GESIS Leibniz Institute for the Social Sciences

Representation Bias in Probability and Non-Probability Surveys -

Theoretical Considerations and Practical Applications Using the sampcompR R-Package

Leibniz Association

Björn Rohr and Barbara Felderer 2025

ESRA 2025, July 14th Rohr & Felderer

GESIS Leibniz Institute for the Social Sciences

Probability and Nonprobability Surveys

Definition Probability Surveys

A probability survey is a survey for which every member of the **target population** has a known and positive (i.e., non-zero) **probability of being sampled** for the survey.

Definition Nonprobability Surveys

A nonprobability survey is a survey for which the **probability of participating** is (a) **unknown**, or (b) ***zero***, for at least some members of the **target population.**

Example Nonprobability Surveys

- Online Access Panel Survey
- Student Survey
- Social Media Ads-Recruited Survey
- Snowball Sampling

Example Nonprobability Surveys

Online Access Panel Survey:

- A survey based on a (commercial) online access panel
- One of the most common types of nonprobability survey
- Often conducted as quota sample of panel members
- Can be easy and cheap, but very untransparent regarding participant selection

Example Nonprobability Surveys

Student Survey:

- University students are interviewed voluntarily (sometimes in exchange for credit points)
- Commonly used method in psychology
- Often in combination with experiments
  - Experiments guarantee internal validity
  - But do not guarantee external validity

Example Nonprobability Surveys

Social Media Ads-Recruited Survey:

- A link to the survey is distributed via social media advertisement (e.g., Facebook-Ads, Instagram-Ads,…)
- Can be a very cheap method to conduct a nonprobability survey.
- Can be used to survey hard-to-reach populations
- Researchers have some control on survey recruitment
- Requires knowledge and monitoring-effort from researchers.

Example Nonprobability Surveys

Snowball Sampling:

- Initial survey participants are recruited (probability or non-probability) and asked to distribute the survey (or survey link) to others in the target population
- This is useful for hard-to-reach populations
- A similar method (Respondent-Driven Sampling) can even lead to probability surveys if certain assumptions are met

GESIS Leibniz Institute for the Social Sciences

Exercise

Think about the definition of nonprobability surveys again:

Why are Online Access Panel surveys nonprobability surveys?

What about student surveys? Do student surveys exist that are probability surveys?

GESIS Leibniz Institute for the Social Sciences

Definition Nonprobability Surveys

A nonprobability survey is a survey for which the **probability of participating** is (a) **unknown**, or (b) ***zero***, for at least some members of the **target population.**

Solution:

Why are Online Access Panel surveys nonprobability surveys?

Solution:

Why are Online Access Panel surveys nonprobability surveys?

- The panels are typically not randomly selected, and we do not know how the selection process works.
- Further, they most often exclude non-Internet users.
- Finally, even the recruitment for the specific survey is often not based on a random sample of panel members but done with quota sampling.

Solution:

What about student surveys? Are there student surveys that are probability surveys?

Solution:

