Title: Estimating nonresponse bias and mode effects in a mixed-mode survey.

To be published in International Journal of Market Research (in press).

Peter Lugtig
Utrecht University
Gerty J.L.M. Lensvelt-Mulders
University of Humanistics
Remco Frerichs and Assyn Greven
TeamVier BV

In mixed-mode surveys, it is difficult to separate sample selection differences from mode-effects that can occur when respondents respond in different interview settings. This paper provides a framework for separating mode-effects from selection effects by matching very similar respondents from different survey modes using propensity score matching. The answer patterns of the matched respondents are subsequently compared. We show that matching can explain differences in nonresponse and coverage in two Internet-samples. When we repeat this procedure for a telephone and Internet-sample however, differences persist between the samples after matching. This indicates the occurrence of mode-effects in telephone and Internet surveys. Mode-effects can be problematic; hence we conclude with a discussion of designs that can be used to explicitly study mode-effects. (120 words)

Keywords: propensity score matching, Internet survey, telephone survey, mode-effects, nonresponse bias.

4353 words

About the authors (include at end)
Peter Lugtig is a researcher at the department of methods and statistics at Utrecht University, the Netherlands, where he specializes in survey methods. Gerty Lensvelt-Mulders is full professor at the University of Humanistics in Utrecht, the Netherlands. Her research focuses on meta-analysis, mixed-mode surveys as well as the role of research in society. Remco Frerichs and Assyn Greven are director and principal investigator at TeamVier research B.V. They specialize in mixed-mode studies and public opinion research.

Correspondence to: Utrecht University, methods and statistics, P.O. Box 80.140, 3508 TC, the Netherlands
E-mail: p.lugtig@uu.nl

Introduction

It is becoming more difficult and costly to conduct surveys among the general population (Groves 2005). This is mainly because of the fact that response rates have been slowly decreasing over the past decades (de Leeuw & de Heer 2002). Although this does not necessarily mean that coverage and nonresponse bias have been increasing as well (see Groves & Peytcheva 2008), survey researchers are nowadays trying to tailor survey designs to limit survey costs, keep up response rates and limit nonresponse bias. One of the ways in which surveys are tailored is by implementing mixed-mode survey designs. This paper discusses how to study one of the possible downsides of mixed-mode surveys: the mode effect. A mode effect occurs when respondents give different answers solely because of the method of interviewing. Studying mode-effects is difficult, because they are easily confounded with selection effects that occur when conducting surveys with multiple modes. This paper proposes Propensity Score Matching (PSM) as a method to disentangle mode effects from sample composition differences and shows how mode effects occur in a mixed Internet-telephone study.

Mode effects in mixed-mode surveys

In mixed-mode surveys, two or more methods of survey data collection are combined. The most prominent modes in current survey research are face-to-face, telephone, paper and the Internet (de Leeuw 2005). These modes can be combined in different stages of the survey process: to contact people, in the initial response phase, and also in following up on respondents.

While mixed-mode surveys intend to reduce potential coverage and nonresponse bias, this advantage may be offset by the occurrence of a mode-effect. A mode-effect occurs when respondents answer differently to a survey question, solely because of the mode in which the question is being administered. Mode effects might stem from differences in question administration: whether an interviewer is present, the media in which questions are administered and the way in which information is transmitted (de Leeuw 2005). These differences have led survey designers to worry about three related types of mode effects.

In situations where there is an interviewer, some people adjust their answers to what they expect the interviewer wants to hear. This social-desirability effect increases with the sensitivity of the question (Kreuter, Presser, & Tourangeau 2008). This leads to generally more positive answers when respondents evaluate a question on a negative-positive dimension, as will be the case in this study.

The second type of mode-effect can occur because of a difference in auditive versus visual transmission of data. In telephone surveys, interviewers read out the survey-questions along with all possible answer categories. The respondent listens and typically awaits the interviewer's instructions before answering. Those answer categories that are read out last, are more likely to be memorized and chosen (recency effect). In contrast to this, respondents in mail or Internet surveys read the questions and answer categories themselves. They read top-down or left-right and pick the first answer category that is thought to be appropriate (primacy effect) (Dillman & Christian 2005)

Finally, another mode-effect can occur with the choice for a "don't know" response category in telephone and Internet surveys. In telephone surveys, this option is generally not offered to respondents, but can be registered by the interviewer when respondents have trouble answering a question. In an Internet-survey the "don't know" option however is either explicitly offered or not offered, leading to differences in the frequency of "don't know" answers in a mixed-mode survey (Dillman & Christian 2005).

Although worries about mode effects have been extensively discussed in the survey literature, there is mixed evidence for their existence (for an overview, see de Leeuw 2005). Partially, this may be due to the fact that mode effects depend on the topic and specific structure of the question and response scale (Dillman et al. 2009). It also depends however on the fact that mixed-mode surveys lead to different compositions of the sub-samples. A difference that is found between two samples in a mixed-mode survey might be due to different levels of nonresponse or coverage bias in the different survey modes, but it could also be caused by a mode-effect.

