Research & Methods
ISSN 1234-9224 Vol. 18 (1, 2009): 3–43
Institute of Philosophy and Sociology
Polish Academy of Sciences, Warsaw
www.ifispan.waw.pl
e-mail: publish@ifispan.waw.pl@

Non-Response Bias in Cross-National Surveys: Designs for Detection and Adjustment in the ESS

Jaak Billiet
Centre of Sociological Research (CeSO) of Katholieke Universiteit Leuven

Hideko Matsuo
Centre of Sociological Research (CeSO) of Katholieke Universiteit Leuven

Koen Beullens
Centre of Sociological Research (CeSO) of Katholieke Universiteit Leuven

Vasja Vehovar
Social Sciences, University of Ljublijana

This paper is focused on process and output aspects of the obtained sample and deals with the measurement of non-response, and the study of non-response bias from a viewpoint of comparative research in which the concept of “equivalence” in measurement is central (Jowell et al., 2007). The paper starts with a theoretical reflection on several designs for the detection of non-response bias: comparing sample statistics with population statistics; using information from reluctant respondents based on converted refusals; asking a small set of crucial questions at occasion of first contact (and refusal) or in a period after the main survey, and collecting observed information of the house and neighbourhood of the sampling units. Each of these methods are used in the past three round of ESS, but only the first and second approaches are fully documented for Rounds 1 and 2 till now. Problems related to each of these methods are considered, and the application of each of the procedures is (empirically) evaluated using information of past ESS surveys as far as the data are available. Methods that can be used for data based adjustment of the sample measures are considered.

Key words: Data quality assessment • cross-cultural surveys • measurement error • non response bias

INTRODUCTION

The implementation of strict quality standards and the pursuit of survey quality criteria such as high response rates and low nonresponse bias are not unusual in national surveys but rare in cross-national surveys (O'Shea et al. 2003). High standards and optimal comparability, as well as the evaluation and improvement of response and contact procedures have, from the outset, been an important focus in the *European Social Survey* (ESS). Actually, comparability of obtained response between countries is one of the most serious challenges of comparative and longitudinal cross-nation research (Jowell 1998). Large differences in response rates between country samples may result in nonresponse bias which differs across countries. In view of this challenge, in past rounds of ESS great efforts were made to reduce nonresponse and to obtain strict comparable estimates of the response rates.

The norm for ESS response rates was by the *Preparatory methodological committee* set to 70 percent. The *Central coordination team* (CCT) of ESS applied the definition of AAPOR (2000, Lynn et al. 2002) in order to obtain standardised survey outcome categories and response rate calculations for the different kinds of samples (individual named, household, address). This is a rather severe definition of response rates that considers as non-respondents sample units that are temporarily absent, who are not able to cooperate because of illness, and those who cannot be traced. The way of calculating response rates and the outcome categories are centrally prescribed and provided to the *National coordinators* (NC). The logic behind the target response rate of 70 percent was firstly that several countries could have even higher response rates, secondly that fixing this norm could reduce the differences in response rates, and thirdly that it should be used as a beacon to guide the countries round after round in the right direction. What is the real situation after three rounds? In Round 1 (2002), the highest response rate was 79.6 percent and the lowest 33.0 percent. Thirteen countries obtained response rates higher than 60 percent. The mean response rate for 22 countries was 60.4 percent.[1] (Standard deviation: 10.6 points). The mean response rate for 26 countries in the 2004 survey (Round 2) was 61.5 percent (SD: 7.7 points). The highest response rate was then 79.3 percent, and the lowest 43.6 percent, with 15 countries obtaining response rates over 60 percent (Billiet et al. 2007; Billiet and Pleysier 2007). In Round 3 (2006) the lowest response rate was somewhat higher (46 percent) and the highest somewhat lower (72.7 percent) than in previous rounds. The mean response rate was 65 percent (SD: 7.1 points). The differences in response rates can still lead to nonresponse bias, even in the case that the bias would be invariant over countries which is an unrealistic hypothesis.

The large variation in response rates raises certainly the questions about nonresponse bias, especially about differences in bias between the countries. It

is clear that there is no complete relationship between degree of nonresponse and degree of bias (Groves 2006), however, it is also found that the likelihood of bias increases when response rates are smaller, especially when the factors that effect nonresponse are related to crucial variables in the population (see for ESS R2 Vehovar and Zupanič 2007). The work done in ESS in order to enhance response rates has been thoroughly documented in a number of studies (see Jowell et al. 2007; Billiet et al. 2007). In a joint research activity (JRA2) of the infrastructure project ESSi (EU framework programme 6), special attention is paid to the analysis of bias. Much of the work has still to be done but we have already a sound view on the crucial challenges in a cross-nation situation. The strategies that were developed in order to be able to detect bias (post-stratification weighting, refusal conversion, observable data recorded in the contact files) have been improved in the most recent round, and a specific survey among samples of nonrespondents was organised. The paper starts with a definition of nonresponse bias and a short overview of current approaches to nonresponse bias with focus on the approaches used in ESS. Then we will elaborate each of the applied approaches, illustrate these, and discuss the strong and weak points. This paper concludes with critical reflections about the limitations and profits of the applied approaches to nonresponse bias.

APPROACHES TO NONRESPONSE BIAS

As survey researchers know (Groves and Couper 1998) low nonresponse rates limit the possible nonresponse bias but there is no clear-cut relationship between response rate and nonresponse bias (Groves 2006). In actual discussion about improving response rates, some scholars argue that better than aiming for high response rates one should try to minimize nonresponse bias. This is however a much more difficult enterprise than enhancing response rates. Firstly, whereas a response rate is a relatively straightforward target aiming for minimal bias will be difficult to implement in practical fieldwork protocols. Secondly, within a survey nonresponse bias can vary substantially across variables (Groves 2006). Finally, it is often very difficult to assess nonresponse bias as it requires either population information with respect to the core variables of a survey, or similar information about the nonrespondents. Both are rarely available, at least in surveys on opinions, attitudes and values. In cross-national surveys the situation is even more complicated than in a single country situation. In order to compare survey results one would ideally have minimal nonresponse bias in every country. As this will not be possible, a second best option is a similar or comparable nonresponse bias in every country. This appears however to be equally problematic when the response rates across are very different.

Nonresponse bias in a cross-nation context

As we all know, nonresponse is an important threat to the validity of survey research. It is the failure to obtain responses (or measurements) for all sample units. Why is this a threat? The answer is simple: nonresponse *can* produce bias in the results. Nonresponse bias is a function of the amount of nonresponse and the difference between respondents and nonrespondents (Groves 2006: 648)[2] as is shown in the following expression.

$$\bar{y}_r - \bar{y}_n = (\frac{m}{n})[\bar{y}_r - \bar{y}_m] \quad (1)$$

In this expression, $\bar{y}_n$ refers to the sample mean, $\bar{y}_r$ indicates the respondent mean, $\bar{y}_m$ is the nonrespondent mean, and *m/n* is the nonresponse rate. Theoretically, then, the biasing influence of nonresponse is eliminated under two conditions: either (a) the nonresponse rate is zero (there are no nonrespondents) or (b) there are no differences between respondents and nonrespondents on the statistic of interest (Couper and de Leeuw 2003: 166).[3]

In cross-nation studies, these two factors, and thus non-response bias, may differ from one country to another. The cross-nation and longitudinal character of ESS renders this already complex matter even more difficult. Formula (2) illustrates the effects of nonresponse on cross-national survey estimates, in this case of the difference of two country means:

$$\bar{y}_{1r} - \bar{y}_{2r} = (\bar{y}_{1n} - \bar{y}_{2n}) + (\frac{m_1}{n_1}[\bar{y}_{1r} - \bar{y}_{1m}] - \frac{m_2}{n_2}[\bar{y}_{2r} - \bar{y}_{2m}]) \quad (2)$$

The subscripts 1 and 2 indicate two countries. The difference in estimated bias between the two countries is then (Groves, 1989):

$$B(\bar{y}_1 - \bar{y}_2) = \left(\frac{m_1}{n_1}\right)[\bar{y}_{1r} - \bar{y}_{1m}] - \left(\frac{m_2}{n_2}\right)[\bar{y}_{2r} - \bar{y}_{2m}] \quad (3)$$

When one analyses differences in country means and variances, valid international comparisons cannot be made without adjustments for non-response bias. Most cross-national or cross-cultural research however implicitly assumes that the bias is stable across countries or subgroups. Such an assumption would give evidence of an infinite naïveté since the hypothesis of equal bias and comparable response rates is very unlikely as it is shown in previous ESS rounds. Nonresponse error is relevant not only for simple descriptive statistics such as country means and differences between these means, i.e. proportions but possibly also for the estimation of correlations between variables (Couper and de Leeuw 2003: 166). Moreover, nonresponse bias also relates to the estimation of variances that are used to estimate standard errors.

Methods for assessing nonresponse bias

Groves (2006: 654-656) distinguishes *five* general methods for assessing nonresponse bias in household surveys. The attempt to obtain estimates for the missing observations in order to be able to adjust the survey estimates for nonresponse bias is the core philosophy behind these approaches.

The most easy approach is *response rate comparison across subgroups* even though it does not yield direct estimates of nonresponse bias in key statistics. This indirect method heavily rests on the assumption that there is no bias if the hypothesis of no difference in nonresponse between subgroups, is not rejected (Groves 2006: 654). The comparison of nonresponse rates among subgroups is however only possible for a small number of grouping variables (gender, age, urbanicity subgroups, region). Asserting that constant response rates over subgroups imply no nonresponse bias rests on the assumptions that the sub-grouping variables are the only systematic sources of nonresponse and that other variables only produce random nonresponse. This is the so called '*missing at random'* (MAR) assumption. Nonresponse can however covarie with more crucial variables that are not observed. A more specific problem in a cross-nation context relates to the differences in sampling designs. The method is not applicable when no individualized information about the complete raw sample is available.[4]

A second method for assessing nonresponse bias consists of *comparing response based estimates with similar estimates from other more accurate sources* (Groves 2006: 655). These sources are official population statistics or very large reliable surveys with minimal nonresponse rates (the so called '*Gold standard'*). Bias is then understood as the amount of deviation between the 'true' population distributions and the distributions in the obtained sample. The method relies on the same MAR assumption as previous method. This method of bias estimation and adjustment has been used in ESS (Meuleman and Billiet 2005; Vehovar and Zupanič 2007) and will discussed thoroughly further in this paper.

A third method is based on *variation within the existing survey* (Groves 2006: 655). Actually, this method covers a variety of variants some of which are very easy to implement while others require extra efforts and budgeting. Most simple and straightforward is the comparison of estimates from early and late cooperation in a mail survey or a web survey in which several recalls are organised. It is assumed that the late respondents are informative for final non-respondents. It is possible to compare the respondents from the first phase with those from the full respondent data set when a survey is planned in several phases (Curtin, Presser and Singer 2000). Another variant of the third method which is applicable for face-to-face surveys is the study of converted refusals (Smith 2002; Burton, Laurie and Lynn 2006). This method has been used in the *European Social Survey* and will

be profoundly discussed further in this article. Another variant of the third method makes use of follow up studies. One can obtain additional information about the non-respondents using the "*Basic Question Procedure*" method (Bethlehem and Kersten 1985: 292; Voogt 2004) at occasion of the 'normal' controls, or by means of a new survey among both respondents and non-respondents from the original survey after a short time period. One tries to obtain information about additional auxiliary variables in order to change the NMAR situation into a MAR situation (Little and Rubin 1987).[5] This approach has also been used in some countries in Round 3 of ESS.

