# Differences in General Health of Internet Users and Non-users and Implications for the Use of Web Surveys

Rainer Schnell Sabrina Torregroza Marcel Noack

City University London & University of Duisburg-Essen

ESRC Research Methods Festival 5–7 July 2016, University of Bath

# Acknowledgement

Work reported here is part of the article:

*Rainer Schnell, Marcel Noack and Sabrina Torregroza (2016): Differences in General Health of Internet Users and Non-users and Implications for the Use of Web Surveys*

under review

# Introduction

- Web Surveys are increasing in fields of research like psychology, sociology, election studies and health.
- Web surveys suffer from higher rates of undercoverage and nonresponse than traditional modes.
- This is compensated by different weighting strategies like raking, post-stratification, GREG and propensity score weighting.
- These weighting procedures are based on the assumption that the missing data is either missing completely at random (MCAR) or can be explained using the observed data (missing at random, MAR).

# Basic problem of the study

- If the target variable is related to the cause of missing data and can not be explained by observed data, resulting estimates will be biased.
- Simulations and empirical studies suggest that strong correlation between target variables and response mechanism will cause biased estimates despite weighting.
- Surveys on health related topics will suffer from nonresponse and undercoverage caused by health issues.
- Therefore, non-sampling errors of web surveys might be an example of MNAR which cannot be corrected by weighting.

# Undercoverage in Web Surveys

- Undercoverage is usually considered as refering to Internet access by the target population.
- Proportion of households with internet access increases rapidly, but still differences between industrialized countries.
- Bias by noncoverage depends on proportion of those excluded from Internet access ($\frac{N_{NI}}{N}$) and on differences regarding the target Variable $Y$ between people with ($I$) and without Internet ($NI$):

$$B(\bar{y}_I) = \frac{N_{NI}}{N}(\overline{Y}_I - \overline{Y}_{NI})$$

- Data for USA and Europe suggest that those with Internet access differ from those without access in regard to education, income, age, gender and ethnicity.

# Nonresponse in Web Surveys

- Surey Response Rates are declining in all industrialized western countries.
- Response Rates for web surveys are even lower than those of other collection methods.
- Bias due to nonresponse will be higher with increasing proportions of of nonrespondents and increasing correlations between target variable and cause of nonresponse.

# Nonresponse in Web Surveys

If

- $\overline{p}$ is the average probability of participation in target population,
- $S_{pY}$ the covariance between the values of the variable of interest and the probability of participation,
- $R_{pY}$ the corresponding correlation and
- $S_p$, $S_Y$ the variances of the probability of participation and the target variable

the resulting bias can be estimated as

$$B(\overline{y}_R) = \tilde{Y} - \overline{Y} = \frac{S_{pY}}{\overline{p}} = \frac{R_{pY} S_p S_Y}{\overline{p}}.$$

# Bias Reduction by Weighting Procedures

- Weighting techniques are used to adjust for undercoverage and nonresponse.
- The basic approach is cell weighting (post-stratification).
- In academic research and official statistics the calibration approach is widely used.
- Most often, propensity weights are used for web surveys.
- Combinations of the approaches have been published.
- However, all approaches assume MCAR or MAR.

# Missing Data Mechanisms

- Let $\boldsymbol{Y}_i = (Y_{i1}, Y_{i1}, \ldots, Y_{ip})^T$ be a set of $p$ variables of interest for respondents $i = 1, \ldots, n$.
- For each respondent $\boldsymbol{Y}_i$ can be partitioned in two parts, a part of observed data ($\boldsymbol{Y}_{i,obs}$) and a part that is missing ($\boldsymbol{Y}_{i,mis}$).
- Furthermore, $\boldsymbol{R}_i = (R_{i1}, R_{i1}, \ldots, R_{ip})^T$ is a set of binary variables, indicating if the data on variable $p$ for case $i$ is missing ($\boldsymbol{R}_{ip} = 0$) or observed ($\boldsymbol{R}_{ip} = 1$).
- Then a missing data mechanism is defined by the conditional distribution of $\boldsymbol{R}_i$ given $\boldsymbol{Y}_i$.