Separating mode–effects from differences in sample composition

There are a number of ways to separate sample composition effects from mode-effects. Every approach has its disadvantages, and it is generally difficult to separate the two effects. The first and most straightforward way to assess nonresponse bias and mode-effects uses an experimental setting, in which a random group of respondents changes survey modes during the interview (Heerwegh 2009). It is essential in such a design that none of the respondents who have to switch, drop out, and so this approach is difficult to use in a study among the general population.

A second approach is the comparison of survey estimates from mixed-mode studies to a 'golden' standard (de Leeuw 2005; Kreuter, Presser & Tourangeau 2008). One of the problems is that we seldom have validation data on attitudinal questions, which is the type of question where survey researchers worry about mode effects.

The third approach relies on statistical modeling. The goal of this approach is to make the two samples from a mixed-mode study equivalent. This can be done by weighting (Lee 2006), or by using a multivariate model that corrects for differences between the samples (Dillman et al. 2009). Finally, Latent Variable models can be used in combination with re-interviewing (Biemer 2001) or validation data on voter turnout (Voogt & Saris 2005) to correct for nonresponse bias. The disadvantage of these modeling approaches is that they assume that every survey mode can potentially cover the entire population. We know however that for example telephone and Internet coverage rates are not universal (Blyth 2008).

This paper takes a different approach and will show how Propensity Score Matching (PSM) can be used to match respondents from two sub-samples in a mixed-mode survey and study mode-effects. The idea of PSM stems from quasi-experimental research, and is used to eliminate differences in sample composition using a set of covariates. In this paper, we use PSM to correct for sample differences in levels of coverage and nonresponse. An illustration of this idea for the Netherlands is shown in figure 1.

- insert figure 1 about here –

As opposed to weighting, PSM does not try to make the two samples equivalent. In fact, one of the main advantages of PSM is the fact that we can identify those respondents who are unique to a specific survey mode, and those who are found in both modes.
Matched respondents from the two survey modes share the same background characteristics. We argue that after matching they then should also be similar on other aspects related to the variables used in the matching process. Substantive differences that we find after matching for the matched respondents should be small, if there are no mode effects. If large differences remain after matching, they are likely due to mode-effects.

In the next section, we explain how we use three samples in this study: we first compare a probability-based Internet-sample to a quota sample drawn from an Internet-panel. We choose to first compare two Internet-samples in order to show how propensity score matching can be used to explain differences due to coverage and nonresponse bias between two samples. As all respondents in these two modes receive the same Internet questionnaire, mode effects cannot exist. We will show that differences between the matched Internet samples disappear after matching. In the second part we return to the primary objective of this paper, and match a telephone sample to the probability-based Internet-sample. We expect a mode effect after matching: the telephone respondents should respond more positive to a set of rating scales than the Internet respondents due to the presence of an Interviewer. Second, we expect the matched samples to differ in the proportion of extreme positive as well as negative answers due to a recency effect in the telephone sample.

Methods

Sample

Our data stem from a mixed-mode survey conducted between April and June 2008 in the province of Zuid-Holland in the Netherlands. In the survey, respondents were asked how they experience environmental pollution from industry, traffic and agriculture.
For the survey two random samples were drawn from the central database containing all postal addresses in the Netherlands. The Computer Assisted Telephone Interview (CATI) sample consists of 6118 households which have a known landline phone number. They received a letter sent by the province government. A week later these households were called and the household member with the next birthday was asked for the telephone interview, with no incentive offered. Five recall attempts were made, if no contact was established. This procedure resulted in 2685 complete CATI-interviews and a response rate (RR1) of 47 per cent (AAPOR 2008).
The Internet-sample was also drawn from the central address-database. Because we lack a sampling frame of e-mail addresses for the general population we used a two step approach. 7090 households were sent a letter, which included an URL and an individualized login code to complete the survey on the Internet. Two weeks later, nonrespondents to the letter were sent a reminder by mail, and again two weeks later, nonresponding households who had a known telephone number, were phoned and asked to participate. Among those who participated in this Web Assisted Personal Interview

(WAPI) hundred gift vouchers each worth fifty Euros were raffled. This resulted in 1347 complete interviews and a response rate (RR1) of nineteen per cent (AAPOR 2008).

In order to investigate nonresponse bias and mode-effects, we drew a third sample in addition to the two probability-based samples. A quota sample stratified on age, gender and employment situation was drawn from the TeamVier access panel. Five hundred respondents took part in the exact same Internet-survey as the WAPI-respondents.

Instruments

The questions of the CATI and WAPI surveys were identical, except for the introduction and end of the questionnaire. Both surveys contained socio-demographic questions, including age, gender, highest level of education (7 point scale), composition of the household and employment status. From the postal code provided by the participants, we coded the degree of urbanization and average income in the street of the respondent on the basis of the registry of Statistics Netherlands (2009).

A set of seven questions asked how respondents experience environmental hindrance; our dependent variables. Respondents had to indicate on a scale from 1 (a lot of hindrance) to 10 (no hindrance at all), how much hindrance they generally perceived. The items asked for hindrance in the form of 1) dust from industry 2) bad smell from industry 3) noise from industry 4) bad smell from traffic 5) noise from traffic 6) noise from airplanes and 7) light pollution. A 'don't know' option was implicitly offered, both in the CATI and Internet-survey, where respondents could skip a question. We will assess mode-effects for all seven variables separately by evaluating the response patterns for all these variables in detail. We will also look at the combined composite score of the seven environmental hindrance questions to see whether mode-effects are consistent across variables, or cancel each other out[1].