The fourth general approach exists in *enriching the sample by matching the individual records of a sample with individual records from other sources* (see for example Lin and Schaeffer 1995). In these cases, much more information becomes available than what was available in the sampling frame (Groves 2006: 654). This opens the possibility of a more effective weighting procedure. When data at individual level does not exist, a weaker but more general applicable variant of the matching method is possible by matching the individual level records in the sample with records at a aggregate level.

Actually, one is never sure about the direction and size of nonresponse bias for all variables in the sample. It is however possible as a fifth approach to *compare several distributions each based on another hypothesis about nonresponse.* The adjusted data are then compared with the not adjusted sample data. The common idea behind this class of methods is that one attempts to measure the amount of nonresponse bias that might be eliminated by several post-survey adjustments (Groves 2006: 656). Four classes of adjustment methods are distinguished: *weighting, extrapolation, imputation and modelling* (Voogt 2004: 133; Brehm 1993; Bethlehem 2002). *Weighting adjustments* are based on the use of auxiliary information available for the whole population or for at least for the gross sample (nonrespondents included). *Post-stratification* (PS) has been used in all rounds of ESS and will be discussed extensively later in this paper. *Extrapolation* is based on the idea that certain groups of respondents are more comparable to the nonrespondents than others are (Voogt 2004: 134). Actually, extrapolation is under certain conditions useful when information about converted refusals has been obtained (see Potthoff, Manton and Woodbury 1993). *Imputation means* that the missing values among the nonrespondents are substituted by estimates (see Kalton and Kasprzyk 1986; Little and Rubin 1987). The post-survey adjustments for nonresponse bias are all based on distinctive models of response probabilities (Gelman and Carlin 2002; Rizzo, Kalton and Brick 1996; Laaksonen and Chambers 2006; Knot 2006).

The strength of post-survey adjustment methods is that a large set of alternative estimators supposed to measure the same population parameter can be compared.

When the alternative estimators are based on very different assumptions about the nature of the nonresponse, and when these are very similar in size, one can have more confidence in the conclusions of the survey. If they differ seriously, then the researcher should try to understand why. The weakness of these methods is that they lack a '*gold standard*' or an external benchmark that makes it possible to test the assumptions behind the alternative estimation methods (Groves 2006: 656).

Population based as well as sample based approaches were used in ESS, covering the following four methods for bias detection and estimation: post-stratification weighting; the comparison of cooperative with reluctant respondents; information from observable data among all sampled units; information from follow up surveys among nonrespondents. We will not further deal with the latter since the analysis was not yet completed at the time that this paper was prepared.[6] The three other methods are thoroughly discussed and illustrated using data from ESS Round 2.

BIAS AS DEVIATION BETWEEN SAMPLE AND POPULATION DISTRIBUTIONS

A rather easy way to obtain a view on nonresponse bias is the comparison between the obtained sample distributions with the population distributions on a number of variables of which the joint distributions are documented. The source of the "expected" distributions are reliable official population statistics, or other trustworthy sources that may be conceived as "*gold standards*". A chi-square test in which a (joint) distribution in the sample is tested against the expected (joint) distribution gives a first impression of the amount of bias. This method is useful when the sample is not stratified on variables that are used in the comparison (e.g. region, gender, age category), and when one can expect covariance between these variables and the variables of interest that are not documented in the population statistics. In some cases when a joint distribution of n variables is not available in the population statistics of some countries, but only for (n-1) variables, raking ratio can be used when the marginal distribution of the n-th variable is available (Kalton and Kasprzyk 1986). In this approach, bias has been defined as the difference between the means and distributions in the two samples. However, PS is not without discussion. Some authors have showed how PS of samples reduce possible bias due to nonresponse (Thomsen 1973), but others are somewhat more restrictive and admit that this is not always the case (see Kalton and Kasprzyk 1986). We will discuss this after the presentation of PS in ESS.

Post-stratification weights in ESS

In the context of the assessment of nonresponse bias, the function of post-stratification is double. One can first of all study the effects of post-stratification weightings on the distributions of the post stratification variables and on many

other variables in the sample. That way one has under certain assumptions[7] an view on the amount of bias in the sample, and on the variables that are most sensitive for nonresponse bias. Secondly, the weighted samples are, again under certain assumptions, considered as (somewhat) adjusted for nonresponse bias. A complete report of PS weighting in rounds 1 and 2 of ESS is provided in the studies of Vehovar (2007) and Vehovar and Zupanič (2007). To evaluate the effect of post-stratification weighting, a comparison was made between the unweighted and weighted means, which we – with some simplifications – attributed to nonresponse bias. The simplifications deal with the fact that non-random differences between estimates based on the realised samples and the population statistics are not only due to nonresponse, but also because of some defects in the applied sampling procedures or different categorisations of the PS variables in the population statistics. This means that the method from this point of view overestimates somewhat the amount of nonresponse bias as such, although it is in general underestimated because of the weak relationships between post-stratification variables and target variables. We should also note that, when we speak about the unweighted samples, we always refer to samples that are already be weighted by the so called design weights[8] that one should always apply when analysing ESS data sets. The design weights correct for the expected differences in design effects attributed to the differences in sampling design over countries.

In order to compare results across different countries, the data should be optimally weighted for variables that covarie strongly with the target variables in a study. But because the joint population distribution of such variables is unknown, the data are weighted with respect to gender, age and education.[9] These three variables are common in post-survey adjustments and in general available for all ESS countries in the survey documentation delivered by the *National Coordinators*. The approach used to estimate nonresponse bias and to correct the data for nonresponse is PS based on *strict post-stratification* or on the *raking method.* The latter has been used when post-stratification was impossible because no joint distribution about the three mentioned population variables used was documented in reliable population statistics. Concerning 24 countries involved in Round 2 of ESS, the raking method was applied in 10 country samples (Vehovar and Zupanič 2007). In order to undertake the weighting for (joint) distribution of gender, age and education, the age variable was grouped into three categories (15-34, 35-54, 55 and older). Gender has two categories. Information about the education level of the population was also grouped in three categories separately for each country: lower secondary or less; higher secondary; post-secondary. The education variable is somewhat more problematic than gender and age because in a number of cases the joint distribution of age and gender with education is not available in the population statistics, and because the coding systems are not

universal. In Round 2 of ESS a three category variable *eduvla* based on ISCED 1997[10] was used (Vehovar 2007: 338).[11] The categories are: (1) Not completed primary education, Primary or first stage of basic, Lower secondary or second stage of basic; (2) Higher secondary; (3) Post secondary, non-tertiary, First stage of tertiary, Second stage of tertiary.[12]

Step 1: Weight calculation and the impact on stratification variables (ESS Round 2)

In a first step, several statistics are computed in order to evaluate the size of deviations of the sample from the population distributions of the PS variables. In a second step, the effect of the weightings for gender, age, and education on a large number of substantial variables has been analysed. Post-stratification weights are computed by dividing the cell proportions in the multivariate table (gender x age w education) in the population by the corresponding cell proportions in the obtained sample (Rässler, Rubin and Schenker 2008: 375). Weights have a value 1.0 when the sample cell proportion is identical to the population proportion; weights are in the range $0.0 < w <= 1.0$ when the sample proportion is higher than the proportion in the populations, and they are higher than 1.0 when the sample proportions are lower than expected (in the population). This way of computing post-stratification weights is somewhat modified in the ESS datasets because of the design weights that are assigned to the sample units in the ESS datasets. The post-stratification weight factors are computed by dividing the population cell probabilities by the (design weighted) sample cell probabilities. When we refer to the *unweighted sample* (W1) in this article, we always mean the sample which is only weighted by the design weights.[13] The *final weighted sample* (W2) is then the sample weighted for the design weights and the post-stratification weights. This is done by multiplying the post-stratification weight coefficients with the design weights coefficients before analysing the data. It is possible to use this product of both weights since it is very likely that both are independent. Samples weighted by W2 reflect the population distribution of the stratification variables gender, age, and education.

The effect of weighting on the post-stratification variables

How serious is the impact of post-stratification weighting on the distributions of the post-stratification variables? There are several ways to answer this question. One can compare the distributions of the stratification variables between the W1 and W2 samples in all 24 countries.[14] We will only summarize the main findings and move then to the amount of variance inflation (VIF) in all countries since this is an indication of the impact of post-stratification weighting on the PS variables.

The marginal distributions of the variable 'gender' do not differ strongly between population and sample. There are somewhat more deviations in the age distribution. Most important is Spain where the size of the youngest category is seriously underestimated in the sample and the older age categories firmly overrepresented. In UK the oldest age category is overrepresented too, but in Belgium, the respondents over 55 years of age who are strongly underrepresented in the sample. There are other countries with rather high deviations in age distribution (Austria, the Netherlands, Iceland, and Luxemburg), but all by all, the deviations in age distribution are rather moderate. It seems all by all most likely that the size of the youngest age category (15-35 year) is underestimated in the samples, and the oldest is more often overrepresented. This finding is in line with the contact ability hypothesis (Groves and Couper 1998: 133-136). Older respondents are more easy to contact. It is also possible that older respondents are more cooperative because they have a greater sense of civic duty than youngsters.

If one takes the number of serious deviations between populations and samples into account, education seems then much more related to nonresponse. There are more serious deviations in no less than fourteen countries. The lower educated are seriously underrepresented in the samples of eight countries (CH, CZ, DE, EE, HU, NO, SK, UA) and considerably overrepresented in four countries (AT, LU, SE, UK). With exception of Austria, the size of the higher educated is mostly seriously overestimated (CH, EE, FR, HU, IS, NO, UA). The proportion of middle educated is also more often substantially overestimated (AT, FR, IS, NL, SE, UK) than underestimated (CZ, EE). The largest deviations between sample and population according to education is observed for the middle category of education in France (FR) and Iceland (IC), and the high level of education in Iceland (IS) and Ukraine (UA).

The most important conclusion from a cross-nation point of view is that there is no stable pattern of overrepresentation or underrepresentation in the categories of the post-stratification variables. Since all the samples were random samples, and since they are comparable over sample designs because the design weights were always applied in the computations, the deviations are reasonably assigned to nonresponse.[15] This finding means that there is no universal relation between nonresponse and background variables.

PS weighting and variance inflation[16]

Weighting does not only have an effect on the precision of the estimates (means and percentages), it has also consequences for the sampling variance and thus the estimation of the standard errors (Little and Vartivarian 2005). Weighting reduces bias but at the same time it may results in a loss of precision (Sturgis 2004). One should realise that the weights themselves are estimates (Rässler,

Rubin and Schenker 2008: 375). Whether or not one should use weights depends of the question whether the reduction in bias outweighs the loss in precision.[17] The estimated variance in the weighted sample is usually, but not always, inflated by the variation of the weights (Little and Vartivarian 2005). The coefficient of variation (CV_w) and the variance inflation factor (*VIF*) are used for evaluating this effect of weighting on the estimate of the variance (Vehovar 2007: 340-343).

The estimate of the increase of the sample variance is based on the well-known Kish (1965) formula for the coefficient of variation of the weight variable. (CV^2_w) expresses the ratio between the elementary variance of the weight variable *w* and the square of the arithmetic mean for the same weight variable *w.*

$$CV^2{}_w = \frac{s^2{}_w}{\overline{w}^2} \tag{4}$$

VIF expresses the increase of the sampling variance of a weighted sample in comparison with the sample variance (with the same sample size) where there would be no need for weights.

$$VIF = 1 + CV^2{}_w \tag{5}$$

According to this expression, the minimum value of *VIF* is 1.0 in case of zero variation of the post-stratification weights. The consequence of weighting is an increase of the sampling variance (unless *VIF* has it minimum value).

$$Var(\overline{y}_w) = Var(\overline{y}) \times VIF \tag{6}$$

These statistics are calculated separately for the W1 and W2 samples. The increase in sampling variance is one of the most important consequences of weighting for statistical analysis since it has implications for the rejection of the null hypotheses.