# MCAR, MAR, MNAR

**MCAR** Missing of $\boldsymbol{Y}_i$ does not depend on $\boldsymbol{Y}_i$

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) = Pr(\boldsymbol{R}_i).$$

**MAR** Missingness is independent from unobserved data $\boldsymbol{Y}_{i,mis}$, but dependent on the observed data $\boldsymbol{Y}_{i,obs}$:

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) = Pr(\boldsymbol{R}_i|\boldsymbol{Y}_{i,obs}).$$

**MNAR** The probability of a value being missing depends on the underlying missing value itself even given the observed data:

$$Pr(\boldsymbol{R}_i|\boldsymbol{Y}_i) \neq Pr(\boldsymbol{R}_i|\boldsymbol{Y}_{i,obs})$$

# Internet and Health

- Empirical results suggest that a person's health might affect undercoverage and nonresponse in all survey modes.
- There is limited evidence that this is also true in Web Surveys, for example Schonlau/vanKapteyn/Couper (2009, SMR) based on SHARE (50+).
- Data from same study confirmed group differences in regard to health variables even after controlling for socio-demographics.
- Most studies are based on subgroups of the population and one country.
- We want to show the effect for other countries using general population surveys.
- Furthermore, this effect has not been framed as MNAR before.

# Data

- The ESS round 5 conducted in 2010 contains information on subjective health and Internet use.
- ESS incorporates 27 countries and sampled about 52.000 persons.
- For including United States Data from the Behavioral Risk Factor Surveillance System (BRFSS, 2013) is used which contains 485.000 respondents.

# Measurement of general health

| | |
|---|---|
| ESS | *How is your health in general? Would you say it is . . . (1) very good, (2) good, (3) fair, (4) bad, or, (5) very bad?* |
| BRFSS | *Would you say that in general your health is: (1) Excellent, (2) Very good, (3) Good, (4) Fair, (5) Poor.* |