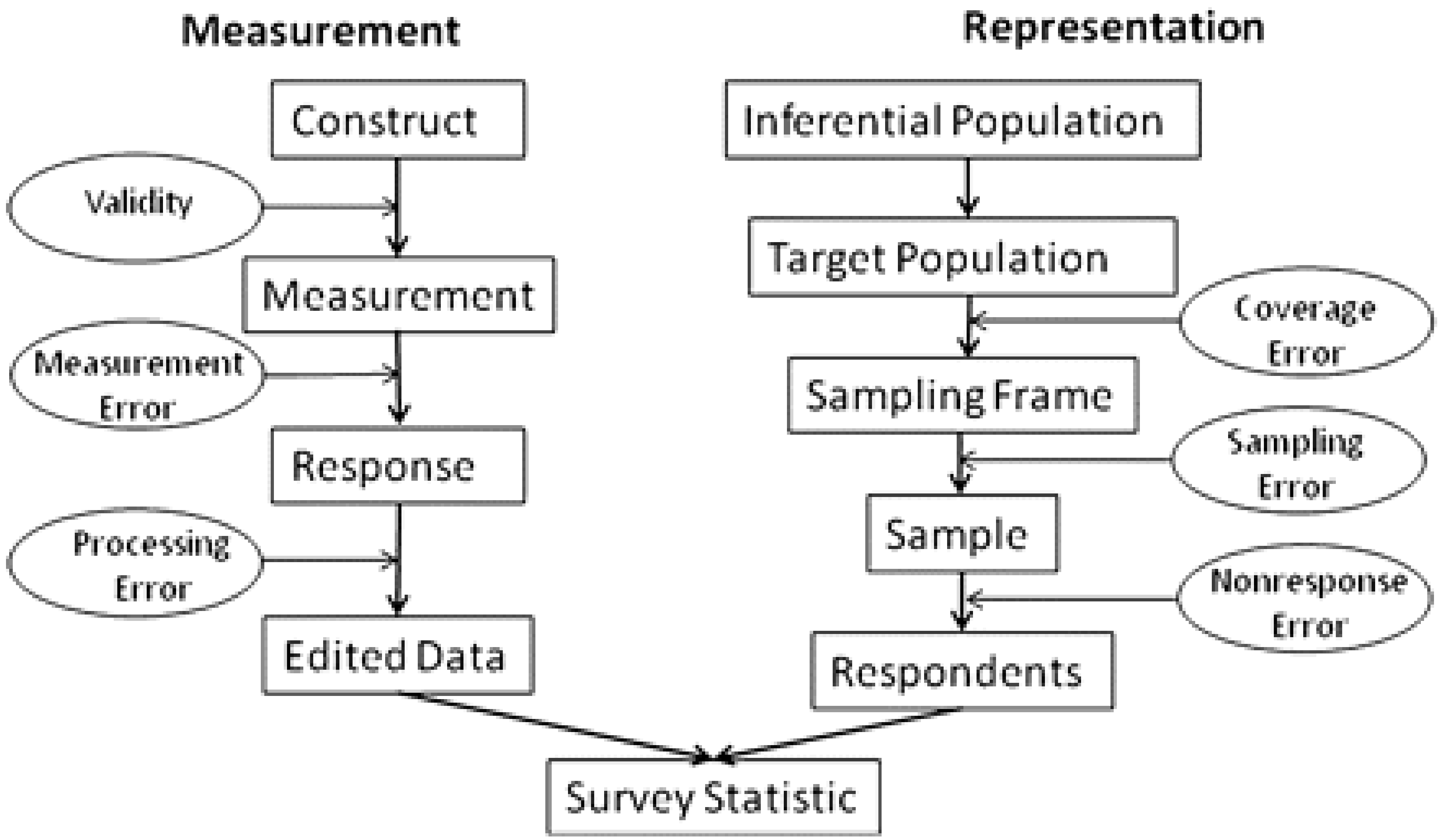
What about student surveys? Are there student surveys that are probability surveys?

- Typically, students are not randomly selected but self-selected, for example, when they conduct the surveys for credit points in some specific classes.
- Additionally, the psychological experiments are often intended to be generalized to a population bigger than students (e.g., the whole country).
- Student surveys might be a probability survey if the study population is students of a specific university, and they are invited with known and non-zero probabilities for all students.