Table 1 Variance inflation factors for final weights in ESS Round 2 (Vehovar 2007: 343)

country	Vif	country	Vif	country	Vif	country	Vif
FI	1.02	DK	1.22	IE	1.45	UK	2.38
PL	1.02	DE	1.25	SE	1.62	HU	2.50
SI	1.02	GR	1.31	LU	1.82	CZ	2.81
ES	1.03	CH	1.36	AT	1.99	IS	3.03
BE	1.07	PT	1.42	SK	2.05	UA	3.31
NO	1.17	NL	1.44	EE	2.18	FR	4.02

Table 1 gives an idea of the variance inflation that might be assigned to the final weights (design weights times PS weights). The effect of PS weighting on itself is however somewhat smaller since the increase in the sampling variance due to the clustering, i.e. the design effect, is generally around 1.2 – 1.5 for this type of surveys, but can also be higher than 2 or even 3 for some variables that are related to the neighbourhoods that were used as primary sampling units (PSU's). After squared-rooting of the final *VIF*, the confidence intervals would however rarely expand to more than 10 percent (Vehovar 2007: 341). The left upper part of the table shows the country samples for which the design weights have their minimum value (1.0) and where the variance is only inflated because of post-stratification weights.

Step 2. The effects of the weightings as indication of nonresponse bias

In the PS approach, the size of the weights can be conceived as an indication of the amount of nonresponse bias. However, we must assume then that the differences between the weighted estimates and the unweighted ones are merely attributed to the nonresponse. This may be a risky assumption because we do not know about other potential sources of bias such as noncoverage bias, field work errors, processing errors, and measurement errors. The bias related to nonresponse is thus overestimated. However, given the reasonable controls of the central coordination of ESS over other error sources, we may assume that the bulk of the bias in the samples originates from the nonresponses. The moderate negative correlation between the response rates and estimated biases support somewhat this assumption (Vehovar 2007: 344).

In Round 2 of ESS no less than 45 items are included in the study of the deviations between W1 and W2 samples. The items are selected in each of the core sections and rotating modules according to their importance, relevance and appeal of the variables they intend to measure. The items cover the following issues: media (3 items); social trust (3 items), politics (12 items); well-being (3 items); religion (3 items); economic morality (6 items); family and life-work balance (6 items); socio-demographic profile (5 items); human values (4 items). The items are listed in Table A1 in Appendix 1. In this approach to nonresponse, bias is defined as the difference between the estimate in the unweighted and the weighted sample.

$$bias(\bar{y}) = \bar{y} - \bar{y}_w \tag{7}$$

Two measures are important, the relative bias and standardised bias. The *relative bias* (*Rbias*) provides a measure of the magnitude if the bias magnitude of the bias in comparison with the estimate itself.

$$Rbias = Relative\ bias = \frac{bias\ (\overline{y}_w)}{\overline{y}_w} \tag{8}$$

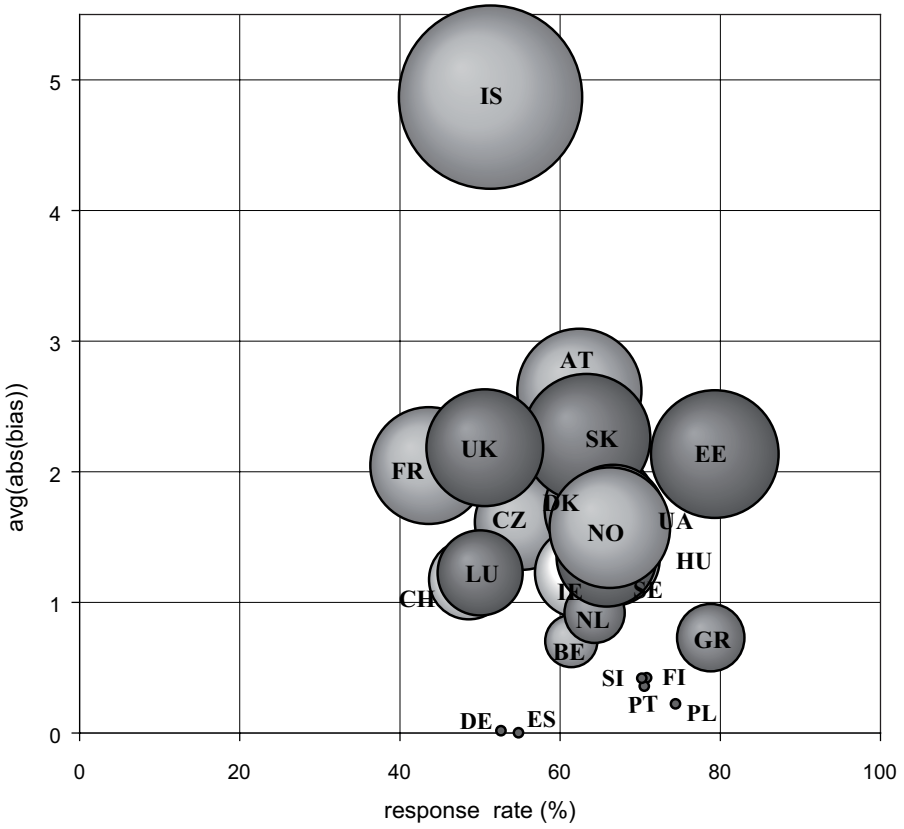
This ratio is needed if one wishes to compare the bias in several indicators which are different according to their mean values. The second measure, named *standardised bias*, compares the nonresponse bias with the standard error of the estimate, i.e. the sampling error. It is the ratio of both:

$$Sbias = Standardised\ bias = \frac{bias(\overline{y}_w)}{se(\overline{y}_{SRS})} \tag{9}$$

In this expression the standard error is calculated under the assumption of a simple random sample (*SRS*), which of course underestimates the sampling variability for the design effect and for the *VIF*. However, further refinements of *Sbias* could be obtained if the design weight and proper *VIF* are included. The results of the analysis of nonresponse bias according to these estimates contain a very large amount of information since there are estimates for 45 items in 25 countries, this means no less than 1,125 bias estimates.[18] In the remaining part of this section we focus only on the standardised bias (*Sbias*) which expresses the nonresponse bias as ratio to the standard error. We can then in principle use the usual 5 percent level of significance and take the value $t = 1.96$ as the benchmark, which denotes the statistically significant biases at the 0.05 significance level. The absolute average standardised bias (*absolute ASbiases*) for each item is reported; this is the absolute average value per item over all countries (see Table A1 in Appendix 1).[19] We observe that there are only six items with an average *Rbias* larger than our benchmark (> 1.96) – four of these are from the module of political attitudes. The majority of the absolute *ASbiases* are below one standard error of the corresponding estimate. Two of the items with largest estimated nonresponse bias are items on immigration. These findings about items that are most sensitive to nonresponse bias are comparable with previous research with Round 1 data (Billiet et al. 2007: 153-155).

The large amount of information is compactly shown in Figure 1. This is a three-dimensional presentation with the average absolute standardised bias in the vertical axis, the response rate at the horizontal axis, and the number of items with absolute *Sbias* larger than 1.96 in each country expressed in the size of the bubbles. The country with the largest number of items biased by nonresponse (according to our approach) is Iceland (IS) with no less than 37 items (out of 45) with a *Sbias* larger than 1.96. At the other end, we find six countries with no items of which the absolute *Sbias* is larger than 1.96. These are Germany (DE), Spain (ES), Finland (FI), Poland (PL), Portugal (PT) and Slovenia (SI). Four of these

Figure 1 The absolute average standardised bias in relation to the response rate of the country samples



Source: Based on Figure 3 in Vehovar (2007: 352)

countries had response rates over 70 percent. The relation between the amount of bias and obtained response is however disrupted by Estonia (EE) that has a very high response rate of 79.3 percent but nevertheless 18 items with absolute value of *Sbias* larger than 1.97. The response rate in the sample of Greece (GR) is nearly as high (78.8 percent), but the average standardised bias is small and only five items have *Sbias* larger then 1.96.

The impression that there is a negative correlation between the bias estimates and the response rates at the country level is supported by the correlations between the average absolute standardised bias and the response rate (r = -0.29), and between the numbers of items with standardised bias > 1.96 per country and the response rate (r = -0.26). Country samples with higher response rates are more likely to be characterised by smaller nonresponse bias.

As was already mentioned, only one component of *VIF* arising from fixed weight variation is included in our computation of *VIF*. The design effect from clustering is missing which means that the total *VIF* would be somewhat higher for the majority of countries in which cluster samples are used (Vehovar 2007: 352-353).

Does post-stratification on gender, age, and education matter for substantive findings?

This is a crucial question for the substantial users of ESS data who are not interested in the methodological sophistication and findings but in the implications for their substantive findings when they are comparing countries, or when country variables play a role as explanatory variables in their explanatory models. In these cases it is necessary that one can rely on the estimated statistics (means of latent variables or constructs, correlations, regression parameters, standard errors) concerning the relevant variables measured at individual level in each of the countries.

Good candidates for evaluating the effect of weighting on substantive findings are the distributions of the latent variables "*interest in politics*" (POLINT) and "*positive consequences of immigration for the country*" (CONSEQUE). The first variable *interest in politics* is a dimension of (internal) political efficacy and is also measured by three indicators referring to the respondent's interest in politics, the respondent's understanding what is going on in politics, and how difficult it is to form an opinion on political issues. The central indicator ("*how interested are you*…?") shows the highest bias of all checked indicators (see bottom of Table A1 in Appendix). The second variable can be interpreted as '*the evaluation of the consequences of immigration*'.[20] It was measured by three items concerning the consequences of immigration for the country's economy, for the cultural life, and for living conditions in general. Two of the three indicators of this variable are significantly biased (see bottom of Table A1 in Appendix). The central indicator ("*how interested are you*…?") shows the highest bias of all checked indicators. As we can expect because of the low weight factors in a number of countries, the W1 and W2 estimates are not very different. There are however some exceptions. For *political interest*, there are eight countries with a difference in W1 and W2 scores of 5 percent or more but only one exceeds 10 percent (Estonia). Concerning *consequences of immigration*, there are three countries with effects of final weights larger than 5 percent but none exceeds 10 percent. The largest differences are found in Estonia for POLINT and in Iceland for CONSEQUE.

In social research, we are not as much interested in descriptive statistics as means but in the comparison of explanatory models. We will therefore compare the W1 and W2 weighted samples in a substantive explanatory model for these two variables in the countries Estonia (political interest) and Iceland (consequences).[21]

One should also realise that the effects of weighting does not only concern the means but also the variances. In Figure 1 we found that IS (*VIF* = 3.03) and EE (*VIF* = 2.18) are among the countries with large amount of final weight (W2) variance inflation. Variance inflation for the design weights (W1) is zero for these countries since the weights are 1.0. When we compare regression models for these countries, we should multiply the standard errors by the *square root* of *VIF* in order to obtain adjusted standard errors (*adjust. SE*). However, we know that the tests are now somewhat conservative since *VIF* is overestimated for the post-stratification weights. Let us see what the effects are on substantive conclusions based on the adjusted standard errors of the unweighted[22] and weighted samples.

