.

4. Method

The goal of the present study was to determine the level of item nonresponse in each of the two modes, mail and web, and investigate circumstances contributing to it in each of the modes. Specifically of interest was to determine whether one of the modes generated a rate of item nonresponse consistently higher than the other.

All contact strategies in the 2007 experiment (Holmberg et al. 2009) – the standard and the alternatives – implied each its own different mixed-mode response model. In order to have, at the beginning of an investigation regarding item nonresponse, as simple a situation to deal with as possible, only one of the contact strategies was chosen for the analysis. It was the standard strategy, as that sample size was by far the largest, about 14,500 sampled persons, compared to 2,000 sampled persons per each of the alternatives.

It is known that different demographic and other auxiliary variables are related to mode choice in studies conducted under concurrent mixed-mode models of the kind implemented in the current standard strategy. For instance, the variables sex, age, income, education level and country of birth have all been found to be correlated with mode choice in a similar data collection setup (Lorenc 2008). As these auxiliary variables could also be correlated with the propensity to generate item nonresponse, it was important to control for this.

One way to proceed could have been to build full-fledged models involving mode choice, auxiliary information and dependent variables, of the kind met in econometrics (Heckman 1979). Strong dependence on the assumed model and difficulties of estimation of its parameters call for caution with respect to this approach (Wainer 1986). With the recent advances in Bayesian methodology and tools, such approaches are becoming more widespread (e.g., Munkin and Trivedi 2003). However, the assumption of conditional independence between a “treatment” (nonresponse, mode selected for participating in a survey, etc) and the study variable, given the auxiliary information, is standardly assumed in a number of current approaches and works in survey methodology – for instance in propensity score weighting for coverage and self-selection issues (Duncan and Stasny
2001; Lee 2006) and in calibration for nonresponse (Särndal and Lundström 2005).
The assumption of conditional independence and the associated propensity score theory
(Rosenbaum and Rubin 1983, 1984) were used in the present study, in order to correct
for the effect of self-selection.

For each of the two modes, \( m = \{1 = \text{mail}, 2 = \text{web}\} \), the interest was in the
mode’s tendency to generate item nonresponse, \( \theta^{(m)} \). Ultimately, the interest was in
the difference \( d \),

\[
d = \theta^{(1)} - \theta^{(2)}
\]

between the two modes.

Define \( \pi^{(m)}_g \) to be the tendency associated with item group \( g \) to generate item
nonresponse in mode \( m = \{1, 2\} \). This magnitude was estimated by the average observed
item nonresponse rate in the mode \( m \) for item group \( g \),

\[
\hat{\pi}^{(m)}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} \frac{1}{n_{r,im}} \sum_{k \in r^{(m)}} I^{(m)}_{ik}
\]

where \( n_g \) is the number of items that belong to the group, \( r^{(m)} \) denotes the response set in
mode \( m \), \( n_{r,im} \) is the size of that response set in the mode, and \( I^{(m)}_{ik} \) is the nonresponse
indicator for respondent \( k \) on item \( i \). Note that a respondent is responding through only one
of the modes \( m \).

As different aspects of an item can affect its item completion rate, it was necessary to
build an explanatory model that distinguished between the main effect (base item
nonresponse rate) and additional explanatory factors. For this, a logistic regression model
with indicator regressors was used:

\[
\log \left( \frac{\pi^{(m)}_g}{1 - \pi^{(m)}_g} \right) = \theta^{(m)} + [\mathbf{B}^{(m)}]^{\prime} \mathbf{x}^{(m)} + \mathbf{e}^{(m)}
\]

that is, the item group \( g \)’s tendency to generate item nonresponse in mode \( m \), \( \pi^{(m)}_g \) was
modelled using an intercept \( \theta^{(m)} \), a mode specific column vector of regression coefficients
\( \mathbf{B}^{(m)} \) (excluding an intercept term), and an item group specific, as well as mode specific,
column vector of indicators \( \mathbf{x}^{(m)} \). For the error term \( \mathbf{e}^{(m)} \), \( E(\mathbf{e}^{(m)}) = 0 \).