Selective Survey Participation

Total Survey Error Framework

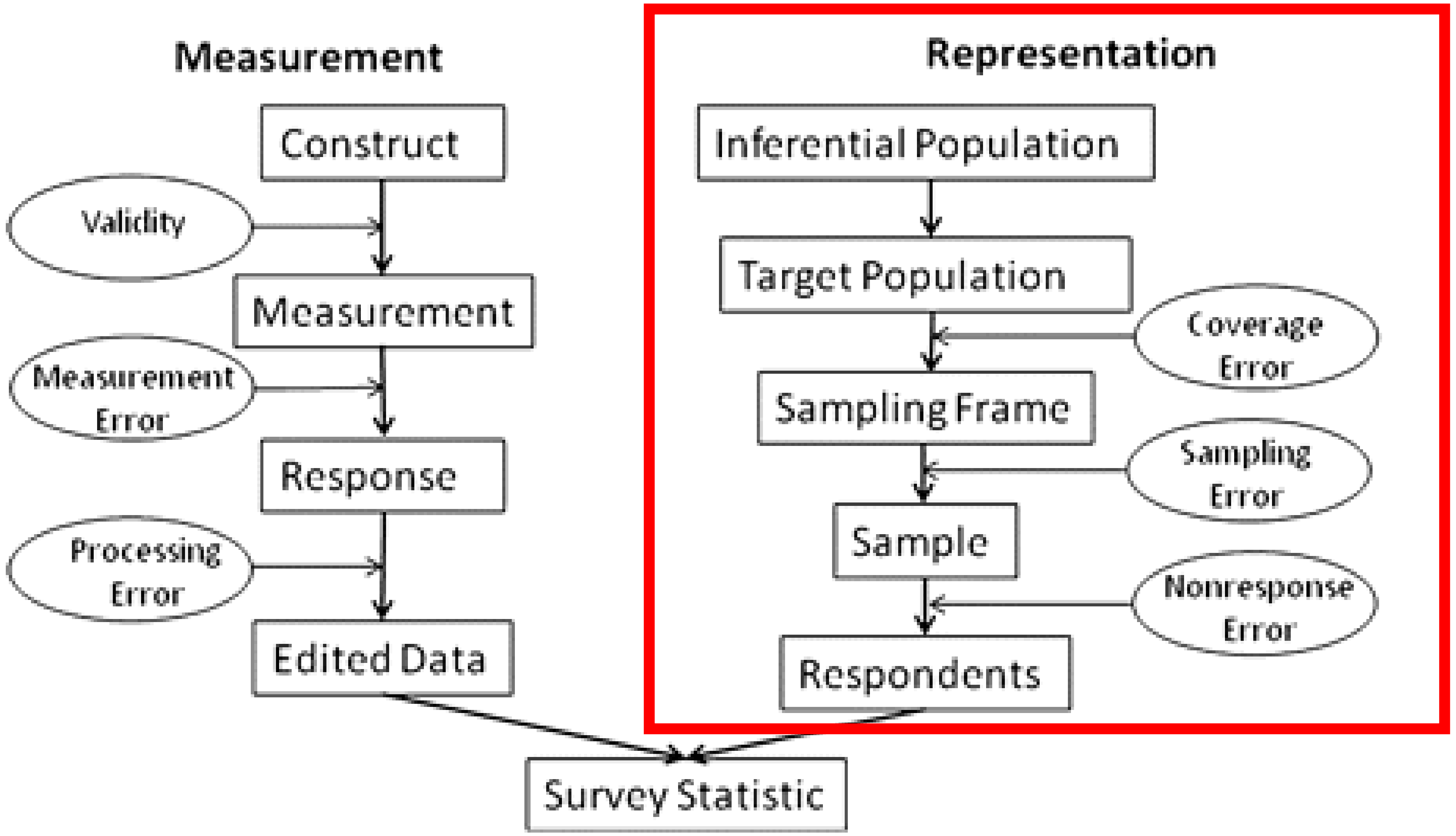
GESIS Leibniz Institute
for the Social Sciences



Total Survey Error Framework (Groves and Lyberg 2010) Reprinted from “Total Survey Error: Past, Present, and Future” by Author R. M. Groves & L. Lyberg, 2010, Public Opinion Quarterly, 74(5), 856. Copyright [2010] by the Oxford University Press. Reprinted with permission.