Let us first compare the W1 and W2 samples on *political interest* in Estonia (see upper part of Table 2). The standard errors for the W1 sample in Estonia is not adjusted since there are no design effects because of clustering in the sample and the design weights are all 1.0. The regression parameters are more or less the same in size in the W1 and W2 samples. In Estonia there are somewhat larger differences between W1 and W2 regression regression coefficients. The effect of age (the older the less interested) is also stronger in the W2 sample where it, contrary to the W1 sample, is statistically different. The effect of the higher secondary (versus higher education) is also somewhat stronger. Those with higher secondary education are somewhat less interested then those with higher education, and this effect seems stronger in the W1 sample. The strength of the effect of ever having a job is no longer significant in the W2 sample. All by all, in Estonia researcher may arrive to very small differences in the substantive model when a post-stratified sample (W2) is used, but the differences are very small.

It may surprise that in a country (Iceland) with a larger variance inflation factor in the W2 sample, and with a much smaller response rate than Estonia, does not show substantial differences in the conclusions of the regression analysis with adjusted standard errors. We observe serious differences in the size of some regression coefficients for gender, for those who had only lower educated or who finished higher secondary education, and for those who never had a job. These effects of these explanatory variables are stronger in the W2 samples, but this is not reflected in the probabilities (level of significance). These probabilities of obtaining a zero coefficient given the estimated values under the regression model are always lower, but still significant at 0.05 level. The reason might be that the sample in Iceland is much smaller than the other samples.

Does this all mean that there is nearly no nonresponse bias in ESS, or should we rather conclude that the assumptions behind the post-stratification method are responsible for the failure to detect bias in the target variables and adjust for it? It is reasonable to accept that the answer is partially yes on both questions. At one hand, ESS sampling and data collection are very well prepared and as much

Table 2 Comparison of explanatory regression models for *political interest* and *consequences of immigration* in samples unweighted (W1) and weighted for post-stratification (W2) samples in Estonia and Iceland (ESS Round 2)

Estonia (political interest)								
Explanatory variables	Unweighted sample (design weight = 1)				Final weighted sample			
	Unstand. coeff.	SE	t-value	Prob.	Unstand. coeff.	Adjust. SE	t-value	Prob.
Intercept	1.488	0.091	16.36	<.0001	1.578	0.115	13.781	<.0001
Male (= Yes)	0.329	0.032	10.37	<.0001	0.247	0.047	5.300	<.0001
Age	-0.001	0.001	-1.43	ns	-0.004	0.002	-2.683	<.01
Education Lower	-0.754	0.072	-10.49	<.0001	-0.744	0.102	-7.274	<.0001
Lowsec	0.467	0.054	8.68	<.0001	0.485	0.086	5.627	<.0001
Highsec	-0.282	0.040	-7.12	<.0001	-0.310	0.086	-3.611	<.01
Higher	ref.	ref.	ref.		ref.	ref.	ref.	
Urban	0.033	0.013	2.61	<.01	0.042	0.019	2.186	<.01
Active	-0.030	0.040	-0.77	ns	-0.102	0.062	-1.641	ns
Ever had a job	-0.225	0.072	-3.10	<.01	-0.127	0.094	-1.348	ns
Job control*	0.053	0.006	8.64	<.0001	0.052	0.010	5.114	<.0001
R^2	0.20				0.18			
Iceland (consequences of immigration)								
Explanatory variables	Unweighted sample (design weight = 1)				Final weighted sample			
	Unstand. coeff.	SE	t-value	Prob.	Unstand. coeff.	Adjust. SE	t-value	Prob.
Intercept	5.138	0.553	9.29	<.0001	5.566	0.953	5.843	<.0001
Male (= Yes)	-0.041	0.166	-0.25	ns	-0.214	0.280	-0.766	ns
Age	0.006	0.005	1.24	ns	0.015	0.009	1.726	ns
Education Lower	-1.076	0.363	-2.97	<.001	-1.426	0.692	-2.060	<.01
Lowsec	0.743	0.206	3.61	<.0001	0.757	0.448	1.691	<.05
Highsec	-0.559	0.202	-2.77	<.001	-1.077	0.425	-2.532	<.01
Higher	ref.	ref.	ref.		ref.	ref.	ref.	
Urban	-0.012	0.076	-0.16	ns	0.028	0.131	0.218	ns
Active	0.251	0.212	1.18	ns	0.371	0.359	1.035	ns
Ever had a job	-0.372	0.440	-0.85	ns	-0.881	0.698	-1.262	ns
Job control*	0.009	0.032	0.27	ns	-0.056	0.050	-1.125	ns
R^2	0.05				0.08			

* A latent variable measuring the amount of control one has over one's job.

as possible standardised. This may be a reason for minor bias. But at the other hand, in the post-stratification approach, the amount of bias reduction in the target variables depends on the strength of the correlation between the post-stratification variables and the target variables. In the cases we analysed, where the largest bias in the target variables was observed, the explained variance is rather moderate to low. It is somewhat lower when we keep only the three post-stratification variables in the models. Post-stratification weights can thus only reduce a small portion or the bias related to sampling and nonresponse since the covariance of the PS variables with our target variables is low.

The post-stratification approach to nonresponse bias: concluding remarks

There are several problems related to post-stratification: no distinction can be made between nonresponse bias and sampling bias; this method assumes MAR (missing ad random) within each combination of the stratification variables, and when this is not the case because of non-random missingness within these classes (NMAR) there can be still a serious undocumented bias; the size of the bias in the target variables can be seriously underestimated when the correlation of these variables and the post-stratification variables is low; and finally there is strictly no guarantee that the adjusted sample reflects better the distribution of the target variable in the population.

The PS estimator in a sample characterised by nonresponse may be biased in itself when the source (or "*gold standard*") does not accurately reflect the population distributions. The bias in the PS estimator only disappears if there is no relationship between response probabilities and values of the target variable within each stratum since all stratum covariates are then zero (Bethlehem 2002: 277-277). This is the case in the situation in which the strata are homogeneous with respect to the target variable, or in which the strata are homogeneous with respect to the response probabilities. But precisely this is mostly the weak point of the method since the covariance of variables like gender and age with the target variables is mostly very weak. This means that the target variables are heterogeneous within the strata. PS on the variable "level of education" is sometimes more effective but the joint distribution of this variable with the other demographics in the population is not always available with enough precision. Moreover, in cross-country research the classifications of the education variable are often not comparable in both population statistics and surveys. Bethlehem (2002: 279) reports however on grounds of his practical experience that nonresponse often seriously affects estimators like means and totals, but less often affects estimates of relationships between variables.

Despite the observation that the nonresponse bias as estimated here is in general relatively small in ESS Round 2 or, with some exceptions not very dramatic, we

must be aware that with our weighting (age/gender/education) we actually removed only one specific part of the nonresponse bias. Further improvements could be obtained if we include more control and more detailed variables and perform more sophisticated adjustments techniques which would more fully incorporate the auxiliary information. The existing control variables may also be improved. Since Census 2001 data available in most of the countries are becoming increasingly outdated (Vehovar 2007: 355).

Despite of these deficits, a major advantage of the post-stratification approach to the estimation of bias and adjustments in a cross-national context is that it is applicable to all country samples. One has at least an idea about the existence and direction of bias even when one cannot take it that a substantive part of the bias is removed, Moreover, standardised and comparable procedures are more easy to conduct. The approaches discussed in next sections of this article need to be considered with much more caution in this respect. There we have only data for a few number of country samples, and the procedures used differ much more between countries since these are dependent of many actors involved in the process of data collection.

BIAS AS THE DIFFERERENCE BETWEEN COOPERATIVE AND RELUCTANT RESPONDENTS

A second approach used to assess bias in ESS is obtaining additional information about the respondents who refuse cooperation by trying to convert them. The call record data related to ESS surveys contain detailed information on actual recruitment procedures followed by interviewers and the outcomes obtained for each sample unit. In view of analysis, the call record data are merged with the main data files. It is important to note that the researcher has then complete information about (nearly) all questions for both cooperative respondents who participated directly and reluctant respondents who are 'converted'. One can distinguish two perspectives with respect to the profile of reluctant respondents: one can assume that reluctant respondents are more similar to real refusers than respondents who were immediately cooperative (the *'continuum of resistance'* model); other scholars assume that reluctant respondents don't necessarily resemble those who finally refuse because people refuse for various reasons (the '*classes of non-participants model*') (Stoop 2005: 105-112). The underlying assumption of the *'continuum of resistance'* model is that with less field efforts these reluctant respondents would have been final refusals, and that with even more field efforts additional refusals could have been converted (Lin and Schaeffer 1995; Groves and Couper 1998). When one uses this approach for estimating bias, one should keep in mind that, without any additional assumptions, the converted respondents

are utmost comparable with those who refuse to cooperate and not with the non contacted sample units. Because of these necessary assumptions in this approach to nonresponse bias, we prefer to name it "*traces*" of nonresponse bias, knowing that we do not have a complete view on nonresponse bias.

In what follows, the focus will be on establishing whether there are actually substantial differences between cooperative and reluctant respondents in a selected number of country samples in which the amount of reluctant respondents exceeds 100. These countries are listed in the Table 3 below.[23] In a cross-national perspective, we do not only ask the question whether one can estimate the direction of bias in some countries but also whether this can be done at a comparable way for all countries involved in a cross-nation survey as ESS. Until now, we find in the three past rounds of ESS that the successes of refusal conversion are very different over countries. There is some improvement in final response rates, but this increase in response due to refusal conversion is minimal to moderate in most of the countries (see: Billiet and Pleysier 2007; Billiet et al. 2007; Beullens et al. 2008). The effect of refusal conversion, expressed as proportion of the initial refusals that are converted, ranges from 0.02 to 0.41. In ten countries, this effect is lower than 0.05. This clearly demonstrates the large differences in refusal conversion practice between countries. In some countries virtually all initial refusals are re-approached in view of refusal conversion while in other countries only a small portion is re-approached, and that selection process is not random at all (Beullens et al. 2008). This has serious consequences for the usefulness of refusal conversion for bias detection (and adjustment) in a cross-nation context.

Methodological decisions

The classification of the respondents in cooperative and reluctant respondents, and a further refinement of kinds of reluctant respondent, is based on the information obtained by means of call record data. Concerning Round 2 of ESS, the reluctant respondents are compared to cooperative respondents on a number of background variables, attitudinal variables[24] and indicators of media use. The focus is on Switzerland, Germany, Estonia, the Netherlands and Slovakia. In these five countries, refusal conversion efforts have led to a considerable number of additional respondents (See Table 3).

The background variables under study include gender, age, level of education, partnership status, number of household members, urbanisation level, labour market status, religion, health status and citizenship. Additionally, the group of cooperative and reluctant respondents will be compared with regard to attitudinal variables and indicators of satisfaction and integration which are believed to be related to refusal to participate in the survey. These attitudinal variables comprise

attitudes towards political and social institutions, social trust, attitudes towards immigrants and perceived ethnic threat. The indicators of satisfaction and integration refer to satisfaction with government and own life, feel comfortable about income, feel discriminated, social isolation and feeling safe. Finally, also the distribution of media use is considered. Cross-cultural equivalence of the multiple indicator latent variables was tested and documented in previous studies (Billiet and Meuleman 2008; Davidov et al. 2008). The survey questions that showed a rather large absolute average standardised bias according to the post-stratification approach are all included.

Table 3 Number of cooperative respondents, initial refusals, percent re-approaced, and reluctant respondents in five countries*

	Cooperative respondents	Initial refusals	Percent initial refusals re-approached	Reluctant respondents
Switzerland	2059	2190	76.0	175
Germany	2378	2340	48.5	494
Estonia	1789	485	67.6	201
Netherlands	1358	1375	87.8	526
Slovakia	1407	652	40.0	105

* Slovenia had also a large number of converted refusals, but because of defective identifications in main file and contact forms file, these data could not be analysed.