Now, the above would have held if there were no self-assignment issues and
the associated differences in auxiliary variables (demographic and other properties of the
respondents) between the two mode groups. In order to accommodate also this aspect
in the estimation, a propensity score approach was used.

The propensity score is, in general, the conditional probability of occurrence of one of
the outcomes a binary variable \( b \), given the set of auxiliary variables, \( \mathbf{z} \),

\[
e(\mathbf{z}) = \Pr (b|\mathbf{z})
\]

Used in observational studies, the propensity score can – with the appropriate
conditions fulfilled – give unbiased estimates even if groups receiving the two outcomes
differ on relevant background properties (Rosenbaum and Rubin 1984).

While in the present study there were three outcomes \( \{1 = \text{mail}, 2 = \text{web}, 3 = \text{unit}
nonresponse}\), the data regarding unit nonrespondents were not used in the analysis.
In order to estimate the propensity scores, a logit model,
\[
\log \left( \frac{\Pr(m = 1|z_k)}{1 - \Pr(m = 1|z_k)} \right) = \alpha'z_k + \nu_k
\]
was fitted, where \( \alpha \) was a vector of regression coefficients and \( z_k \) was a vector of available auxiliary variables thought to be able to account for the choices regarding participation and mode, \( k \in r \) (that is, respondents in both modes). For the error term \( \nu_k \), \( E(\nu_k) = 0 \). The model was fit using the \textit{glm} function in the \textit{stats} package of R (R Development Core Team 2009), from which were obtained the associated probabilities for each respondent, that is, the estimated propensity scores. Choosing mail respondents as the reference population, the empirical distribution of \( \hat{e}_1^{(1)}(z_k) \) – the estimated propensity score for respondent \( k \) of being a mail respondent (the “1” in the index) among the mail respondents (the parenthesised “1” in the exponent) – was divided into \( H = 4 \) intervals with approximately equal mass by identifying the quantiles corresponding to 0.25, 0.5 and 0.75 of the area under that distribution.

Denote the quantiles by respectively \( q(h)_1^{(1)} \), \( q(2)_1^{(1)} \) and \( q(3)_1^{(1)} \) and – for simplicity of reference – add the denotations \( q(0)_1^{(1)} = \min(\hat{e}_1^{(1)}) \) and \( q(4)_1^{(1)} = \max(\hat{e}_1^{(1)}) \).

Then the propensity score strata \( r_h \), for \( h = \{1, 2, \ldots, H-1\} \), were for all respondents \( k \), defined as
\[
k \in r_h \quad \text{iff} \quad q(h)_1^{(1)} \leq \hat{e}_1^{(1)}(z_k) < q(h+1)_1^{(1)}
\]
and for \( h = H \) as
\[
k \in r_H \quad \text{iff} \quad q(H)_1^{(1)} \leq \hat{e}_1^{(1)}(z_k)
\]

Now, the model fitting in (3) and the associated calculation of average observed item group nonresponse rates in (2) were in fact performed within each of the propensity score strata, resulting in \( H = 4 \) estimates of the intercept of interest per mode, \( \hat{\theta}_h^{(m)} \), one per stratum. The final estimates were then obtained using
\[
\hat{\theta}_ps^{(m)} = \frac{1}{H} \sum_{h=1}^{H} \hat{\theta}_h^{(m)}
\]
Likewise for the other regression coefficients, those in the vector \( \hat{\beta}^{(m)} \).

Producing the estimates in (7) for both modes, it was possible to estimate the magnitude of primary interest, presented in (1), namely
\[
\hat{d} = \hat{\theta}_ps^{(1)} - \hat{\theta}_ps^{(2)}
\]
Corresponding differences were calculated for the other regression coefficients in \( \hat{\beta}^{(m)} \), presented in the last row of Table 4 below.