Total Survey Error Framework

GESIS Leibniz Institute for the Social Sciences

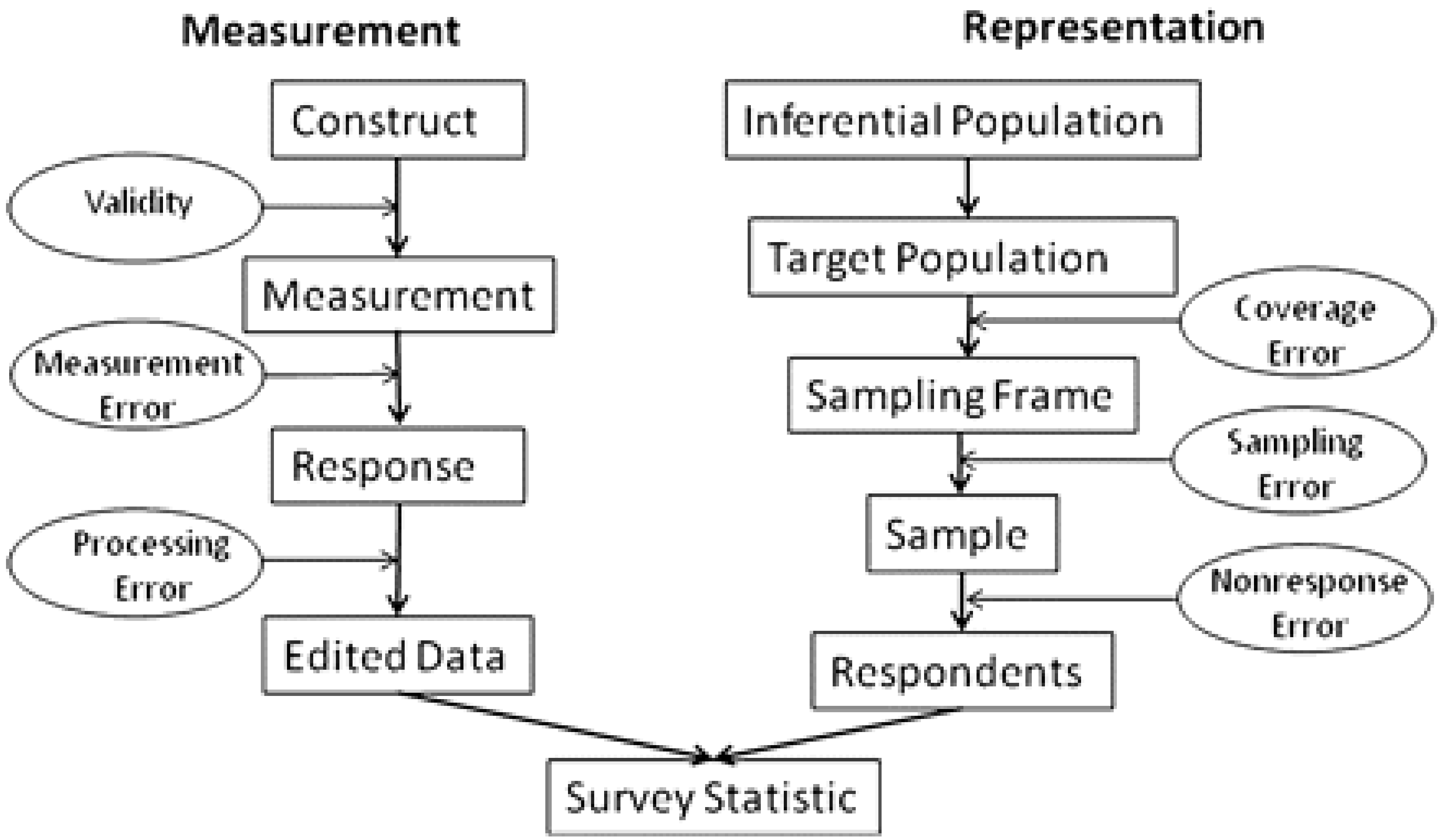


Total Survey Error Framework (Groves and Lyberg 2010) Reprinted from "Total Survey Error: Past, Present, and Future" by Author R. M. Groves & L. Lyberg, 2010, Public Opinion Quarterly, 74(5), 856. Copyright [2010] by the Oxford University Press. Reprinted with permission.