Detection of nonresponse bias in multivariate logistic regression models

We directly focus on the relation between some of these variables and the kind of respondent (cooperative/reluctant) in the context of multivariate logistic regression models. In such models the most dominant relationships emerge, while spurious relations vanish. Table 4 gives the results of a logistic regression model[25]. The response variable is the type of respondent (reluctant versus cooperative). For categorical explanatory variables with more then two categories, effect coding has been used.

In Switzerland, the model reduces to one single parameter that relates to the number of household members. Sample persons from larger Swiss families seem to be more reluctant to participate in ESS Round 2. In Germany, reluctance is associated with being female, being aged, living in big city dwellers, internet surfing, having a history of unemployment, and less political participation. In Estonia, a similar relationship between gender and reluctance is observed, together

with having a paid job. Living in a village and ever been unemployed is more connected to cooperativeness.

In the Slovakian sample, the estimate for reluctance versus cooperative increased somewhat with age. Respondents who obtained a middle level of education, who are more religious, and feel comfortable with their family income are somewhat more likely to belong to the converted refusals. Job experience in the past (ever had a job) and feeling safe in the neighbourhood after dark have both a negative effect on reluctance, and vice versa, a positive effect on cooperative respondent behaviour.

Table 4 Logistic regression estimates (β-parameters) for reluctant versus cooperative respondents

	Switzerland	Germany	Estonia	Netherlands	Slovakia
BACKGROUND VARIABLES					
Male (=Yes)		-0.2275*	-0.3478*	-0.4872***	
Age		0.0185***			0.0184*
Household size	0.1630**				
Urbanisation level					
Countryside – village		-0.0684	-0.3908**		
Town		-0.2032**	0.2571*		
City		0.2716***	0.1337		
Level of education					
Education low				-0.2016	-0.3833
Education middle				0.2176**	0.5642**
Education high				-0.0161	-0.1810
Labour market status					
Paid job (1 = yes)			0.6224***		
Ever job (1 = yes)				0.5273*	-0.8563*
Ever unemployed (1 = yes)		0.3051*	-0.5380*	-0.5296*	
Good health (1 = yes)				0.2589*	
Comfortable income (1=yes)					0.6352*
Religious involvement (0-10)					0.2881***
ATTITUDES					
Perceived ethnic threat (0-10)				0.1278***	
Trust political inst. (0-10)				0.0962*	
Political participation (0-10)		-0.0934*		0.0915*	
Civil obedience (0-10)				0.0648*	
SATISFACTION & INTEGRATION					
Satisfied with life (0-10)				-0.1488***	
Social isolation (0-10)				0.0701*	
Safe after dark (=yes)					-0.7675**
MEDIA USE					
TV watching (minutes/day)				0.0036***	
WWW (no to daily = 0-7)		0.0627**		0.0501*	

*** p < 0.001; ** p < 0.01; * p < 0.05

In the Netherlands, the likelihood of being a reluctant respondent increases when the respondent is a female,[26] has an average education level, watches more television and surfes more frequently on the internet. Reluctance is also more likely when one ever had a job and feels more healthy. Respondents in the Netherlands who see immigrants more as a threat are more likely to belong to the converted refusals than those who feel less threatened. That was already found in the Round 1 study on reluctance (Billiet et al. 2007).

The effects of trust in political institutions, adhering civil obedience and participate more in (political) organisations were not expected. Respondents who share these attitudes are somewhat more likely to belong to the reluctant respondents. The effects of feeling socially isolated and dissatisfaction with own life are in the expected direction.

In sum, we find that the type of respondent 'reluctant' versus 'cooperative' is related to social-demographic variables, attitudinal indicators and other interesting variables, and that this effect still exists after controlling for the background variables. However, the bias induced by 'reluctant' respondents is not comparable over countries under study. We do find traces of bias in some countries, and not in others, and the predictors are not the same everywhere. It is possible that the quality of the obtained samples differs, but it is also possible that the differences between countries are artefacts of differences in the practice of the survey organisations, and of interviewer behaviour and decisions in the field.

This would mean that the category of converted respondents is not comparable over countries because in one country nearly all refusals are re-approached while in another country a selection is made at basis of information collected in previous contacts (see Table 3). Some segments of refusals may prioritized when in a survey organisation, the field supervisor (or the interviewer) select cases for refusal conversion that are most likely to cooperate at occasion of a refusal conversion attempts. This may result in an over-representation of 'soft' refusals among the converted refusals, that are not representative for all final refusals.

Are there differences in soft and hard refusals among the reluctant respondents?

We will now try to find out whether there are differences between countries in the proportion of soft and hard refusals according to countries and the decisions by survey organisations and interviewers. Best candidates for answering this question are the samples of the Netherlands and Germany since the amount of reluctant respondents is large enough to differentiate.

We have found in previous research by means of correspondence analysis that in the German sample of initial refusals optimal distinction between kinds of reluctant respondents is best made at basis of the responses of the interviewers to

the question how likely they estimate future cooperation of the target respondent. In the German sample, the probability of reissuing is largest among the target refusals who are classified as "probably cooperates", those with missing estimation of future cooperation, and the refusals by proxy. However, conversion success is highest among those who "probably cooperate" or those without estimation, especially when a new interviewer is mobilized (Beullens et al. 2007: 16).

The reissue probabilities are much higher in the Dutch sample of refusals. The crucial variable of differentiating between kinds of reluctant respondents is not the estimation of future cooperation by the interviewers, neither proxy refusal nor household refusal before case selection, but the number of refusals that occurred before refusal conversion. In the Netherlands, 44 percent of the reluctant respondents refused twice or more before they were convinced to cooperate (Beullens et al. 2007: 23). These differences in reissuing of refusals between the German and Dutch initial refusals, resulting in differences in composition of the reluctant respondents can explain the finding that in the Dutch sample we can find much more multivariate effects of background variables and attitudes on the probability ratio of being a reluctant respondent versus cooperative respondent (odds ratio's). Table 5 reports the significant covariates of a multinomial regression model, wherein the likelihood of kind of refusal (once or twice) versus direct cooperation according to relevant characteristics in the Netherlands is investigated.

Table 5 Multinomial baseline logit estimates (β) and odds ratio's* of belonging to soft and hard refusals versus cooperative respondents (reference) with respect to background, attitudinal, and media use variables (ESS Round 2, the Netherlands)

Predictors	Refused once		Refused twice	
	β	Odds ratio	β	Odds ratio
Male	-0.1301	0.878	-0.9279***	0.395***
Single	-0.2908	0.748	-0.4962**	0.609**
Level of education				
Low education	-0.2019	0.817	-0.4006*	0.670*
Middle education	0.2314*	1.260*	0.1146	1.121
High education	-0.0295	0.971	0.2860*	1.331*
Minutes watching television / day	0.0028*	1.003*	0.0023	1.002
Minutes reading newspaper / day	-0.0013	0.999	0.0077***	1.008***
Perceived threat by immigrants	0.0781	1.081	0.1562***	1.169***
Trust in political institutions	0.0782	1.081	0.1450**	1.156**
Social isolation	-0.0028	0.997	0.1123**	1.119**
Satisfied about own life	-0.0929	0.911	-0.2040***	0.815***
Max-rescaled R-Square		0.0755		
-2 Log Likelihood		2786.151		

Odds ratio = exp(β). Effect coding has been used for categorical explanatory variables with more than two categories.
*** p < 0.001; ** p < 0.01; * p < 0.05

It is striking that in most of the cases the parameters are only significant for the ratio 'refused twice/cooperative', and not for the ratio 'refused once/cooperative'. This means that the effect of the background variables on cooperation is most pronounced when cooperative respondents are compared with the reluctant respondents who were most difficult to convince. The reluctant respondents who refused twice (hard refusals) may be most informative for final refusals since in the Netherlands nearly all refusals are re-approached. The following variables are significant in the model: gender, family size (single versus multiple), education, TV watching and newspaper reading, perceived threat from immigrant, political trust, social isolation and the satisfaction with his own life.

How to interpret these parameters? An odds ratio of 1.0 means that there is no effect at all of a predictor. The larger the deviation from 1.0 the larger the effect of a category of a predictor compared with the reference (in case of categorical variables), or the stronger the probability ratio changes for one unit change in the predictor (for quasi metric variables). Parameters between 0 and 1 indicate a decrease in the ratio, while parameters larger than 1 indicate an increase in the ratio compared. One should realise that odds ratio's are proportional to changes in probabilities but they do not express changes in probabilities but in probability ratio's between a category of the dependent variable and the reference.[27] In this case are the ratio's 'soft (refused once) and 'hard' refusals (refused twice) versus cooperative respondents (reference). The odds ratio 'refused twice/cooperative' when the respondent is male is only 0.395 of the ratio for a female. Or vice versa, the ratio 'cooperative/refused twice' is 2.532 higher for females than for males. Among all respondents, male are thus less likely to belong to the 'hard' reluctant respondents. Does this mean however that women are less likely to cooperate in the survey than men? Or does it simply mean that females are much more inclined to participate after repeatedly insisted by the interviewer?

How to adjust the samples using information from reluctant respondents?

This question is an excellent start for reflecting on the way the refusal conversion approach can lead to adjustment of the data for nonresponse bias in a cross-national context. It all depends on the question whether it is possible to obtain (non)response probabilities for all sample units (respondents and nonrespondents) in all samples of a cross-nation survey? The way of proceeding is not that straightforward as in the case of poststratification weighting since several conditions must be fulfilled and even more assumptions must be made enough plausible.

A common way of adjusting for nonresponse bias at basis of logistic regression in situations where more than the classical poststratification variables (from population) are available, is the computation of weights based on response

propensity scores. This weighting technique aims to correct for differences caused by the varying inclination of individuals to participate in a survey. In order to obtain propensity scores, one should rely on a source which provides unbiased estimates. This source is normally a probability-based reference survey with much better response rates that is believed to produce unbiased estimates (Bethlehem and Stoop 2007). This is the so called "*Gold standard*" which is used to improve the target survey. Through logistic regression, the probability of each respondent participating in the target survey that has to be adjusted is estimated according to a set of relevant variables (Lee 2006; Loosveldt and Sonck 2008).

A number of serious problems must be solved before applying propensity scores based on information of reluctant respondents in order to adjust the samples for nonresponse bias. First of all, one should realise that the information obtained from the reluctant respondents (such as reason of refusal) has not been measured among final nonrespondents. The information is only based on a selection of initial refusals. Moreover, refusing is only one kind of nonresponse. The failure of contacting selected sampling units may be caused by other factors. One should combine information of refusals with information about non-contacts for adjusting the realised samples. It is therefore better to use the term '*refusal bias*' than '*nonresponse bias*' when this combination is not possible.

Second, the information obtained from a sample of converted refusals is only useful for the computation of propensities scores when all respondents who refused, or a random sample of them, are re-approached. An analysis of several rounds of call record data in ESS, shows that this is not the case in ESS. Both, National Coordinators (or Field Directors) and interviewers made systematic choices based on information about the refusals and subjective estimates about their future cooperation (Beullens et al. 2008).

A third problem deals with the validity of the logistic regression parameters for reluctant respondents as indices of the likelihood of nonresponse within classes of the predictors (independent variables). Actually, the information has been obtained among initial refusers who are ready to cooperate after new interventions. The assumption that the reluctant respondents offer valid estimations of parameters among the final refusals is rather weak. Using the reluctant respondent method, one can obtain a view on the direction of bias in some variables but the method does not provide precise estimates of response propensities.