A rigorous theoretical derivation of this approach is not within the scope of the present paper.
5. Results

Of 14,477 sampled persons, 9,342 (65%) participated using mail, 1,602 (11%) using the web, and 3,533 (24%) were unit nonresponse. The item completion rate was generally high, with 62% of the respondents having full item completion, 33% having item nonresponse on less than 10% (but more than 0%) of the applicable items, and only the remaining 5% having an item nonresponse rate larger than 10% of the applicable items (as the last pair of columns in Table 2 indicate).

Upon an exploratory preview of the data it became obvious that participant auxiliary information was associated with participation choice on the one hand (Table 1) and with a proxy for our variable of interest (item nonresponse), a per respondent average item nonresponse rate, on the other (Table 2). Thus, balancing the difference between the two mode groups was necessary, carried out using the propensity score technique. There was a sufficient overlap of the mail and web populations in each of the propensity score strata, as Figure 1 illustrates. The propensity score reweighting succeeded in balancing the distributions of auxiliary variables between the modes, shown in Table 3.

The exploratory preview of the data showed further that in both modes there existed factors that were probably accountable for some of the item nonresponse in that mode. For instance, Item 36 in HS07 had 12 subitems. The way that the web data collection instrument of Statistics Sweden works, with lower screen resolutions some of the content of the item may initially be presented below the lower screen edge, with only the scroll band along the right-hand side edge of the screen indicating that there are more items below (Figure 2). Thus, item nonresponse on items below the lower screen edge would probably be caused by a different and additional mechanism than that on items initially displayed above the lower screen edge.

In all, six factors were investigated:

- **Below** - whether the item or subitem is initially presented below the lower screen edge in the 1,024 × 768 pixels resolution (Figure 2), or not. This factor should only be relevant in the web mode.
- **Branch** - whether the item is on a branch (Figure 3), or not (“not” implying that the item is on the main path of the questionnaire); an item on a branch is to be completed conditional on responding in a specific way on a previous item. From a respondent’s perspective, this factor should only be relevant in the mail mode as in the web mode respondents are presented with just those items that are applicable to them, given the already submitted responses.
- **BrRoot** - whether the item is the root of a branch (Figure 3), or not. From a respondent’s perspective, this factor should only be relevant in the mail mode.
- **Embed** - whether the item has a response field embedded in the text of a response alternative (Figure 4), or not.
- **Indent** - whether the item is indented (Figure 3), or not. Due to the way the web questionnaire is constructed, this factor was less expressed in the web version (e.g., item 22_y in the right pane of Figure 3) than in the mail version.
- **TypeIn** -whether the response requires either writing or keying in a response (Figure 3), or not.
Table 1. Distribution (in percentages) of the mode of participation (mail response mode, web response mode or unit nonresponse) among categories of available demographic and socio-economic variables. All the tables contain significant differences. (The smallest group size in the calculations is 1,533)