Total Survey Error Framework

GESIS Leibniz Institute for the Social Sciences

Probability Surveys

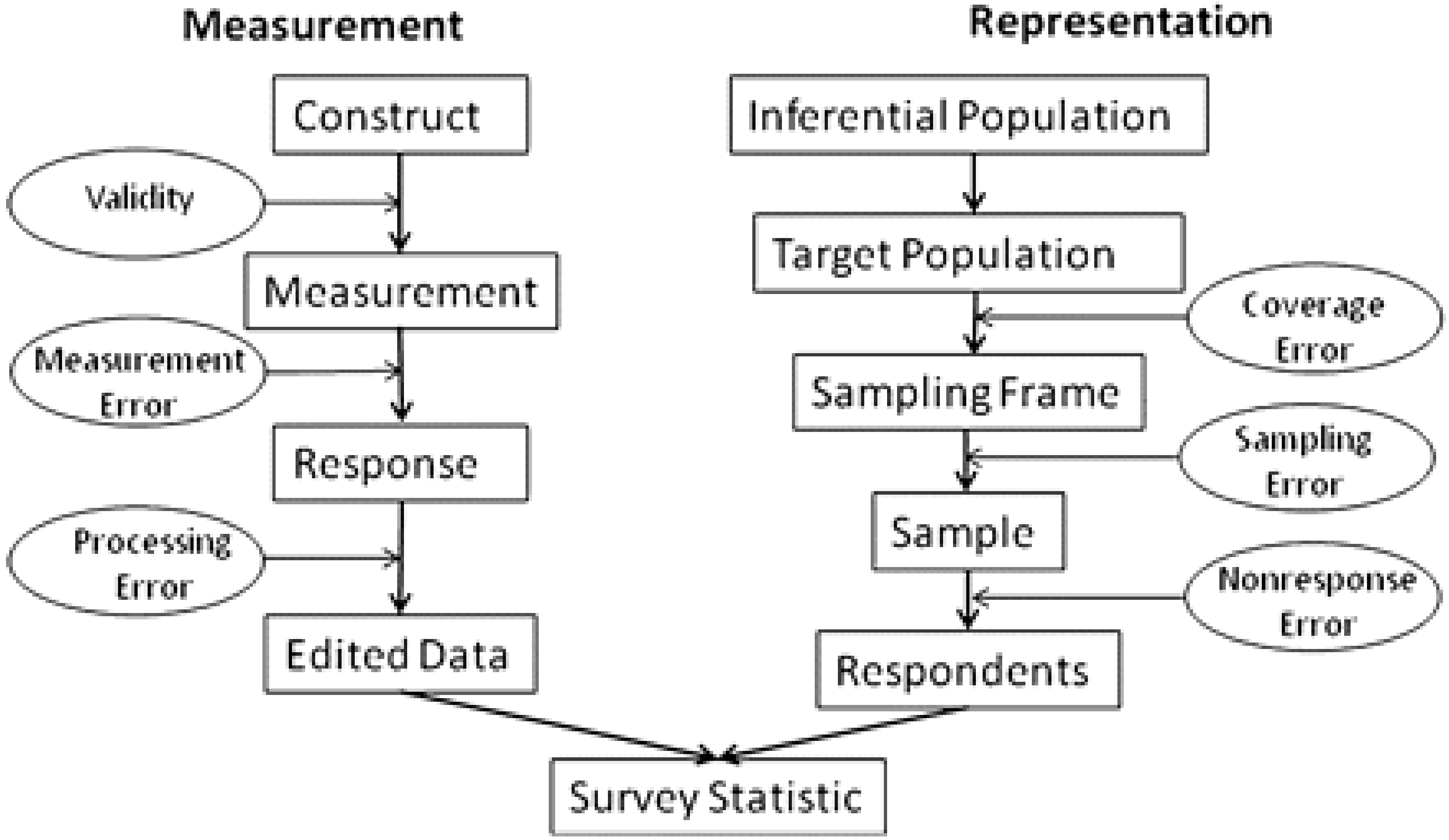


Total Survey Error Framework (Groves and Lyberg 2010) Reprinted from "Total Survey Error: Past, Present, and Future" by Author R. M. Groves & L. Lyberg, 2010, Public Opinion Quarterly, 74(5), 856. Copyright [2010] by the Oxford University Press. Reprinted with permission.