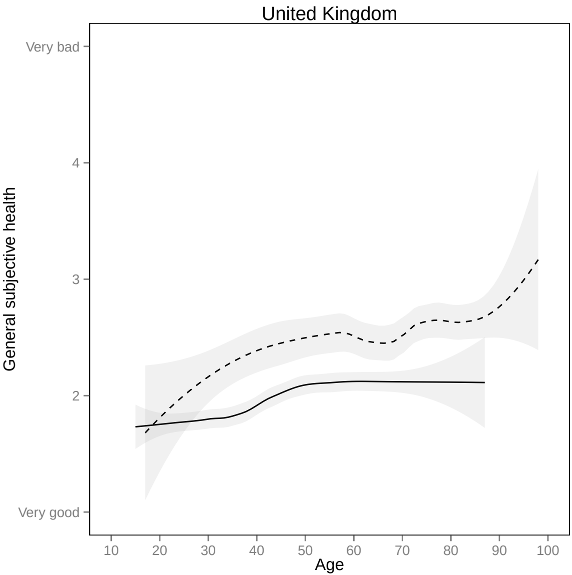
# Measurement Internet use

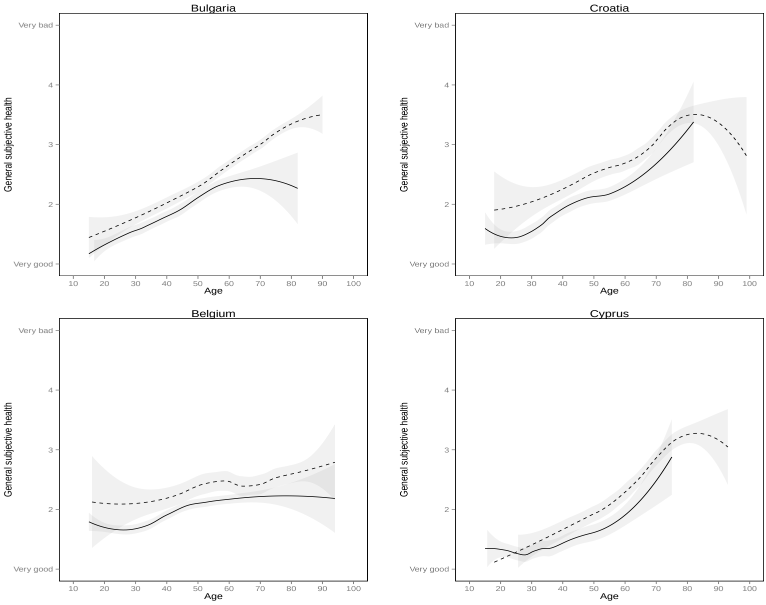
| | |
|---|---|
| ESS | *Now, using this card, how often do you use the Internet, the World Wide Web or e-mail – whether at home or at work – for your personal use? (0) No access at home or work, (1) Never use, (2) Less than once a month, (3) Once a month, (4) Several times a month, (5) Once a week, (6) Several times a week, (7) Every day.* |
| BRFSS | *Have you used the Internet in the past 30 days? (1) Yes, (2) No.* |

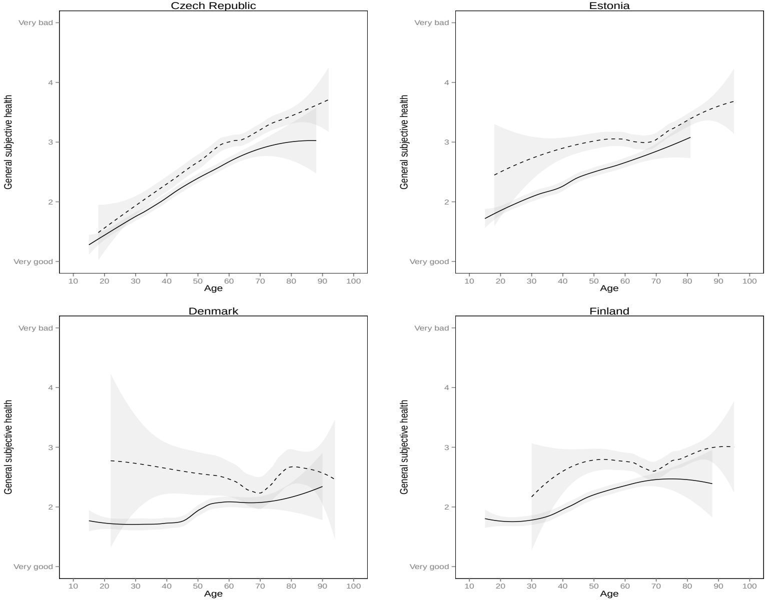
# Method

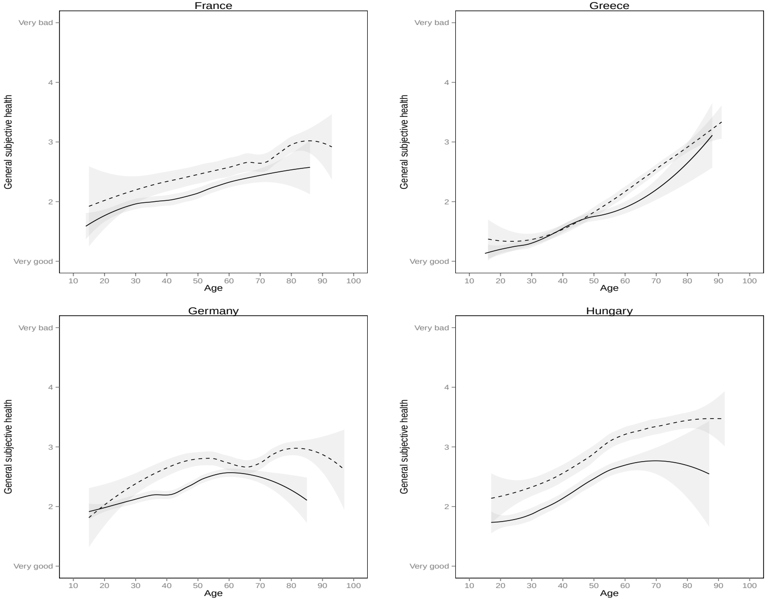
- Internet use is strongly related to age in many countries.
- Age is related to health.
- Therefore, age has to be controlled for in the analysis.
- To explore expected nonlinear relationships, separate nonparametric regressions (Loess, $bw = 0.8$) for each country are used.
- The solid line in each plot is the regression estimate for Internet users and the dashed line the regression estimate for non-users.
- The shaded areas are 95% confidence bands.