Finally, and most important from the perspective of cross-national comparison, even when an improved method with valid estimates is possible within one single country sample, or a couple of country samples, one still misses the comparable data and adjustments for *all* countries. The proportion of converted refusals is still very disparate over country samples, and where sufficient cases are available, the results are not stable over countries and rounds.

USING OBSERVABLE DATA

The call record data that are collected among all selected sample units contain information that has been collected by means of observations by the interviewer. The interviewers are charged to record (estimated) age category and gender of each contacted sample unit by means of observation. This information is in principle available for all contacted nonrespondents, and not only for converted refusals. Other information that is in principle available for all selected cases in the sample, refusals and not contacted included, deals with the context of the selected sampling units: the type of housing where the sampled person lives in, and some neighbourhood characteristics. This information was precisely collected for assessing nonresponse bias. The quality of this data was rather weak in some countries, but it is better in later rounds of ESS (Cincinatto et al. 2008).

In this approach we obtain additional information about all sample units, both respondents and nonrespondents. This is an advantage over the reluctant respondent approach. A major weakness is however that there is only information on a very limited number of variables, and that the measurements can be less reliable since these are interviewer observations without very strict observation scheme's or training of them. The measurement of these variables is done by means of subjective estimation and appreciation by the interviewers. The observations about housing and neighbourhood must be classified in a limited number of pre-coded categories about the type of housing and state of the neighbourhood. The utility of this approach in view of bias estimation will be discussed after a summary of the main findings of ESS Round 2 observable data[28].

Measurements and analysis

Since the collection of additional observable information about all selected sample cases is a very demanding task, we expected a larger amount of missing data than in other sections of the contact forms. It is mostly impossible to obtain data about gender and age of the selected sampling units in case of no contact at all, or in case of refusal before the respondent selection took place in household and address samples. Country samples in which the amount of missing data is too high will be dropped from the analysis. The threshold for the combined gender and age variables, and for the combined housing and neighbourhood variables was set at maximum 10 percent missing. For the respondents' age and gender, only five country samples are below this threshold. We will therefore not use these variables in the analysis. The situation of the housing and neighbourhood variables is much better. Fourteen country samples are useful for analysis.

As in previous approach on refusal conversion, we prefer to proceed with multiple indicator construct instead of single questions whenever possible. The

question about the type of housing the respondent lives in, is clearly a variable on itself. The three remaining questions about the physical state of the buildings and dwellings in the array, about the presence of litter, and vandalism are related. Factor analysis shows that an equivalent configuration of two latent variables, litter and vandalism, applied to all fourteen countries. The third question about the physical state of the buildings does not belong to this construct. The two factors 'litter/vandalism' and 'physical state' are used in further analysis. The correlation between these two variables ranges from 0.22 in Switzerland to 0.60 in Poland. Neighbourhoods in which litter, rubbish or vandalism is common, are more likely to have buildings and dwellings in bad conditions.[29] A strange outlier is Austria where the correlation between the two variables is negative (-0.19). When problems of multicolinearity are met during analysis, the two variables are combined into one 'neighbourhood condition' variable. This is the case in the samples of Portugal and Czech Republic (Cincinatto et al. 2008: 15).

In order to study the effects of the housing and neighbourhood variables, multinomial logistic regression *(baseline category logit)* modelling[30] is used with type of respondent as dependent variable. Possible outcomes are *initial refusal* or *final non-contact* versus *cooperative*. The explanatory observed variables are mentioned in previous paragraph. Commonness of litter and/or vandalism in the neighbourhood as reported by the interviewers are (quasi) metric variables. Higher values correspond to neighbourhoods that are in a relatively bad condition and that are more prone to litter and/or vandalism. The type of housing is a categorical variable. The most optimal categorization is in two classes, 'apartments' and 'other houses' containing the remaining housing types. The interaction effects between the housing type on the one hand and the remaining neighbourhood variables on the other, are always tested. In order to find out what interactions must be included in the final model, a stepwise regression has been performed.

The basic idea behind the analysis was to find for each country a *parsimonious* multinominal logistic regression model for explaining the outcome variable under consideration. Step by step, those variables that did not had any significant contribution to the model were eliminated, respecting the hierarchical structure in the model: first non-significant interaction terms were dropped, and then the additive terms. This was done until the model did not significantly deteriorate. After the *parsimonious model* was determined for each country sample, the analysis of variance statistics (*degrees of freedom, Wald chi-square and the probability level*) for each variable were examined in order to obtain an idea of the explaining power of each explanatory variable. Finally, the parameter estimates of the retained multinomial logistic regression model were reported and discussed (Cincinatto et al. 2008). Only the final parameters estimates are shown in this article.

The effects of neighbourhood characteristics on initial nonresponse and final noncontact

Table 6 contains only the odds ratio's belonging to the variables that had a global significant effect ($p < 0.05$) on the probability ratio's *'initial refusal/cooperative'* and '*noncontact/cooperative'*. The significant effect of the global variable does not mean that all categories of this variable are significant, but these are still reported in this case. Non significant main effects of variables are also reported when these variable are in a later step included in a significant interaction. In some countries, the parameters '*noncontact/cooperative'* are not tested when the size of the noncontact category is too small. At bottom of the table, two countries (CZ and PT) are separately reported because of the somewhat different operationalization of the independent variables.

In eight out of fourteen country samples, the effect of living in an apartment as compared with other types of housing on the probability ratio 'initial refusal/ cooperative' is positive. This means that in most countries (CH, EE, ES, FI, GR, NL, PL, and PT) it is more likely that a selected sampling person initially refused to cooperate when she or he is living in an apartment and not in another types of housing. The effects are also positive in five other countries but not significantly different from zero at 0.05 level. Austria is an amazing exception. This country will not be discussed in this paper because a profound study on this is needed before concluding something on this. The findings in Austria are different from the countries (BE, CH, GR, IT, and CZ) where significant effects on noncontacts were found. It is in these countries more likely that the contact attempt fails when the sampled person is living in an apartment. In sum, one can conclude that the housing situation clearly plays a role in the level of response rates. A rather serious effect of housing on initial nonresponse is observed in Poland (odds ratio is 1.484).

The largest effect on failing to contact is observed in Greece where the ratio '*noncontact/cooperative'* is 2.358 times higher if one lives in an apartment than in another housing type. When reflecting on nonresponse, one should realise that the non-contacts are final noncontacts in contrast to a number of initial refusals that were afterwards converted into (reluctant) respondents.

When looking at the physical condition of houses and dwellings in the neighbourhood, and at the presence of litter and/or signs of vandalism, one finds out that the physical condition is more often related to survey participation than the presence of litter or vandalism is.[31] The effect of the physical condition of houses and dwellings in the neighbourhood is positive on *initial refusing* in eight cases, including Portugal where the two neighbourhood variables are combined. This indicates that sampled persons who live in neighbourhoods in a relatively bad

Table 6 Multiplicative logistic regression parameters (odds ratio's) of the neighbourhood variables on the contact outcomes ESS R2. (R = *initial refusals/cooperative;* NC = *final noncontacts/cooperative*)

Country	Housing type: apartment Ref: other houses		Physical condition		Litter and/or vandalism		Interaction housing type and physical condition Ref: other houses		Interaction housing type and litter and/or vandalism Ref: other houses	
	R	NC	R	NC	R	NC	R	NC	R	NC
AT	0.835**	0.848*	0.873***	0.916	1.461***	1.462***	0.680***	1.261***	0.918	0.798**
BE	1.083	1.525***	1.158***	1.501***						
CH	1.158***	1.434***	1.046	1.390***						
EE	1.274***	1.186								
ES	1.276***	1.081	1.229***	1.426***	1.023	0.883	1.098*	1.272**	0.829***	0.765**
FI	1.205***	()	1.251***	()						
GR	1.350***	2.358***								
HU	1.151	()	0.834*	()						
IT	1.020	1.811***	1.272***	1.472***	0.865*	0.818				
NL	1.122***	()	1.163***	()						
PL	1.484***	()		()						
SK	1.035	1.484	1.252***	0.397	1.132	2.045	1.191**	1.552	0.792***	1.398

Country	Housing type: apartment Ref: other houses		Overall condition of neighbourhood		Interaction housing type and overall condition of neighbourhood Ref: other houses	
	R	NC	R	NC	R	NC
CZ	1.028	1.117**	0.936	1.055	0.939	0.869**
PT	1.151***	()	1.471***	()		

() Non-contacts excluded from the analysis because number of observations in category is too small to obtain stable estimation.
*** p < 0.001; ** p < 0.01; * p < 0.05; *Empty cells*: variable not in parsimonious model.

condition are more likely to initially refuse cooperation in the survey. The situation in Austria is again surprising since it goes in the opposite direction.

The ratio '*noncontact/cooperative'* could be tested in eight cases. In five of these, the effect is significantly positive indicating that failing to establish contact with the sampled person is more likely in neighbourhoods characterized by bad physical conditions. There is apart from Austria only one main effect found. The direction of the weak significant effect in the Italian sample is not in the expected direction.

There are significant interaction effects found in four countries. These interactions between housing type, the physical state of the buildings, and the response variable are in the expected direction in Spain and Slovakia. An additional effect of living in an apartment on *initial response* and *noncontact* is detected in neighbourhoods where the houses are in bad physical condition or in neighbourhoods that are characterised by litter and vandalized. The main effects of living in an apartment on *noncontacts* was not significant in these two countries, but it is significant in interaction with the neighbourhood characteristics.

The significant interaction effects between housing type, litter or vandalism, and initial refusal or noncontact in Spain and Slovak Republic are in the expected direction. This means that in these countries those who live in an apartment in a neighbourhood characterized by litter and/or vandalism, and not in other housing type, are less likely to refuse initially or not to be contacted. A comparable interaction effect on the final non-contact is observed in Czech Republic. One explanation for this mitigating effect of litter and/or vandalism for apartment-dwellers is that 'apartment' is too broad a category and could mean different kinds of dwelling types for people with different activity patterns.

Discussion: The effect of neighbourhood variables on response outcomes

The main advantage of the approach based on observable data in the contact forms is that it is in principle possible to obtain auxiliary information about all the selected sampling units, the respondents and the nonrespondents. One can find traces of bias as far as the neighbourhood variables are related to other substantive variables in the survey. This is anyway the case for education that is in a number of countries substantially correlated with housing and neighbourhood characteristics. Main disadvantage is however that the observations about type of housing and neighbourhood characteristics are subject to interviewer's interpretation. This results in a larger amount of missing data. Some unexpected findings might be caused by negligence in recording the information or even inappropriate instructions. Improving of data quality is certainly possible by an additional training of interviewers and by stricter controlling the quality of the observable information in the contact forms.

The basic finding concerning the kind of housing which was meaningfully reduced to two main categories 'apartment' and 'other houses', states that those who are living in apartments are in most countries more likely to refuse initially to participate in a survey. The effect of housing type on the likelihood of not contacting selected units is mostly in the same direction. Where the neighbourhood variables have an effect, it is dominantly in the expected direction. Refusal and non contacting sampled units is more likely in areas characterized by bad physical condition of the houses or by the presence of litter and/or vandalism. Is it because of characteristics of the sampling units living in these neighbourhoods, or do prior expectations of interviewers play a role?

In principle one can use the complete sample with additional information about the observable variables among respondents and nonrespondents as auxiliary variables in order to correct dataset with respondents by means of weightings based on propensity scores. Given the differences in quality of the data recorded in the contact forms, it is at this moment not possible to formally correct the observed samples for bias, but we have the possibility to find indications in what other variables nonresponse bias is likely. We have found in nearly all countries a significant and moderate negative correlation between level of education and the physical state of the houses in the area. In a number of counties there are also correlations between some housing types (detached houses, farms…) and the education level of those who live in these houses.