<table>
<thead>
<tr>
<th>Particip.</th>
<th>Gender</th>
<th>Age</th>
<th>Partnership</th>
<th>Income</th>
<th>Education</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>18–29</td>
<td>30–44</td>
<td>45–64</td>
<td>Y</td>
</tr>
<tr>
<td>Mail</td>
<td>60.2</td>
<td>72.5</td>
<td>51.9</td>
<td>62.0</td>
<td>74.1</td>
<td>70.4</td>
</tr>
<tr>
<td>Web</td>
<td>15.1</td>
<td>8.6</td>
<td>15.1</td>
<td>14.0</td>
<td>8.9</td>
<td>10.6</td>
</tr>
<tr>
<td>Unit NR</td>
<td>24.7</td>
<td>18.9</td>
<td>33.0</td>
<td>24.0</td>
<td>17.1</td>
<td>19.0</td>
</tr>
</tbody>
</table>
Table 2. Distribution (in percentages) of respondents’ categorised average response rates (full item completion, item completion \(> .9\), item completion \(< .9\)) among categories of available demographic and socio-economic variables (except Partnership, which did not contain significant differences). Added is the relation between mode and average response rate. All displayed tables contain significant differences. (The smallest group size in the calculations is 1,103)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Income</th>
<th>Education</th>
<th>Origin</th>
<th>Mode</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>18–29</td>
<td>30–44</td>
<td>45–64</td>
</tr>
<tr>
<td>Resp. r.</td>
<td>63.8</td>
<td>58.5</td>
<td>67.0</td>
<td>67.6</td>
<td>54.9</td>
</tr>
<tr>
<td>(.9,1.0)</td>
<td>31.7</td>
<td>36.6</td>
<td>30.0</td>
<td>29.5</td>
<td>38.8</td>
</tr>
<tr>
<td>(.0,9)</td>
<td>4.5</td>
<td>4.9</td>
<td>3.0</td>
<td>2.9</td>
<td>6.3</td>
</tr>
</tbody>
</table>
Each of these factors was represented as an indicator variable, coding with “1” an occurrence of the phenomenon in question, and with “0” its absence.

In the first round of the analysis, a separate model was fitted for each mode group and each propensity score stratum. This most often gave models with an intercept and two or three regressors – for the mail, generally Branch, Embed and Indent were included, and for the web Below and TypeIn; BrRoot, which was included in only two of the selected models and with weak coefficients, was left out of the further analysis. In the second round, however, all the other factors were included in the models for both modes, in order to enable comparison between the modes. This generally happened to result in little change to estimates of the coefficients originally in the models.

The results of this analysis are presented in Table 4. Regression coefficient estimates in the propensity score strata are those obtained using the \texttt{glm} function in the \texttt{stats} package of \texttt{R} (R Development Core Team 2009), then carefully transformed into probabilities using $\frac{\exp(\beta(m))}{1 + \exp(\beta(m))}$. As the associated standard errors (SE) cannot be interpreted per se, confidence intervals are reported instead, calculated as $\hat{\beta} \pm 1.96SE\hat{\beta}$ and then the end points transformed into probabilities as above. Standard errors of point estimates on Summary and Difference rows are simple averages of the end points in the corresponding propensity score strata confidence intervals. $D^*$ is a measure of model fit (Agresti 2002),

$$D^* = \frac{G^2(0) - G^2(M)}{G^2(0)}$$

where $G^2(0)$ is deviance of the null model (only intercept included) and $G^2(M)$ deviance of the model under consideration.

The table indicates that there was a mode specific base item nonresponse rate, expressed in the intercepts, in the mail mode of 1.5% and in the web mode of 1.1%. Thus the difference between the two modes in base item nonresponse rate was 0.5%. (As all reported values are rounded to one decimal, the reported summands occasionally appear not to add up correctly).
Table 3. Distribution in Table 1 after reweighting using the propensity scores. (Unit nonresponse is excluded as it is not used in further analysis)

<table>
<thead>
<tr>
<th>Particip.</th>
<th>Gender</th>
<th>Age</th>
<th>Partnership</th>
<th>Income</th>
<th>Education</th>
<th>Origin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male</td>
<td>Female</td>
<td>18–29</td>
<td>30–44</td>
<td>45–64</td>
<td>Y</td>
</tr>
<tr>
<td>Mail</td>
<td>86.5</td>
<td>87.9</td>
<td>87.5</td>
<td>83.1</td>
<td>86.6</td>
<td>85.7</td>
</tr>
<tr>
<td>Web</td>
<td>13.5</td>
<td>12.1</td>
<td>12.5</td>
<td>16.9</td>
<td>13.4</td>
<td>14.3</td>
</tr>
</tbody>
</table>
An item initially presented, in the web mode, below the lower screen edge added about 4% to the base probability of item nonresponse in that mode and had a small, practically immaterial effect in the mail mode. Among the web respondents, the effect monotonically increases with the propensity score stratum, thus probably being associated with some of the auxiliary variables.