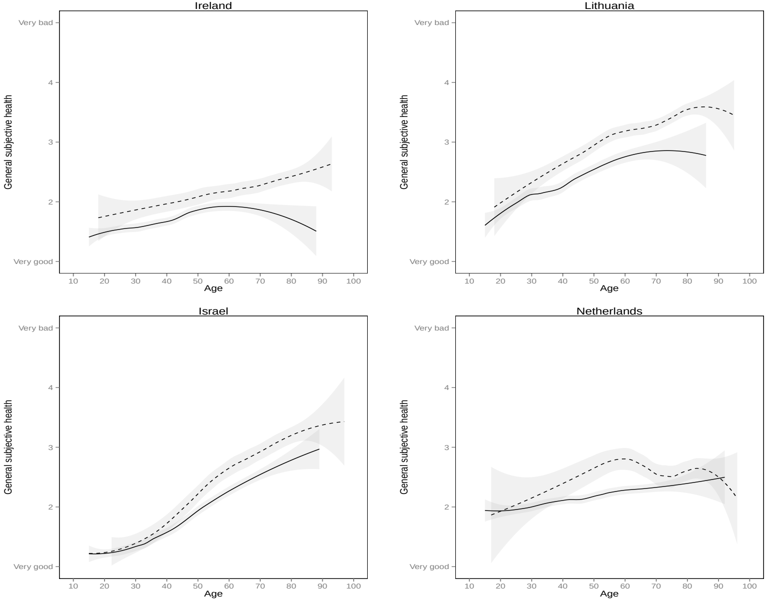
# Prototype Example Result: UK

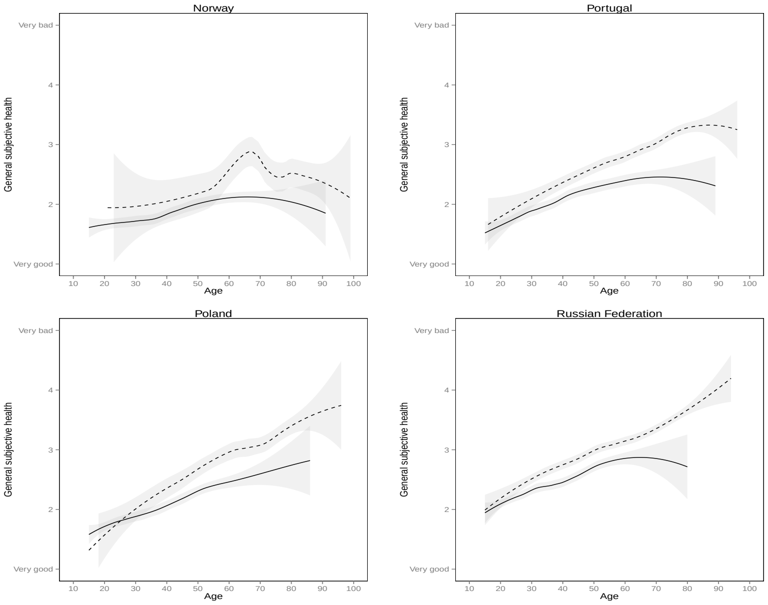



Bulgaria
Croatia
Belgium
Cyprus
General subjective health
Very bad
4
3
2
Very good
Age
10 20 30 40 50 60 70 80 90 100