Propensity score matching

Originally, propensity score matching models were developed to solve a problem in quasi-experiments. Individuals cannot always be assigned randomly to a treatment or control condition, as a result of which the estimation of treatment effects may be biased (Cook, Shadish, & Wong 2008; Deheji & Wahba 2002). This problem is similar to the situation in a mixed-mode study, where random assignment to one survey mode is in practice not possible because a respondent might not be able to respond in a specific survey mode (Schonlau, van Soest, Kapteyn, & Couper 2009).

The propensity score in our study summarizes the conditional probability to be a respondent in the CATI-sample, the WAPI-sample or the panel-sample. The propensity score indicates the differences between these samples pair wise. This means we compute three propensity scores of which two are of interest: first, the propensity to be a member of either the CATI or WAPI sample, and second the propensity to be a member of the WAPI or panel-sample. We do not compare the CATI-sample to the panel sample. After

[1] A Confirmatory Factor Analysis using Amos 7.0 (2006) yielded factor loadings for the composite score between .52 -.80, Cronbach's α .82. We computed a weighted mean score and use that variable as a composite score.

propensity scores are computed, similar respondents from the WAPI and panel samples are matched based on their propensity scores which summarize their socio-economic background. Similarly, the CATI and WAPI respondents are also matched.

Results

Composition of the samples before matching

As expected, inability to participate and nonresponse in the CATI- and WAPI-samples lead to different coverage and nonresponse biases. Table 1 shows that the composition of the CATI and WAPI-samples differs significantly from the population before matching. The CATI respondents are older, are less often employed and single, are more often female, live more in non-urban areas and have a slightly lower monthly income than the general population. These results are in line with other nonresponse analyses of CATI-surveys (de Leeuw & van der Zouwen 1989). The only variable for which we somewhat surprisingly find no bias is level of education.

- insert Table 1 here -

The WAPI-sample is also biased. There is a significant difference for six of the seven demographic variables we tested. The only estimate that is in line with the population value is the proportion of people who is employed. For five variables (gender, household situation, education, urbanicity and income) the CATI-sample produces a less biased population estimate than the WAPI-sample, while the WAPI-sample is less biased on age and employment situation. For two variables (gender and income), the combined CATI and WAPI surveys would produce a good estimate for the population values, but for the other five variables, substantial biases would remain.

The WAPI and CATI-samples also differ on our dependent variables. Respondents in the CATI-sample consistently score significantly higher on all seven environmental hindrance questions. The differences are large. Table 1 shows that the means for CATI respondents are about 1 full point or 10% higher than the means in the WAPI-sample. There are also differences between the WAPI and panel sample, although these differences are somewhat smaller. The question we now turn to is whether these differences in our dependent variables are caused by differences in sample composition or mode-effects caused by the different interviewing strategies.

Results from propensity score matching

As the propensity score is computed using a set of covariates, the choice of covariates is extremely important. We chose to use a basic set of socio-economic characteristics to compute the propensity score for ach individual: gender, being employed (dummy), age, household composition (single or not), education (1-7 scale), urbanicity (1-5 scale), income and knowledge of an environmental complaints phone number. We also use all possible two- and three-way interactions between these variables in a logistic regression analysis and compute a propensity score for every individual. These covariates produce a Nagelkerke R^2 of 0.16 in a logistic regression with survey mode (CATI-WAPI) as dependent variable. With WAPI-panel as dependent, we find a Nagelkerke R^2 of 0.17[2]. The socio-demographic variables produce a R^2 of 0.31 when the composite score serves as the dependent variable in a regression analysis. All these coefficients indicate that at least part of the nonresponse biases in the different samples can be explained by our covariates. The inclusion of more covariates would possibly increase the probability to explain all differences (Cook et al. 2008). The reasons why we constrain ourselves to this set of covariates are threefold. First, socio-economic variables are routinely used in marketing and social sciences to weight data. Second, socio-economic variables are highly correlated with access to both the Internet and a landline-phone. Finally, attitudinal variables are themselves subject to possible mode-effects, and therefore, we deem them unsuitable as covariates in this analysis.

Propensity score matching is implemented in the statistical programme R 2.9.1 (R Core Development Team 2009) along with the package 'MatchIt' (Ho, Stuart, Imai, & King 2009). Apart from being flexible, open-source and user-friendly, the ‘Matchit’ package offers many different ways to match respondents. We chose to use the technique of Coarsened Exact Matching (CEM) for two reasons. First, with CEM, the balance between the treatment groups is defined ex ante. This prevents the user from adjusting imbalances through repeatedly running the matching procedure with different specifications for average treatment effect estimation error and number of matches. Second, CEM can deal with missing data, by discarding those cases from the matching procedure (Iacus, King, & Porro 2009)[3]. As a result, about 5 per cent of all respondents were not included in the matching procedure[4].

About sixty per cent of the Dutch population has access to both a landline phone and the Internet (Kool, Maris, & Munck 2009). For this reason we chose to match about sixty per cent of the sample members in our smallest sample (WAPI). For comparison reasons we specified about the same number of matches in the panel-sample[5]. Those respondents that were matched were as expected very similar on the covariates, leading to a balance improvement of 99 per cent. In other words, we managed to match about sixty per cent of the respondents in the panel and WAPI-sample to a very similar respondent in the WAPI and CATI-sample.