FINAL CONCLUSIONS AND DISCUSSION

In previous rounds of ESS, traces of bias were explored and models were tested in order to find out whether nonresponse could have an effect on relevant constructs (Billiet et al. 2007). Several problems emerged in each of the approaches that were used in ESS until now, and several questions must still obtain a fair answer.

The post-stratification approach has the advantage of estimating for some variables complete nonresponse bias in all its components (refusal, non-contact, other). The method for correction via PS-weighting is in principle straightforward. However, this approach overestimates nonresponse bias since it contains also sampling deficiencies, and most important, it has no serious effect on the target variables to the degree that the covariation with the post-stratification variables is weak. The challenge is to find additional weighting variables that are stronger related to the target variables and for which the estimated population distributions are reliable. Moreover, the source used as a "*gold standard*" is often problematic and may be biased in itself, which makes correct weighting more complex. Do official population statistics represent fairly the distributions in the population? Are the national coordinator reports optimal sources for comparison between

countries and for comparison over time? Is the information about joint distributions applicable for all countries in a cross-nation survey?

The problems are even larger when one tries to detect bias by comparing reluctant respondents with cooperative respondents. First of all, the numbers of converted respondents are too small in most countries in order to arrive to stable conclusions. The results are mixed since bias is not always found in the same variables in those countries that are usable for analysis (see also Lynn et al. 2002). It was found that the effect of nonresponse on bias is not stable, neither between countries neither in the same countries over time. This makes it impossible to rely on this kind of information when the aim is adjusting for nonresponse bias using comparable information about the target variables in all countries. Reason of the instability can be partially due to the measurement. Too many decisions concerning the selection of original refusals for refusal conversion attempts are arbitrary, or are at least not comparable between countries (Beullens et. al. 2008). The classification into the category 'converted refusal' depends too much on differences in interviewer decisions or even differences in 'fielding culture' concerning the treatment of initial refusals. Differences between countries in privacy regulations may also affect the cross country differences in characteristics of reluctant respondents. Most important defects in the refusal conversion approach are the narrow definition of nonresponse, and the assumption that converted refusals reflect the final nonrespondents. It is at best *refusal bias* that has been studied since we cannot assume that nonrespondents and non contacted sample units are comparable. It has no sense to propose correction methods using the information of reluctant respondents as long as these problems are not solved.

Does this mean that the reluctant respondents approach is useless from the viewpoint of bias estimation in a cross-nation situation? Even when no complete comparable information for all country samples exists, it can warn the researchers against serious bias in some variables. The size of the sample increased substantially in a number of country samples because of refusal conversion. It may also improve the survey climate within a survey organisation since it is communicated to the interviewers that investment in high response rates is taken seriously.

The information about observable variables collected by means of a short observation questionnaire at the end of the contact has the advantage that it provides additional information about all sampling units (cooperative, reluctant, refusals, non-contacted…). Condition is however that all selected units are personally visited at location. So in principle one can produce propensity weights at basis of the augmented sample. Additional to the problem we have already met in the PS approach, the measurement of the observable variables has some problem, to a certain extent comparable with the defects in the PS sources. The answers to the questions concerning observable information depend too much on

the subjective appraisal of the interviewers. That is the reason why there is an serious interviewer effect on these questions. These questions are not very useful for correction techniques as long as there is no special training of the interviewers on recording observable data.

NOTES

1 This is the mean at country level, and not weighted by the sample sizes per country.
2 Groves provides a more general expression that takes account of the idea that everyone has an unobserved "propensity" of being a respondent or nonrespondent. The sample based expression (1) does not have an expected value equal to the population expression, but rather includes a term involving the covariance between the nonresponse rate, on the one hand, and the difference between respondent and nonrespondents means, on the other (Groves 2006: 648).
3 Notice that in expression (1) no distinction has been made between the components of nonresponse such as the refusal rate and the noncontact rate (see: Groves and Couper 1998: 12; Heerwegh et al. 2007).
4 In second round of ESS, there were 13 individual named samples, 7 household samples, and 6 address samples.
5 In case of nonresponse, the completely missing at random hypothesis is not realistic. Missing at random (MAR) means that nonresponse is independent from the study variables given a set of variables of which the population distributions are known. MAR allows the missingness mechanism to be related to covariates and observed survey outcomes of the study variables. In this case one can correct for the non random nonresponse in the variables with known population distributions. The missing data is then called ignorable, and one has not to model the missingness mechanism. Most likely situation however is not missing at random (NMAR) which means that the nonresponse is dependent of the study variables, even within the known distributions in the population. (Pyu-Martikainen and Rendtel 2008).
6 Moreover, inclusion of the analyses of the nonresponse surveys should make the paper much longer. The results will be reported in a book on nonresponse bias in the ESS (Stoop et al. 2010).
7 See the critical reflections on post-stratification later in this article.
8 The design weight corrects for deviations from the equal probability design (EPSEM). Using a design weight, ESS compensates for these discrepancies and apply the EPSEM principle for a weighted sample for individuals aged 15+ in each country.
9 As we will see, one of the main weaknesses of the PS method is the cross-cultural comparability of the education variable in both, population statistics and the survey. In the context of ESS, several researchers are trying to improve the measurement of education in view of cross-country comparison.
10 The *International Standard Classification of Education* (ISCED 1997) designed by the UNESCO (United Nations Educational, Scientific and Cultural Organisation), which is an instrument suitable for assembling, compiling and presenting statistics on education both within individual countries and internationally. It offers standard concepts, definitions and classifications.
11 In the preliminary report on weighting for the three previous ESS rounds, Vehovar had found that the distributions reported by the NC's are not always correct and optimal

presentations of the population. It was therefore proposed to change the source for the weightings and to prefer a "gold standard" instead of the population statistics. This proposal is however still in discussion and as an illustration of PS we are still using the weightings as proposed in the report of Round 2 (Vehovar 2007).

12 For details see Vehovar 2007: 338.

13 The design weights are all 1.0 when no clustering effects are in the sample and when all sampled (secondary) units have equal selection probabilities.

14 A complete overview is published in Vehovar (2007: 342-343).

15 One can however never exclude that some deviations are due to the original coding into the ISCED codes.

16 This section is largely based on Vehovar (2007: 344-353).

17 This can be evaluated by estimating the Mean Square Error (MSE) of the estimates which is the sum of the variance and the square of bias (Groves 1989).

18 The complete information is available in an extensive report at the ESS data website (Vehovar and Zupanič 2007). See the annex at http://mi.ris.org/uploadi/editor/1169212272appendix2.xls.

19 Of course, even the ASbias value of 2.077 for NWSPPOL does not truly mean statistical significance for the following reasons: it is only the average of the country *Sbiases*; and the properly inflated standard errors are not used. As was already mentions, the *Sbiases* are typically overestimated.

20 Functional equivalence of the measures for these two concepts was established in previous research (Davidov et al. 2008; Billiet and Meuleman 2008).

21 These countries were each compared with the German sample which has a much lower response rate than EE (51 percent). We do not find any difference between W1 and W2 models in the German sample.

22 This is the sample weighted by design weights for Germany.

23 This section is largely based on the report on refusal conversion in Round 2 of ESS (Beullens, Vandecasteele and Billiet 2007).

24 The attitudinal variables are all multiple indicator constructs. The measurement models of some of these were tested by structural equation modelling (with Lisrel 8.3). Others are evaluated by factor analysis. A report with information of the measurements is in the Annex of the original report (Beullens, Vandecasteele and Billiet 2007).

25 The variables were selected through a forward selection process. Backward and stepwise selection lead to the same outcome.

26 In more detailed analysis, it was found that females in the Netherlands were much more than male ready to cooperate in the survey after additional insistence of the interviewers.

27 It is possible to compute probabilities and changes in probabilities compared to a reference at basis of the odds ratio's (Allison 1999: 11-14).

28 A complete overview of the findings and analysis appeared in Cincinatto, Beullens and Billiet (2008). This part of the paper strongly relies on this report.

29 We can not guarantee that the observations are independent. The correlations may be high because of the interviewer's impressions.

30 These sampled persons will only be included when their absolute number exceeds 100. To be more specific, in Estonia, Finland, Hungary, the Netherlands, Poland and Portugal the non-contacts will not be accounted for since their respective absolute numbers are 85, 59, 0, 78, 20 and 78. Hungary is a special case. In fact, the non-contact number is

128 but due to inconsistencies in the contact forms they are not present in the dataset for Hungary and they can therefore not be accounted for (Billiet and Pleysier 2007: 51).

31 Remember that these are interviewer evaluations.

REFERENCES

Allison, Paul D. 1999. *Logistic Regression Using the SAS® System*. SAS Institute Inc.: Cary NY.

American Association for Public Opinion Research. 2000. *Standards and Best Practices*. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* Ann Arbor, Michigan: AAPOR.

Bethlehem, Jelke G., and Kersten Hubert M.P. 1985. 'On the Treatment of Nonresponse in Sample Surveys.' *Journal of Official Statistics* Vol. 1 (3): 287-300.

Bethlehem, Jelke G. 2002. Weighting Nonresponse Adjustments Based on Auxiliary Information. Pp. 275-288. In: Groves, Robert M., Donn A. Dillman, Eltinge, John L. and Roderick J.A Little (2002). *Survey Nonresponse*. New York: Willey.

Bethlehem, Jelke G., and Ineke Stoop. 2007. Online Panels – a Paradigm Theft? Pp. 113-131. In: Trotman, Mike et al. (Eds.). *The Challenges of a Changing World*. Southampton, UK: Association for Survey Computing.

Beullens, Koen, Leen Vandecasteele, and Jaak Billiet. 2007. *Refusal Conversion in the Second Round of the European Social Survey*. Deliverable n° 3 of Joint Research Actions 2 of ESSi. CeSO: Working paper Survey Methodology/2007-5, 34 pp.

Beullens, Koen, Jaak Billiet, and Geert Loosveldt. 2008. *Selection Strategies for Refusal Conversion of Four Countries in the European Social Survey, 3rd round*. CeSO: Working paper Survey Methodology, 21 pp.

Billiet, Jaak, and Stefaan Pleysier. 2007. *Response Based Quality Assessment in the ESS - Round 2. An update for 26 countries*. Research report of the Centre for Sociological Research. CeSO/SM/2006-5, 64 pp.

Billiet, Jaak, Michel Phillipens, Roy Fitzgerald, and Ineke Stoop. 2007. 'Estimation of Nonresponse Bias in the European Social Survey: Using Information from Reluctant Respondents.' *Journal Of Official Statistics* Vol. 23: 135-162.

Billiet, Jaak, and Bart Meuleman. 2008. Measuring Attitudes and Feelings Towards Discrimination in Cross-nation Research: Lessons Learned from the European Social Survey. In: *Proceedings of the 33nd CEIES Seminar Ethnic and Racial Discrimination on the Labour Market*, Malta 6-7 June 2007, 26 pp.

Brehm, John. 1993. *The Phantom Respondents. Opinion Surveys and Political Representation.* Ann Arbor: The University of Michigan Press.

Burton, John, Heather Laurie, and Peter Lynn. 2006. 'The Long Term Effectiveness of Refusal Conversion Procedures on Longitudinal Surveys.' *Journal of the Royal Statistical Society*. Series A, vol. 169: 459-478.