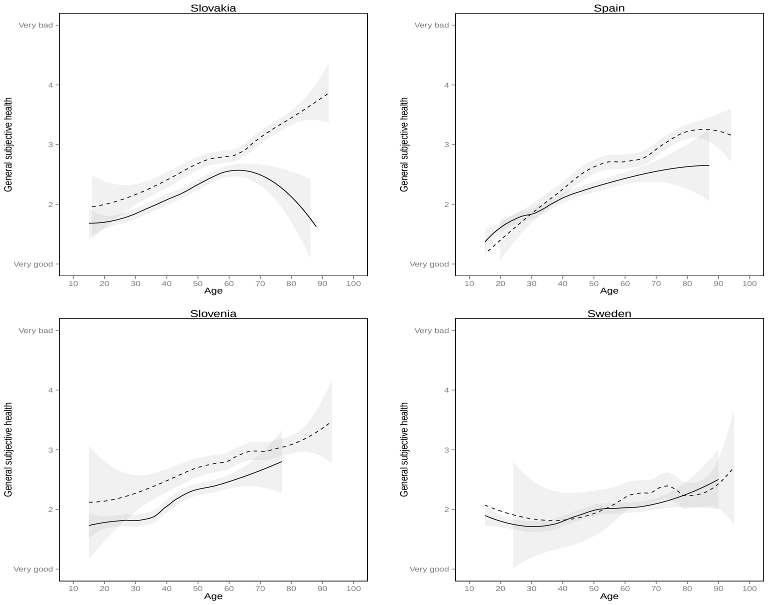

Czech Republic
Estonia
Denmark
Finland
Very bad
4
3
2
Very good
General subjective health
10 20 30 40 50 60 70 80 90 100
Age

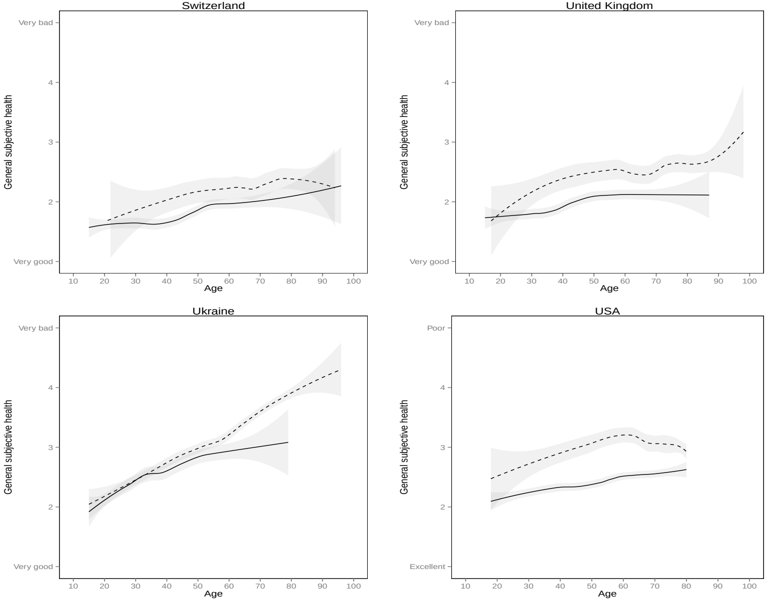

France
Greece
Germany
Hungary
General subjective health
Very bad
4
3
2
Very good
Age
10
20
30
40
50
60
70
80
90
100


Ireland
Lithuania
Israel
Netherlands
General subjective health
Very bad
4
3
2
Very good
10
20
30
40
50
60
70
80
90
100
Age



Norway
Portugal
Poland
Russian Federation
Very bad
4
3
2
Very good
General subjective health
10 20 30 40 50 60 70 80 90 100
Age


Slovakia
Spain
Slovenia
Sweden
General subjective health
Very bad
4
3
2
Very good
10 20 30 40 50 60 70 80 90 100
Age


Switzerland
United Kingdom
Ukraine
USA
General subjective health
Age
Very bad
4
3
2
Very good
Poor
Excellent
10 20 30 40 50 60 70 80 90 100

# Summary of plots

- In all 28 countries, the subjective health reported by the respondents is worse for non-users of the Internet.
- In 20 of the 28 countries the worse health status of non-internet users can be observed in all age groups.
- Increasing differences in reported health between Internet users and non-users with increasing age can be observed for the majority of countries.
- Limitations:
  - small number of observations in subgroups
  - possible different nonresponse bias in different countries
  - statistical problems of bootstrapping confidence bands for nonparametric regressions.