The WAPI and panel samples after matching

After matching, the WAPI and panel-respondents are according to our expectations very similar. From the WAPI-sample 209 respondents are matched to 162 respondents from

[2] Due to the fact that we do not have up-to-date information on income and urbanicity for the panel members, these values are not shown in Table 1, nor were these variables used in matching the panel respondents to WAPI-respondents.

[3] To make sure our results were robust, we also tried ‘nearest neighbour’, ‘exact’ and ‘genetic’ matching and in each of these methods we used various matching- specifications. In most settings, we arrived at the same results, although some settings did produce different results from the results we present here. We come back to this point in the discussion section.

[4] The R-code used for the matching procedures is available from the authors upon request

[5] Due to a smaller number of cases in the Internet-sample and greater imbalance in the propensity score, 45 per cent of all respondents in the panel sample was matched.

the panel-sample. Most of the differences in the dependent variables that we found before matching disappear for these matched respondents.

Table 2 shows the response patterns of both the matched and unmatched respondents. For the first of the seven questions on environmental hindrance, we see that the significant difference that we found before matching (as shown in table 1) is greatly reduced. Before matching, the mean hindrance score in the WAPI-sample was 6.98 and in the panel-sample 8.19. After matching, the hindrance for the matched WAPI-respondents is 7.58 and 8.07 for the panel-respondents. This difference is no longer significant. For the means of the other environmental hindrance questions, we find that the differences that were there before matching are consistently reduced after matching. The only strong difference that remains is for the question about the bad smell of industry. Two other differences remain marginally significant, while the differences on the other questions, as well as the composite score, disappear after matching (see the appendix for all statistical tests).

- insert Table 2 here -

Apart from the means, we also find the response patterns in the matched WAPI and panel-samples to be similar. There are differences in the proportions of positive responses within the matched samples, but neither the matched WAPI-, nor the panel-respondents are consistently more positive (matched WAPI–matched panel differences range between -7.7 and +9.9 per cent).

In the proportion of extreme positive and negative responses we also find no consistent pattern for the seven dependent variables (differences in extreme positives range between -15.5 and +9.2 per cent and the difference for extreme negatives between -2.7 and +1.7 per cent). The response patterns of the matched WAPI and panel respondents are in conclusion very similar. The only indicator where differences persist after matching is the mean score on hindrance from bad smell from industry. We are not able to explain why a difference persists for this variable. All other indicator show that PSM is able to explain the differences caused by different levels nonresponse and coverage errors.

Apart from the matched respondents, Table 2 also shows the response patterns for the respondents that we were unable to match. In short, we find the unmatched panel-respondents to respond more positively in general, and choose the extreme positive answer category more often than the unmatched WAPI-respondents. As expected, the unmatched panel- and WAPI-samples do differ from each other.

Concluding, we find that propensity score matching successfully explains the differences between the matched WAPI and panel samples. Matching can be successfully used to select those respondents that are found in the two modes, as well as identify those respondents unique to a survey mode.

The CATI and WAPI-samples after matching

Matching the CATI and WAPI-samples proved to be more difficult than matching the two Internet-samples. Before matching, the means of all seven dependent variables as well as the composite score were different in the CATI and WAPI-samples. From Table 3 we see that these differences are only slightly reduced by matching. The means for the

1068 respondents from the CATI-sample who are matched are still consistently higher than the means for the 708 WAPI respondents (see the appendix for all statistical tests). This finding holds for all seven environmental hindrance questions as well as the composite score and indicates a mode-effect: respondents in the CATI-sample give consistently more positive answers than WAPI respondents, who are very similar to them. This is likely because of an interviewer effect.

- insert Table 3 here

Unsurprisingly, the higher mean scores in the CATI-sample are accompanied by other differences in the response patterns. We find that the differences in the proportion of positive answers are consistently higher in the matched CATI-sample (differences between CATI and WAPI proportion of positive answers range between +2.2 and +9.4 per cent). We also find the matched CATI respondents are much more likely than matched WAPI-respondents to choose the most extreme positive answer category (differences between +1.7 and +16.2 per cent). In the WAPI-sample, respondents choose the extremely negative answer category more often than CATI-respondents for six of the seven variables. The differences are however small (between +0.4 and +1.5 per cent).

All in all, we believe our findings indicate two related mode effects: respondents in the CATI-sample are more positive than respondents in the WAPI-sample, even after matching. They also pick the extremely positive answer category more often, but this may partially be explained by the fact that CATI-respondents are more positive in general.

The differences in the response patterns of the matched CATI and WAPI-samples are not caused by a failure to effectively match respondents. Table 3 shows that the differences in the response patterns for the unmatched samples are even more pronounced than the matched samples. The differences in means, the proportions of positive responses and extreme responses are all larger in the unmatched samples than the matched samples. In the next section we discuss the implications of these mode-effects.

Conclusion and discussion

When carefully utilized, mixed-mode surveys can both increase coverage- and nonresponse rates and decrease bias resulting thereof. However, using different survey modes results in a confounding of sample selection effects and mode-effects, and separating these effects from each other is difficult. The starting point of this paper was to show how propensity score matching can help to disentangle mode-effects from sample effects.