An item on a branch added about 3% to the probability of item nonresponse in the mail mode; this factor had practically no effect in the web mode, as expected.

Embedding the response field within the response alternative of a previous question resulted in both modes in an effect of adding about 5% to the probability of item nonresponse.

Fig. 2. Initial display of Item 36 in the web questionnaire, on a screen with the resolution 1,024 × 768 pixels. Subitems h–l of the item are below the lower screen edge

Fig. 3. Item 21 is the root of a branch (BrRoot), Items 22, 22_y, 23, 24 and 25 are on a branch (Branch), Item 22_y is an indented, unmarked question (Indent), and Items 22_y, 23, 24 and 25 require writing or typing in a response (TypeIn). A section of the mail mode questionnaire is displayed in the left pane, and of the web mode questionnaire in the right pane.
Further, inclusion in the questionnaire of an indented item, not explicitly marked as a question, added about 2% to the probability of item nonresponse in the mail mode and was practically without effect in the web mode.

Finally, an item requiring a written or keyed in response (rather than a ticked or a clicked one) added 0.5% to the probability of item nonresponse in the mail mode and 0.8% to the probability of item nonresponse in the web mode.

Data collection staff for HS07 were interviewed to find out whether there were any technical difficulties that could account for the base difference, but none could be identified.

6. Validation by a Propensity Score Analysis Only

A way of avoiding application of the propensity score technique to regression coefficients would be to limit the analysis to only those items for which there were no other factors potentially accounting for the observed item nonresponse than the mode itself. In other words, limit the analysis to those items in the questionnaire that were coded “0” on all five item characteristics factors found in the study to be of relevance (Below, Branch, Embed, Indent, and TypeIn). An analysis that discards some data and does not allow inference about magnitudes of effects of several factors at the same time is in general less desirable than one that includes them, but it may serve the purpose of verifying the basic result of that more complex analysis.

There were 91 items, of totally 133, that were not indicated on any of the five factors that had an effect in the preceding analysis. For each of these items, a propensity score estimate of item $i$’s mean item nonresponse rate in mode $m$,

$$\tilde{\mu}_{i, m} = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{n_{r,h,k}} \sum_{j \in R_{h,k}} \mu_{i, h}^{(m)}$$

was calculated, where $r_{j}^{(m)}$ is the set of all participants responding in mode $m$ in the propensity score stratum $h$, $n_{r,h,k}$ is the size of that set and $\mu_{i, h}^{(m)}$ is an item’s mean item nonresponse rate for that mode and stratum.