Total Survey Error Framework

GESIS Leibniz Institute for the Social Sciences

Probability Surveys



Coverage Error Example

Target Population:
All German citizens

Sampling Frame:
A list of German landline numbers

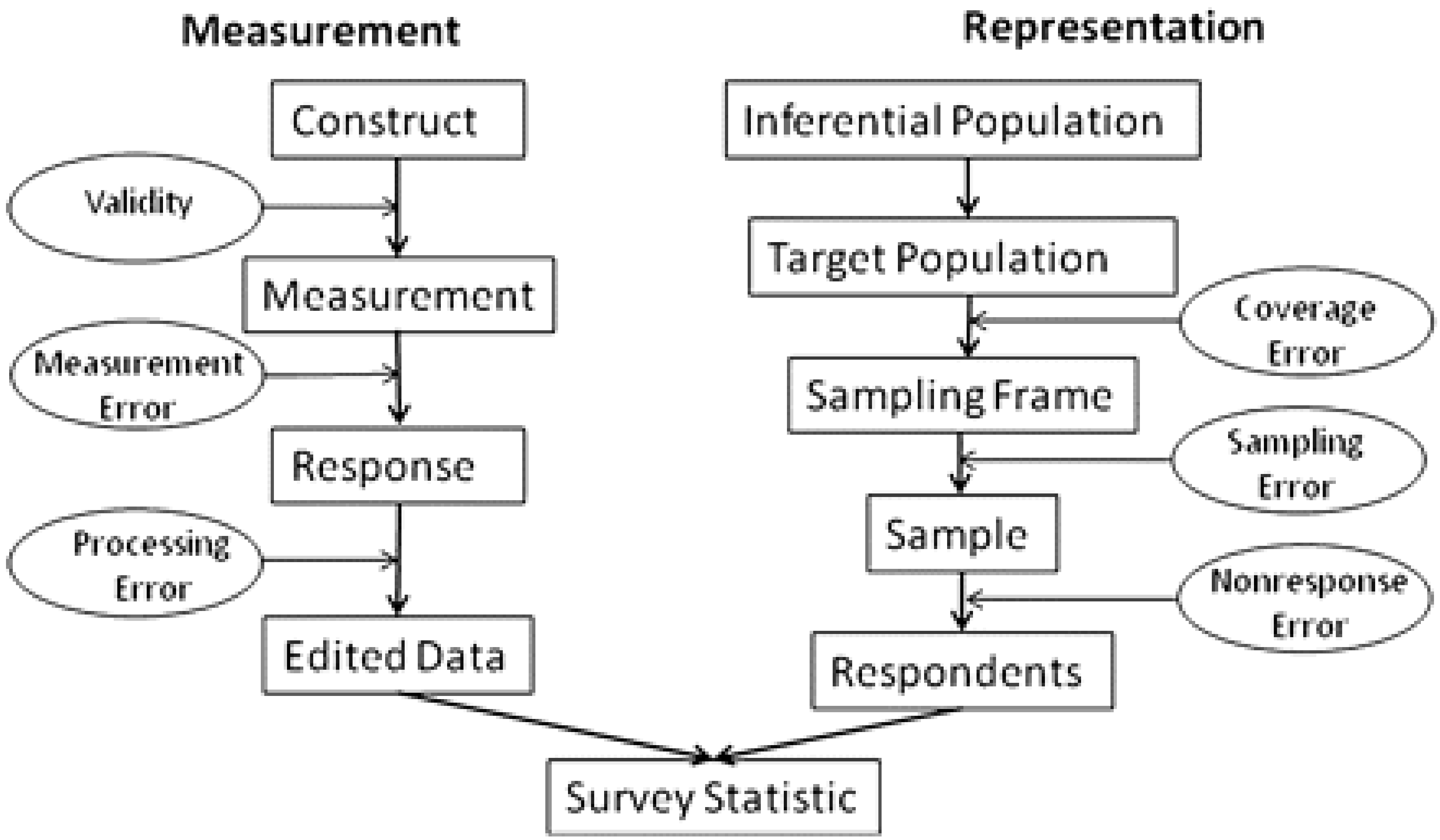
Problem: No Cell Phone Numbers
Many German households do not have a landline connection

Total Survey Error Framework (Groves and Lyberg 2010) Reprinted from "Total Survey Error: Past, Present, and Future" by Author R. M. Groves & L. Lyberg, 2010, Public Opinion Quarterly, 74(5), 856. Copyright [2010] by the Oxford University Press. Reprinted with permission.

Total Survey Error Framework

GESIS Leibniz Institute for the Social Sciences

Probability Surveys



Total Survey Error Framework (Groves and Lyberg 2010) Reprinted from "Total Survey Error: Past, Present, and Future" by Author R. M. Groves & L. Lyberg, 2010, Public Opinion Quarterly, 74(5), 856. Copyright [2010] by the Oxford University Press. Reprinted with permission.

Nonresponse Error Example

Target Population:
All German citizens

Sample:
Randomly selected from population register

Survey Mode:
Self-administered (online or paper)