Propensity score matching can be used to classify respondents who are unique to a certain mode versus respondents who are present in both modes. When two Internet-samples (panel-WAPI) are compared, the *matched* respondents from both samples are similar not only on their socio-economic characteristics; after matching they also show similar answer patterns on our outcome variables. This leads to the conclusion that propensity matching explains differences caused by sample selection effects. As expected, we find no mode-effects comparing the random WAPI-sample and quota sample from an access panel. The differences in outcomes of the unmatched parts of these Internet-samples are due to differences in the compositions of the unmatched samples.

However, when *matched* respondents of the telephone and Internet-sample are compared (CATI-WAPI), respondents that appear to be similar on their background characteristics, still respond differently. Although the magnitude of the differences declines for the matched samples, the answer patterns of the matched samples show mode-effects. The matched CATI-respondents choose the extremely positive category more often and respond more positively in general than their WAPI-counterparts.

Concluding, we showed that mode-effects and nonresponse effects interact in mixed-mode surveys combining telephone and Internet surveys, making it impossible to straightforwardly merge the data from these surveys and analyze them as one dataset.

A limitation of our study is the way in which we studied mode-effects. The different mode effects that we wanted to distinguish (i.e. recency effects, primacy effects and interviewer effects) interact with each other, making it impossible to evaluate which types of mode-effects occur. Recency effects and social desirability in telephone surveys both lead to higher sample means, and in our study, it is impossible to separate the two.

A second limitation of our study is that propensity score matching is a form of statistical modeling related to regression techniques. As such, it suffers from some of the weaknesses that statistical models in general suffer from. A different specification or the inclusion of different covariates could have resulted in different results. We tried various matching specifications, and as long as we chose not to match all sample respondents, our results were robust. However, more research is needed on propensity score matching and its effectiveness in mixed-mode surveys to learn about the differential effects of matching specifications under different circumstances.

Looking forward, the central question that emerges in mixed-mode survey research is whether we can combine data from mixed-mode surveys. Here we offer two directions for further study. The directions both involve the use of external validation data. Adding substantive questions (e.g. newspaper readership), for which the aggregate population estimate is known, can be used to evaluate the quality of mixed-mode samples before and after matching. Moreover, external validation can give insight in the possible trade-off between nonresponse error and mode effects, and ultimately it is the trade off between errors of non-measurement and measurement that researchers need to understand.

The combination of two mixed-mode samples in presence of mode effects is an issue that still needs to be taken up. Simply combining the two surveys and ignoring mode-effects does not seem the most sophisticated solution. The first and best solution to this problem is to try and prevent mode-effects. Unimode-questionnaires try to make questions cognitively equivalent across modes, reducing the problem of mode-effects (Dillman & Christian 2005).

A second method would be to assess mode effects first, and then decide whether the results from two modes should be presented separately or not. Propensity score matching can disentangle mode-effects from sample differences and shed light on this issue.

References

AAPOR. (2008). *Standard definitions: Final dispositions of case codes and outcome reports for surveys, 5th edition*. Lenexa, Kansas: American Association for Public Opinion Research.

Biemer, P. B. (2001) Nonresponse bias and measurement bias in a comparison of face to face and telephone interviewing. *Journal of Official Statistics,* **17,** 2, pp. 295-320.

Blyth, B. (2008) Mixed mode: The only 'fitness' regime? *International Journal of Market Research,* **50,** 2, pp. 241-266.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008) Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management,* **27,** 4, pp. 724-750.

de Leeuw, E. D. (2005) To mix or not to mix data collection modes in surveys. *Journal of Official Statistics,* **21,** 2, pp.233-255.

de Leeuw, E. D., & de Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 41-54). New York: Wiley,

de Leeuw, E. D., & van der Zouwen, J. (1989). Data quality in telephone and face to face surveys: A comparative meta-analysis. In R. M. Groves, P. B. Biemer, L. E. Lyberg, J. T. Massey, W. L. I. Nichols & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 283-299). New York: Wiley.

Deheji, R. H., & Wahba, S. (2002) Propensity score matching methods for non-experimental causal studies. *The Review of Economics and Statistics,* **84,** 1, pp.151-161.

Dillman, D. A., & Christian, L. M. (2005) Survey mode as a source of instability in responses across surveys. *Field Methods,* **17,** 1, pp.30-52.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., & Messer, B. L. (2009) Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the internet. *Social Science Research,* **38,** 1, pp. 1-18.

Groves, R. M. (2005). *Survey errors and survey costs* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons.

Groves, R. M., & Peytcheva, E. (2008) The impact of nonresponse rates on nonresponse bias - A meta analysis. *Public Opinion Quarterly,* **72,** 2, pp. 167-189.

Heerwegh, D. (2009) Mode differences between face-to-face and web surveys: An experimental investigation of data quality and social desirability effects. *International Journal of Public Opinion Research,* **21,** 1, pp. 111-121.