The distribution of these 91 estimates, separately calculated for the two modes, is given in Figure 5. The reference population is mail respondents.
<table>
<thead>
<tr>
<th>Mode</th>
<th>Str.</th>
<th>D*</th>
<th>Intcpt (CI)</th>
<th>Below (CI)</th>
<th>Branch (CI)</th>
<th>Embed (CI)</th>
<th>Indent (CI)</th>
<th>Typeln (CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mail</td>
<td>1</td>
<td>0.997</td>
<td>0.8 (0.8, 0.8)</td>
<td>-0.1 (-0.2, -0.1)</td>
<td>2.2 (1.9, 2.5)</td>
<td>3.2 (2.5, 4.1)</td>
<td>2.2 (1.7, 2.8)</td>
<td>0.1 (0.0, 0.2)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.987</td>
<td>1.2 (1.1, 1.2)</td>
<td>-0.3 (-0.3, -0.2)</td>
<td>3.0 (2.6, 3.4)</td>
<td>2.8 (2.2, 3.6)</td>
<td>2.0 (1.6, 2.6)</td>
<td>0.2 (0.1, 0.4)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.985</td>
<td>1.5 (1.5, 1.6)</td>
<td>-0.4 (-0.5, -0.3)</td>
<td>3.2 (2.8, 3.6)</td>
<td>5.2 (4.2, 6.4)</td>
<td>2.2 (1.7, 2.9)</td>
<td>0.5 (0.3, 0.8)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.977</td>
<td>2.6 (2.5, 2.7)</td>
<td>-0.4 (-0.5, -0.3)</td>
<td>4.0 (3.6, 4.5)</td>
<td>8.3 (7.0, 9.7)</td>
<td>1.9 (1.3, 2.5)</td>
<td>1.3 (1.0, 1.7)</td>
</tr>
<tr>
<td>Summary</td>
<td></td>
<td>1.5</td>
<td>(1.5, 1.6)</td>
<td>-0.3 (-0.4, -0.2)</td>
<td>3.1 (2.7, 3.5)</td>
<td>4.9 (4.0, 6.0)</td>
<td>2.1 (1.6, 2.7)</td>
<td>0.5 (0.3, 0.8)</td>
</tr>
<tr>
<td>Web</td>
<td>1</td>
<td>0.996</td>
<td>0.4 (0.4, 0.5)</td>
<td>2.2 (1.8, 2.8)</td>
<td>0.0 (-1.3)</td>
<td>3.5 (1.6, 7.0)</td>
<td>-0.0 (-3.9)</td>
<td>0.3 (-0.1, 0.0)</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.997</td>
<td>0.7 (0.6, 0.8)</td>
<td>3.4 (2.6, 4.4)</td>
<td>0.0 (-2.4)</td>
<td>3.1 (1.2, 7.1)</td>
<td>-0.3 (-5.1, 0.7)</td>
<td>0.7 (0.2, 0.0)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.996</td>
<td>1.8 (1.6, 2.0)</td>
<td>5.5 (4.4, 6.9)</td>
<td>-0.2 (-6.4)</td>
<td>9.0 (4.3, 16.7)</td>
<td>-1.0 (-14.1, 1.2)</td>
<td>1.5 (1.4, 0.0)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.982</td>
<td>1.3 (1.1, 1.5)</td>
<td>5.9 (4.4, 7.9)</td>
<td>1.0 (3.2)</td>
<td>4.2 (1.4, 10.2)</td>
<td>-0.1 (-7.2, 1.1)</td>
<td>0.6 (-2.2, 2.2)</td>
</tr>
<tr>
<td>Summary</td>
<td></td>
<td>1.1</td>
<td>(0.9, 1.2)</td>
<td>4.3 (3.3, 5.5)</td>
<td>0.2 (-2.8)</td>
<td>4.9 (2.1, 10.3)</td>
<td>0.3 (-7.1, 1.3)</td>
<td>0.8 (-0.2, 0.3)</td>
</tr>
<tr>
<td>Difference</td>
<td></td>
<td>0.5</td>
<td>(0.3, 0.7)</td>
<td>-4.6 (-5.9, -3.5)</td>
<td>2.9 (1.9, 3.7)</td>
<td>-0.1 (-6.3, 3.8)</td>
<td>2.4 (2.3, 4.5)</td>
<td>-0.2 (-2.0, 0.8)</td>
</tr>
</tbody>
</table>
In the mail mode, the mean item nonresponse rate was 1.6%, the median was 1.5% and the first and the third quartile of the distribution of the 91 estimated item nonresponse rates were 1.2% and 1.9% respectively. In the web mode, the mean item nonresponse rate was 0.9%, the median was 0.7%, and the first and the third quartile of the distribution of the 91 estimated item nonresponse rates were 0.5% and 1.2% respectively.

Also the difference per item of the two estimates in (9) was calculated,

\[ \hat{d}_{i, ps} = \hat{\mu}_{i, ps}^{(1)} - \hat{\mu}_{i, ps}^{(2)} \]  

the expression corresponding to (8) but here done pairwise over the 91 items \( i \) in the analysis (estimated difference in item nonresponse rates between the two modes on matched pairs of “same” items). The resulting distribution is displayed in Figure 6. The mean of this distribution of the differences was 0.7%, the median was 0.6% and the first and the third quartile of the distribution were 0.4% and 1.0% respectively.
Thus this estimate of the difference and the one from the propensity score weighted regression analysis (0.5%) agreed reasonably well.