Table: General subjective health (ESS, round 5), Linear Multilevel Model
Level 1: Individuals, Level 2: Countries
Model (1): Not weighted
Model (2): Design weight at L-1, population weight at L-2
Model (3): Poststratification at L-1, population weight at L-2

| | Model (1) | | Model (2) | | Model (3) | |
|---|---|---|---|---|---|---|
| Age | 0.339** | (0.009) | 0.285** | (0.021) | 0.281** | (0.024) |
| $\text{Age}^2$ | −0.057** | (0.005) | −0.069** | (0.009) | −0.066** | (0.007) |
| $\text{Age}^3$ | −0.013** | (0.004) | −0.001 | (0.007) | 0.004 | (0.006) |
| Netuse | −0.065** | (0.006) | −0.065** | (0.015) | −0.066** | (0.014) |
| Age × Netuse | −0.079** | (0.006) | −0.071** | (0.009) | −0.072** | (0.008) |
| Years of Education | −0.020** | (0.001) | −0.014** | (0.003) | −0.015** | (0.002) |
| HH-income (Decile) | −0.047** | (0.002) | −0.044** | (0.003) | −0.043** | (0.004) |
| Female | 0.049** | (0.008) | 0.074** | (0.023) | 0.079** | (0.021) |
| Constant | 2.764** | (0.058) | 2.829** | (0.072) | 2.824** | (0.077) |
| $\sigma_u^2$ | .078 | | .083 | | .079 | |
| $\sigma_e^2$ | .638 | | .606 | | .598 | |
| $\rho$ | .108 | | .120 | | .117 | |
| $R^2_{\text{Level 1}}$ | .239 | | .210 | | .228 | |
| $R^2_{\text{Level 2}}$ | .170 | | .052 | | .032 | |
| $R^2_{\text{Overall}}$ | .232 | | .194 | | .209 | |
| $n$ | 39305 | | 39305 | | 39305 | |

Standard errors in parentheses
Age and netuse $z$-transformed
$^{*}\ p < 0.05$, $^{**}\ p < 0.01$

# Effect size

- More import than the problem of significant differences between groups is the size of the differences.
- We used Cohen's D, adjusted for unequal group sizes:

$$D = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}} \sqrt{\frac{\frac{n_1+n_2}{2}}{\frac{2n_1n_2}{n_1+n_2}}}.$$

According to Ellis (2010, Essential Guide to Effect Sizes):

- If $D \geq 0.5$: medium effect
- If $D \geq 0.8$: large effect

# Cohen's D for general subjective health

| Country | D | Country | D |
|---|---|---|---|
| Estonia | 1.255 | Russian Federation | .907 |
| Lithuania | 1.177 | Israel | .875 |
| Poland | 1.125 | Netherlands | .858 |
| Norway | 1.079 | USA | .843 |
| Hungary | 1.046 | Spain | .828 |
| Czech Republic | 1.045 | Belgium | .814 |
| Portugal | 1.037 | Greece | .780 |
| Finland | 1.027 | Ukraine | .779 |
| Croatia | .991 | Switzerland | .718 |
| Slovakia | .964 | France | .716 |
| Denmark | .960 | Germany | .684 |
| Bulgaria | .957 | United Kingdom | .675 |
| Cyprus | .940 | Ireland | .668 |
| Slovenia | .933 | Sweden | .626 |
| Mean D | .903 | Std. Dev. D | .164 |

groups by internet usage (yes/no)

# Summary for Effect Size

- Ranges roughly between 0.6 for Sweden and 1.2 for Estonia.
- Average D is 0.9.
- Large effect sizes are observed for 20 of 28 countries. The remaining countries show medium effect sizes.
- The supposed difference in subjective health between internet users and non-users seems to be serious.