Ho, D. E., Stuart, E., Imai, K., & King, G. (2009). *Package 'matchit'* , visited on 11-23, 2009, http://cran.r-project.org/web/packages/MatchIt/MatchIt.pdf

Iacus, S. M., King, G., & Porro, G. (2009). *Matching for causal inference without balance checking*, visited on August 30th, 2009, http://gking.harvard.edu/files/abs/cem-abs.shtml

Kool, L., Maris, A., Munck, S. D. (2009). *Marktrapportage elektronische communicatie (market report on electronic communication)*, Netherlands Organisation for the Advancement of Science (TNO).

Kreuter, F., Presser, S., & Tourangeau, R. (2008) Social desirability bias in CATI, IVR and web surveys. *Public Opinion Quarterly,* **72,** 5, pp. 847-865.

Lee, S. (2006) Propensity score adjustment as a weighting scheme for volunteer panel web surveys. *Journal of Official Statistics,* **22,** 2, pp. 329-349.

Schonlau, M., van Soest, A., Kapteyn, A., & Couper, M. (2009) Selection bias in web-surveys and the use of propensity scores. *Sociological Methods & Research,* **37,** 3, pp. 291-318.

Statistics-Netherlands. (2009). *Statline database*. Voorburg: Statistics Netherlands. Visited on July 28th, 2009, http://statline.cbs.nl

R Core Development Team (2009). R: A language and environment for statistical computing, version 2.9.1 Vienna, Austria:www.r-project.org

Voogt, R. J. J., & Saris, W. E. (2005) Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of Official Statistics,* **21,** pp.367-387.

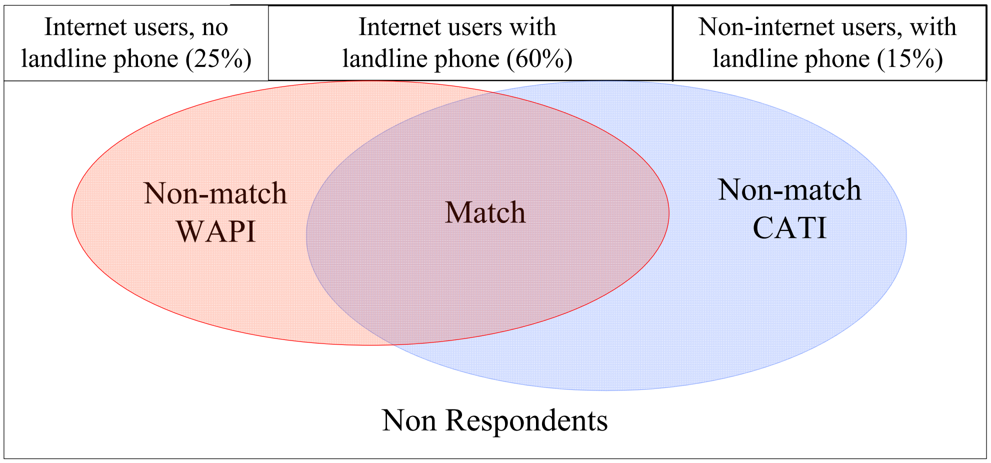
Internet users, no landline phone (25%)
Internet users with landline phone (60%)
Non-internet users, with landline phone (15%)
Non-match WAPI
Match
Non-match CATI
Non Respondents

Figure 1: A mixed-mode survey where respondents from sub-samples are matched.
The sub-samples comprise three strata within the population:
1. Internet users without landline-phone (not covered by CATI),
2. non-Internet users with a landline phone (not covered by WAPI) and
3. those people covered by both Internet and telephone. The strata represent the coverage rates of landline phones and Internet in the Netherlands as of 2009 (Kool et al. 2009).

Table 1: Means and standard deviations for the socio-demographic characteristics of the respondents in the CATI, WAPI and panel-samples and the population

Independent Variables	Means (sd) CATI	Means (sd) WAPI	Means (sd) Panel	Population
Age	**55.1* (16.1)**	***50.1* (14.8)***	*45.9* (14.1)*	47.2
Employed (1=employed)	**.55* (.50)**	**.66 (.48)**	.67 (.47)	.67
Single (1=single)	**.31* (.46)**	**.22* (.42)**	.24* (.42)	.38
Gender (1=female)	**.56*** (.50)	***.45**** (.50)	*.51* (.50)	.51
Education (1-7)	**4.28 (1.73)**	**4.70* (1.54)**	4.85* (1.52)	4.24
Urbanicity (1-5)	2.35* (1.19)	2.44* (1.21)	-	2.22
Monthly net income (600-10000)	**2142*** (673)	**2287*** (687)	-	2200
Worries about society	**2.75** (.68)	**2.49** (.62)	2.53 (.63)	-
Knows environmental complaints agency (1=yes)	**.44** (.50)	***.39*** (.49)	*.56* (.50)	-
Dependent variables (1=a lot of hindrance, 10 – no hindrance at all)				
1) industry dust	**7.97 (2.45)**	***6.98 (2.63)***	*8.19 (2.39)*	-
2) bad smell industry	**8.05 (2.35)**	***6.98 (2.61)***	*8.42 (2.28)*	-
3) noise industry	**8.70 (2.07)**	***7.73 (2.51)***	*8.58 (2.30)*	-
4) traffic bad smell	**7.97 (2.38)**	***7.24 (2.50)***	*7.52 (2.54)*	-
5) traffic noise	**7.46 (2.66)**	***6.56 (2.72)***	*6.94 (2.64)*	-
6) airplanes noise	**8.52 (2.11)**	**7.80 (2.52)**	7.91 (2.40)	-
7) light pollution	**8.87 (2.04**	**8.18 (2.47)**	8.24 (2.49)	-
Composite score 7 items	**8.22 (1.58)**	***7.35 (1.82)***	*7.95 (1.84)*	-