Cincinatto, Sebastiano, Koen Beullens, and Jaak Billiet. 2008. *Analysis of Observable Data in Call Records ESS – R2*. Deliverable n° 6 of Joint Research Actions 2 of ESSi. CeSO: Working paper of Survey Methodology, 43 pp.

Couper, Mick P., and Edith D. De Leeuw. 2003. Nonreponse in Cross-cultural and Cross-national Surveys. In: Janet A. Harkness (eds). *Cross-cultural Survey Methods*. Wiley: New Jersey.

Curtin, Richard, Stanley Presser, and Eleanor Singer. 2000. 'The Effects of Response Rate Changes on the Index of Consumer Sentiment.' *Public Opinion Quarterly* Vol. 64: 413-428.

Davidov, Eldad, Bart Meuleman, Jaak Billiet, and Peter Schmidt. 2008. 'Values and Support for Immigration: A Cross-Country Comparison.' *European Sociological Review* Vol. 24 (5): 17 pp.

Gelman, Andrew, and John B. Carlin. 2002. Post-stratification and Weighting Adjustments. Pp. 289-302. In: Groves, Robert M., Don A. Dillman, Eltinge, John L. and Roderick J.A. Little (2002). *Survey Nonresponse*. New York: Willey.

Groves, Robert M. 1989. *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

Groves, Robert M. 2006. 'Nonresponse Rates and Nonresponse Bias in Household Surveys.' *Public Opinion Quarterly* 70: 646-675.

Groves, Robert M. and Mick P. Couper. 1998. *Nonresponse in Household Surveys*. New York: John Wiley & Sons.

Heerwegh, Dirk, Koen Abts, and Geert Loosveldt. 2007. 'Minimizing Survey Refusal and Noncontact Rates: do our efforts pay off?' *Survey Research Methods* 1: 3-10.

Jowell, Roger. 1998. 'How comparative is comparative research?' *American Behavioral Scientist* 42: 168-177.

Jowell, Roger, Chris Roberts, Fitzgerald, Rory, and Gilian Eva. 2007. *Measuring Attitudes Cross-Nationally. Lessons from the European Social Survey.* London: Sage.

Kalton, Grahm, and Daniel Kasprzyk. 1986. 'The Treatment of Missing Survey Data.' *Survey Methodology* Vol.12 (1): 1-16.

Kish, Leslie. 1965. *Survey Sampling.* New York: Wiley.

Knot, Phillip S. 2006. 'Using Calibration Weightings to Adjust for Nonresponse and Coverage Errors.' *Survey Methodology* Vol. 32 (2): 133-142.

Laaksonen, Seppo, and Ray Chambers. 2006. 'Survey Estimation Under Informative Nonresponse with Follow Up.' *Journal of Official Statistics* Vol. 22: 81-95.

Lee, Sunghee. 2006. 'Propensity Score Adjustment as a Weighting Scheme for Volunteer Panel Web Surveys.' *Journal of Official Statistics* Vol. 22: 329-349.

Lin, I-Fen, and Nora Cate Schaeffer. 1995. 'Using Survey Participants to Estimate the Impact of Nonparticipation.' *Public Opinion Quarterly* Vol. 59: 236-258.

Little, Roderick, J.A., and Donald B. Rubin. 1987. *Statistical Analysis with Missing Data*. New York: John Wiley.

Little, Roderick J.A., and Sonya Vartivarian. 2005. 'Does Weighting for Nonrresponse Increase the Variance of Survey Means?' *Survey Methodology* Vol. 31 (2): 161-168.

Loosveldt, Geert, and Nathalie Sonck. 2008. 'An Evaluation of the Weighting Procedures for an Online Access Panel Survey.' *Survey Research Methods* Vol. 2: 93-105.

Lynn, Peter, Roeland Beerten, Johanna Laiho, and Jean Martin. 2002. 'Towards Standardisation of Survey Outcome Categories and Response Rate Calculations.' *Research in Official Statistics* Vol. 5: 6-84.

Meuleman, Bart, and Jaak Billiet. 2005. *Corrections for Non-response in the ESS Round 1: Weighting for Background Variables. A Simulation*. Research Report of the Center for Sociological Research (CeSO). DA/2005-49, 17 pp.

O'Shea, Ruth, Caroline Bryson, and Roger Jowell (2003) *Comparative Attitudinal Research in Europe*. European Social Survey Deliverable 1.

Potthoff, Richard, Kenneth Manton, and Max Woodbury. 1993. 'Correcting for Nonavailability Bias in Surveys by Weighting Based on Number of Callbacks.' *Journal of the American Statistical Association* Vol. 88: 1192-1207.

Pyu-Martikainen, Marjo, and Ulrich Rendtel. 2008. 'Assessing the Impact of Initial Nonresponse and Attrition in the Analysis of Unemployment Duration with Panel Surveys.' *Advances in Statistical Analysis* Vol. 92 (3): 297-318.

Rässler, Suzanne, Donald B. Rubin, and Nathaniel Schenker. 2008. Incomplete Data: Diagnosis, Imputation, and Estimation. Pp. 370-386 In: de Leeuw, Edith D., Joop Hox and Don A. Dillman (Eds.). *International Handbook of Survey Methodology*. New York: Lawrence Erlbaum ass.

Rizzo, Lou, Graham Kalton, and J. Michael Brick. 1996. 'A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse.' *Survey Methodology* Vol. 22 (1): 43-53.

Smith, Tom W. 2002. Developing Nonresponse Standards. In: Robert M. Groves, Don A. Dillman, John L. Eltinge, and Roderick J.A. Little (Eds.) *Survey Nonresponse*. New York: Wiley, pp. 27-40.

Stoop, Ineke. 2005. *The Hunt for the Last Respondent. Nonresponse in Sample Surveys*. The Hague: Social and Cultural Planning Office.

Stoop, Ineke, Jaak Billiet, Achim Koch, and Rory Fitzgerald. 2010. *Reducing Survey Nonresponse: Lessons learned from the European Social Survey*. New York: Willey (in press).

Sturgis, Patrick. 2004. 'Analysing Complex Survey Data: Clustering, Stratification and Weights.' *Social Research Update*. Issue 43, Autumn 2004.

Thomsen, Ib. 1973. 'A Note on the Efficiency of Weighting Subclass Means to Reduce the Effects of Nonresponse When Analyzing Survey Data.' *Statistisk Tidskrift* Vol. 11: 278-283.

Vehovar, Vasja. 2007. Non-response bias in the European Social Survey. Pp. 335-356. In: Loosveldt, Geert, Marc Swyngedouw, and Bart Cambré (Eds.). *Measuring Meaningful Data in Social Research*. Leuven: Acco.

Vehovar, Vasja, and Tina Zupanič. 2007. *Weighting in the ESS – Round.* University of Ljubljana, Faculty of Social Sciences. http://vasja.ris.org

Voogt, Robert. 2004. *"I'm not interested." Nonresponse bias, response bias and stimulus effects in election research*. PhD dissertation. University of Amsterdam.

Billiet Jacques was until end of 2007 full professor in social methodology at the Katholieke Universiteit Leuven, Belgium, and is now professor emeritus with research tasks. He is a member of the Central Co-ordination Team of the European Social Survey. His main research interest in methodology deals with validity assessment, interviewer and response effects, and the modelling of measurement error in social surveys. He is a member of the Flemish Royal Academy of Sciences & Arts (Belgium). Email address: jaak.billiet@soc.kuleuven.be

Hideko Matsuo obtained a PhD in demography (2003, Groningen University) after Japanese studies at Seisen Univesity and M.I.A at Columbia University. She is since 2007 in the position of a post-doctoral researcher at the Centre of Sociological Research, K.U. Leuven. As member of the Central Co-ordination Team of ESS she is mainly involved in the study response based quality of ESS datasets. Email address: hideko.matsuo@soc.kuleuven.be

Koen Beullens obtained a master degree in Sociology (2004) and Statistics (2007) at Katholieke Universiteit Leuven. He is currently employed as a research assistant at the Centre of Sociological Research (K.U. Leuven). His research interest is in survey methodology, particularly in the analysis of paradata and non-response. Email address: koen.beullenst@soc.kuleuven.be

Vasja Vehovar, PhD, is a full Professor of Statistics at the Faculty of Social Sciences, University of Ljubljana, Slovenia. He teaches courses on Sampling, Survey Methodology, and Information Society. He is responsible for development of the WebSM portal devoted to web survey methodology and was the coordinator of the corresponding EU Framework project. His research interests span from survey methodology to information society issues. Email address: vasja.vehovar@fdv.uni-lj.si

APPENDIX

Table A1 Estimates of the absolute average standardised bias of 45 items across countries

Abs ASbias		Item
0.555	STFGOV	How satisfied with the national government
0.666	PPLHLP	Most of the time people are helpful or mostly looking out for themselves
0.676	IPEQOPT	Important that people are treated equally and have equal opportunities
0.737	LRSCALE	Placement on left-right scale
0.793	IPLYLFR	Important to be loyal to friends and be devoted to close people
0.794	IGNRLAW	Occasionally alright to ignore the law and do what you want
0.801	PFMFDJB	Partner/family fed up with pressure of your job, how often
0.852	STFDEM	How satisfied with the way democracy works in the country
0.864	TRNDNJB	Would turn down a job with higher pay to stay with organisation working for
0.865	CTZHLPO	Citizens should spend some free time helping others
0.901	PYAVTXW	Someone paying cash without a receipt to avoid VAT or other tax, how wrong
0.988	TSTPBOH	Trust public officials deal honestly with you
1.002	HAPPY	How happy are you
1.003	BSNPRFT	Businesses only interested in profit, not improving service/quality
1.035	MNRSPHM	Men should take as much responsibility as women for home and children
1.062	STFECO	How satisfied with the present state of the economy in the country
1.066	STFLIFE	How satisfied with life as a whole
1.100	SMBTJOBA	Get a similar or better job with another employer
1.107	IPFRULE	Important to do what is told and follow rules
1.166	ESTSZ	Establishment size
1.179	STFHLTH	State of health services in the country nowadays
1.190	MUSDOCM	Misused/altered card/document to pretend eligible, last 5 years
1.205	TRSTEP	Trust in the European Parliament
1.255	STFEDU	State of education in the country nowadays
1.260	IMPENV	Important to care for nature and the environment
1.314	WKHTOT	Total hours normally worked per week in main job, overtime included
1.374	TVTOT	TV watching, total time on average weekday
1.444	WKHCT	Total contracted hours per week in main job, overtime excluded
1.474	RLGATND	How often attend religious services apart from special occasions
1.483	AESFDRK	Feeling of safety of walking alone in the local area after dark
1.505	WMCPWRK	Women should be prepared to cut down on paid work for the sake of the family
1.506	GINCDIF	Government should reduce differences in income levels
1.529	RLGDGR	How religious are you

Abs ASbias		Item
1.529	PRAY	How often pray apart from at religious services
1.536	PPLFAIR	Most people try to take advantage of you, or try to be fair
1.562	PPLTRST	Most people can be trusted or you can't be too careful
1.602	SCLMEET	How often socially meet with friends, relatives or colleagues
1.662	WRYWPRB	Worry about work problems when not working, how often
1.857	BRWMNY	Borrow money to make ends meet, difficult or easy
2.077	NWSPPOL	Newspaper reading, politics/current affairs on average weekday
2.082	IMWBCNT	Immigrants make the country a worse or better place to live
2.189	IMBGECO	Immigration bad or good for the country's economy
2.477	NWSPTOT	Newspaper reading, total time on average weekday
3.504	HHMMB	Number of people living regularly as member of the household
3.646	POLINTR	How interested in politics