7. Summary and Discussion

This study set off to investigate whether there was a difference in item completion rates between mail and web data collection modes in the context of a concurrent mixed-mode contact strategy with self-selection of the mode, as applied by Statistics Sweden. In a health survey, on a population fairly well resembling the general population, a small but significant difference of about 0.5% in favour of the web mode was found. Additionally, different factors have been identified as reducing item completion rates in one or both of the modes.

Embedding the response field in the body of a response alternative was in both modes associated with a reduction of item completion rates by about five percentage points. Possibly, an unexpected placement of the response field made it less visible than if it had been placed in its standard position, on the left edge of the item box/window. If so, the use of another placement than the standard would in the practice of questionnaire design need to be supported by arguments that suffice to balance the loss of information due to the item nonresponse. The same conclusion applies to indented response fields, associated with a reduction of item completion rates by about two percentage points (only in the mail mode).

In the web mode, response alternatives shown below the lower screen edge were associated with a reduction of item completion rates of about four percentage points. Advising designers of a web questionnaire to avoid long questions/response alternatives would here sound far too simple. The fact is, what is long depends on the screen resolution of respondents’ displays; and in fact, rather than the tendency being that screen resolutions are increasing, the spread of mobile units carries the consequence of the reduction of screen resolutions. Possibly, new interaction designs for questionnaire completion will need to be developed, that are as efficient for large as for small screen resolutions.

In the mail mode, items on a branch were associated with a reduction of item completion rates of about three percentage points. It does not seem simple to suggest a quick improvement here either. Transition to the web mode seems, however, to resolve this issue, there not being any noticeable increases in overall response rates on the branches.

The cause of the base difference in item completion rates between the two modes could not be found in this study. The effect was observed generally across items, which apparently excludes for instance scanning difficulties pertaining to some specific items in the mail mode as an explanation. On the respondent side, there might possibly have existed variables that have not been included in the study and that influenced in part the selection of the response mode and in part the likelihood of leaving an item without response. In that case, the observed effect could be the bias due to self-selection, uncorrected because some relevant variables were not observed. There are, though, insufficient grounds to speculate on whether and which other variables could have needed to be included in the propensity score weighting procedure, whether assumptions underlying the procedure have been fulfilled or not, and whether another modelling approach could have needed to be used instead. Or, on the other hand, whether the difference is genuine, in which case it is awaiting an elucidation.
The purpose of this study was to compare the two modes beyond the raw unit nonresponse rate measure. It has focused on item completion rates as one important aspect of data quality. While the studies of e.g., Kwak and Radler (2002) and Heerwegh and Loosveldt (2008) had more than just this aspect of data quality, the present study has added an analysis of the potential of item properties to account for item nonresponse.

Parameters to be estimated were in this study near the edge of the parameter space, possibly introducing difficulties in the parameter estimation relying on likelihood maximisation. Further, a hierarchical model could have been in place, where respondents would make up one layer, with nested item group nonresponse rates. However, such models estimated with likelihood maximisation could be even less stable, particularly so with a small number of observations on the lower level, as was the case in this study. An MCMC solution, yet another option, was not completed by the time of finalising this report.

While this study reported some estimates of the differences, it should be seen as an exploratory rather than a confirmatory analysis. The models for mode choice and for individual item nonresponse rates were built using available variables or variables identified in the course of the study rather than variables that a preparatory work established as well accounting for the modelled phenomena. For instance there was likely a difference in cognitive complexity of the items in the survey, a factor that would need to be included in a future model of item nonresponse. However, replication of this work in other contexts would indicate the degree in which the effects identified here are able to carry over into a wider applicability.

8. References