Notes:
*:significant difference from population statistic with *p*=0.05 (one-sample t-test)
- sd: standard deviation
- Statistics in **bold:** significant difference between the CATI and WAPI-samples with *p*=0.05 (independent samples t-test)
- Statistics in *Italics:* significant difference between the WAPI and panel samples with *p*=0.05 (independent samples t-test)
- Population statistics are obtained from Statistics Netherlands (Statistics-Netherlands 2009)

Table 2: differences between WAPI and panel sample after matching

		Mean	Sd.	% Pos.	% Extr. Pos.	% Extr. Neg.	% DK	N
Dust Industry	Match-WAPI	7.58	2.26	79.3	24.0	1.0	.5	208
	Match-panel	8.07	2.47	84.0	39.5	3.7	1.7	162
	Nmatch-WAPI	6.86	2.68	68.5	21.4	3.3	2.6	1082
	Nmatch-panel	8.24	2.36	85.7	44.3	2.7	4.0	300
Bad smell industry	Match-WAPI	7.35	2.41	77.5	22.5	3.3	0.0	209
	Match-panel	8.27	2.34	85.2	43.8	3.1	1.7	162
	Nmatch-WAPI	6.91	2.64	69.2	20.7	4.2	1.1	1099
	Nmatch-panel	8.48	2.28	88.9	48.9	3.6	1.9	307
Noise Industry	Match-WAPI	8.13	2.31	85.2	34.9	2.9	0.0	209
	Match-panel	8.43	2.35	87.8	45.7	3.7	0.6	164
	Nmatch-WAPI	7.65	2.54	79.7	31.8	3.6	1.3	1096
	Nmatch-panel	8.61	2.32	88.6	55.2	2.9	2.2	306
Bad smell traffic	Match-WAPI	7.69	2.34	80.4	29.2	2.9	0.0	209
	Match-panel	7.47	2.39	78.5	20.2	2.5	1.2	163
	Nmatch-WAPI	7.14	2.53	73.5	20.4	3.5	1.1	1099
	Nmatch-panel	7.49	2.64	78.5	28.3	4.6	1.9	307
Noise traffic	Match-WAPI	7.23	2.40	76.0	16.8	3.8	0.5	208
	Match-panel	6.71	2.56	66.1	12.1	3.0	0.0	165
	Nmatch-WAPI	6.45	2.75	64.9	13.0	6.5	0.6	1099
	Nmatch-panel	7.00	2.69	69.9	21.0	3.9	1.2	309
Noise airplanes	Match-WAPI	7.80	2.66	80.4	34.9	5.3	0.0	209
	Match-panel	7.68	2.51	80.0	30.9	3.6	0.0	165
	Nmatch-WAPI	7.79	2.50	81.6	31.5	3.4	0.7	1103
	Nmatch-panel	8.01	2.37	82.6	36.1	2.6	0.9	310
Light pollution	Match-WAPI	8.18	2.57	82.8	42.6	3.8	0.0	209
	Match-panel	8.23	2.33	86.1	43.0	2.4	0.0	165
	Nmatch-WAPI	8.17	2.46	85.1	41.0	3.3	1.1	1099
	Nmatch-panel	8.19	2.61	85.1	45.8	4.2	1.5	308
Composite Score 7 items	Match-WAPI	7.71	1.73	89.5	-	-	-	209
	Match-panel	7.82	1.78	82.7	-	-	-	165
	Nmatch-WAPI	7.27	1.83	88.1	-	-	-	1111
	Nmatch-panel	7.97	1.89	91.4	-	-	-	309

Notes: match-WAPI and match-panel refer to those groups of respondents that could be matched to each other. Nmatch-WAPI and nmatch-panel refer to the groups of respondents that were not matched.

Summary of findings: Nonresponse and coverage bias between WAPI and panel samples are explained

- no differences in means matches panel 4x >WAPI, WAPI 3x > panel
- no recency effect (extreme positives) in matches panel 4x >WAPI, WAPI 3x > panel
- no primacy effect (extreme negatives) in matches panel 3x >WAPI, WAPI 4x > panel
- no acquiescence/social desirability in matches panel 5x >WAPI, WAPI 2x > panel
- no differences in choices “don’t know” too few cases to draw conclusions