# Impact on Estimates

- To quantify the possible impact of the observed effects, we used the Standardized Bias (Collins/Schafer/Kam 2001, Psychological Methods):

$$SB = 100 * \frac{\bar{x} - \mu}{\hat{\sigma}_{\bar{x}}}$$

$\bar{x}$ is the sample estimate

$\mu$ is the population value

$\hat{\sigma}$ is the estimated standard error

- SB is therefore the difference between the estimate and the population value in standard error units.
- Values of SB larger than 40 are regarded as of practical importance.

# Standardized Bias (internet users vs. full sample) for point estimate "General Health" without correction for coverage error

| country | standardized bias | country | standardized bias |
|---|---|---|---|
| USA | -4501 | Lithuania | -1004 |
| Greece | -1752 | Slovenia | -972 |
| Bulgaria | -1560 | Ukraine | -893 |
| Russian Federation | -1458 | Germany | -777 |
| Portugal | -1456 | Finland | -723 |
| Poland | -1451 | Belgium | -715 |
| Hungary | -1444 | France | -620 |
| Czech Republic | -1239 | Ireland | -605 |
| Croatia | -1210 | Switzerland | -510 |
| Cyprus | -1108 | United Kingdom | -499 |
| Spain | -1083 | Denmark | -353 |
| Israel | -1054 | Norway | -316 |
| Estonia | -1054 | Netherlands | -257 |
| Slovakia | -1037 | Sweden | -195 |

# Standardized Bias (internet users vs. full sample) for point estimate "General Health" after correction for coverage error by calibration

| country | standardized bias | country | standardized bias |
|---|---|---|---|
| Bulgaria | -784 | Ireland | -292 |
| Russian Federation | -727 | Croatia | -268 |
| Greece | -492 | Slovenia | -267 |
| Lithuania | -441 | Czech Republic | -241 |
| Ukraine | -440 | Israel | -196 |
| Slovakia | -421 | Switzerland | -185 |
| Hungaria | -407 | Austria | -172 |
| Poland | -398 | Norway | -161 |
| United Kingdom | -372 | Denmark | -155 |
| Cyprus | -360 | France | -145 |
| Estonia | -351 | Finland | -115 |
| Belgium | -306 | Netherlands | -96 |
| Spain | -302 | Sweden | -33 |
| Germany | -302 | | |

# Summary for Standardized Bias

- Before Calibration: Smallest SB for Sweden and Netherlands, but even here SBs of $-195$ and $-257$ indicate differences with practical importance.
- After Calibration: Still smallest SB for Sweden and Netherlands. With the exception of Sweden, all SB still indicate practical differerences.
- In general, the standardized bias seem to decrease with increasing per capita gross domestic product.

# Overall Summary

- Differences in general health between internet users and non-users in all countries examined here are
  - statistical significant,
  - have medium to large effect sizes,
  - seem to be of practical importance and
  - persist even after calibration for Age, Gender, Foreign Birth, Urbanity, ISCED and HH income.

# Discussion

- Health estimates of Internet users and Internet non-users show that people who are less healthy tend to use the Internet not as frequently as healthy people.
- This result was observed in all ESS-countries (27) and in the USA.
- Differences remained (mostly) after controlling for age.
- If the internet users in the ESS and BFSS samples can be seen as a random sample of the frame population of a web survey, the observed health differences are relevant for web surveys in general. will cause biased estimates of health related variables.

# Consequences

- The most alarming finding is the fact that weighting (calibration) does not eliminate the health differences.
- This may be due to unsuited weighted variables, but we used the standard variables usually available in survey research.
- Therefore, we consider the underlying missing data mechanism as an example of MNAR.
- If this holds true, the usual weighting techniques could not be used to correct for this bias, because the fundamental MAR assumption shared by all this techniques would be violated.
- This would be a serious limitation of web surveys concerning health related variables.