Table 3: differences between CATI and WAPI-sample after matching

		Mean	Sd.	% pos.	% Extr. Pos.	% Extr neg.	% DK	N
Dust Industry	Match-CATI	7.88	2.37	81.3	37.6	1.4	.9	1058
	Match-WAPI	7.27	2.43	75.7	22.2	1.9	2.1	688
	Nmatch-CATI	8.03	2.50	82.1	45.4	3.3	1.4	1534
	Nmatch-WAPI	6.63	2.85	64.0	21.7	4.2	2.2	595
Bad smell industry	Match-CATI	7.92	2.34	82.9	38.1	1.5	.6	1062
	Match-WAPI	7.28	2.45	75.9	23.0	2.6	.8	697
	Nmatch-CATI	8.15	2.35	84.8	45.9	1.8	1.1	1539
	NmatchWAPI	6.65	2.75	64.7	18.9	5.8	1.0	603
Noise Industry	Match-CATI	8.63	2.06	85.2	34.9	2.3	.6	1061
	Match-WAPI	7.90	2.38	82.2	33.2	2.7	1.0	696
	Nmatch-CATI	8.74	2.08	90.4	57.9	2.1	.4	1549
	Nmatch-WAPI	7.51	2.67	78.2	31.1	4.7	1.3	601
Bad smell traffic	Match-CATI	7.90	2.35	80.4	29.2	2.1	.6	1062
	Match-WAPI	7.45	2.35	78.2	21.8	2.7	1.0	696
	Nmatch-CATI	8.04	2.39	83.5	42.2	2.2	.5	1548
	Nmatch-WAPI	6.98	2.66	70.7	21.7	4.1	.8	604
Noise traffic	Match-CATI	7.34	2.59	75.6	28.4	3.2	.3	1065
	Match-WAPI	6.71	2.61	68.8	13.0	4.7	.7	698
	Nmatch-CATI	7.56	2.68	78.2	35.2	4.5	.4	1550
	Nmatch-WAPI	6.36	2.85	63.7	14.2	7.9	.5	606
Noise airplanes	Match-CATI	8.35	2.16	87.6	43.6	1.3	.2	1066
	Match-WAPI	7.93	2.33	83.7	31.5	2.5	.6	699
	Nmatch-CATI	8.65	2.06	90.8	52.7	1.5	.3	1551
	Nmatch-WAPI	7.64	2.70	79.0	32.7	4.8	.5	605
Light pollution	Match-CATI	8.78	2.10	90.5	58.1	2.2	.4	1064
	Match-WAPI	8.35	2.29	87.2	41.9	2.9	.8	697
	Nmatch-CATI	8.94	1.98	92.4	62.0	2.1	.3	1552
	Nmatch-WAPI	7.98	2.66	82.1	40.8	3.3	1.0	603
Composite score 7 items	Match-CATI	8.11	1.57	94.9	-	-	-	1068
	Match-WAPI	7.54	1.69	91.2	-	-	-	703
	Nmatch-CATI	8.30	1.59	95.2	-	-	-	1556
	Nmatch-WAPI	7.10	1.95	84.9	-	-	-	609

Notes: match-CATI and match-WAPI refer to those groups of respondents that could be matched to each other. Nmatch-CATI and nmatch-WAPI refer to the groups of respondents that were not matched.

Summary of findings: Differences between CATI and WAPI-samples remain after matching: occurrence of mode-effects.

- differences in means matches CATI 7x >WAPI
- recency effect (extreme positives) in matches CATI 6x >WAPI, WAPI 1x > CATI
- primacy effect (extreme negatives) in matches WAPI 7x >CATI
- acquiescence/social desirability in matches CATI 7x >WAPI
- no differences in choices “don’t know” too few cases to draw conclusions

Appendix:

Table 4 : t-tests for differences between matched and unmatched samples

T-values (df)	Before matching WAPI-panel	Matched samples WAPI-panel	Unmatched samples WAPI-panel
Dust Industry	**-9.19 (917)**	**-1.97 (368)**	**-8.73 (533)**
Bad smell industry	**-11.41 (968)**	**-3.65 (369)**	**-10.30 (558)**
Noise Industry	**-6.75 (928)**	-1.23 (371)	**-6.28 (528)**
Bad smell traffic	**-2.11 (1798)**	0.90 (370)	**-2.10 (1404)**
Noise traffic	**-2.67 (1806)**	**-2.00 (371)**	**-3.16 (1411)**
Noise airplanes	-0.82 (1807)	0.44 (372)	-1.35 (1411)
Light pollution	-0.44 (1801)	-0.21 (372)	-0.10 (1405)
Mean Score 7 items	**-6.23 (1818**	-0.62 (372)	**-5.91 (1422)**
Significant difference	**6/8**	3/8	6/8

- df are rounded to nearest number
- Statistics in **bold**: significant with p<0.05

Table 5 t-tests for differences between matched and unmatched samples

T-values (df)	Before matching CATI-WAPI	Matched samples CATI-WAPI	Unmatched samples CATI-WAPI
Dust Industry	**11.37 (2435)**	**5.25 (1744)**	**10.61(976)**
Bad smell industry	**12.58 (2405)**	**5.49 (1443)**	**11.80 (965)**
Noise Industry	**12.15(2225)**	**6.69 (1331)**	**10.17 (1005)**
Bad smell traffic	**8.86 (2523)**	**3.90 (1486)**	**8.53 (897)**
Noise traffic	**9.87 (2590)**	**4.95 (1485)**	**8.95 (1049)**
Noise airplanes	**9.01 (2271)**	**3.81 (1414)**	**8.28 (889)**
Light pollution	**9.73 (2233)**	**3.98 (1397)**	**8.06 (873)**
Mean Score 7 items	**14.85 (2358)**	**7.06 (1422)**	**13.48 (1942)**
Significant difference	8/8	8/8	8/8

-df are rounded to nearest number

- Statistics in **bold**: significant with $p<0.05$