Problem:
Education bias, more paticipants interested in the topic

Exercise: Coverage and Nonresponse Bias in Non-Probability Surveys

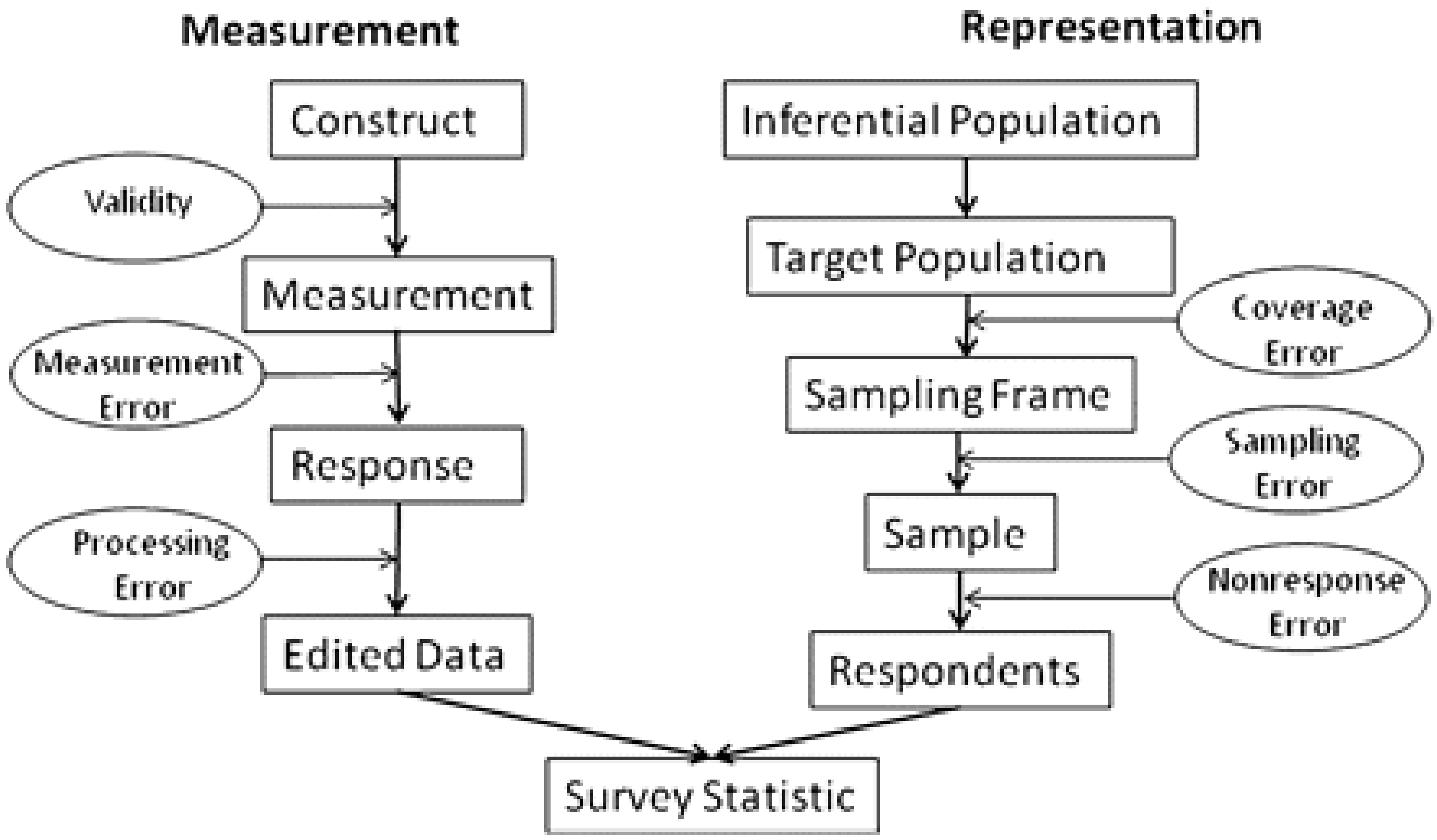
Imagine a commercial online access panel survey.

- Little is known about how the panel was generated.
- The survey was drawn as a quota survey from the panel.

Can you imagine an example of coverage and nonresponse bias?

Total Survey Error Framework

Probability Surveys | **Nonprobability Surveys**



Selection Error

Total Survey Error Framework (Groves and Lyberg 2010) Reprinted from “Total Survey Error: Past, Present, and Future” by Author R. M. Groves & L. Lyberg, 2010, Public Opinion Quarterly, 74(5), 856. Copyright [2010] by the Oxford University Press. Reprinted with permission.

GESIS Leibniz Institute for the Social Sciences

Empirical Examples

Leibniz Association

Empirical Example:
Comparison of Probability and Nonprobability Surveys

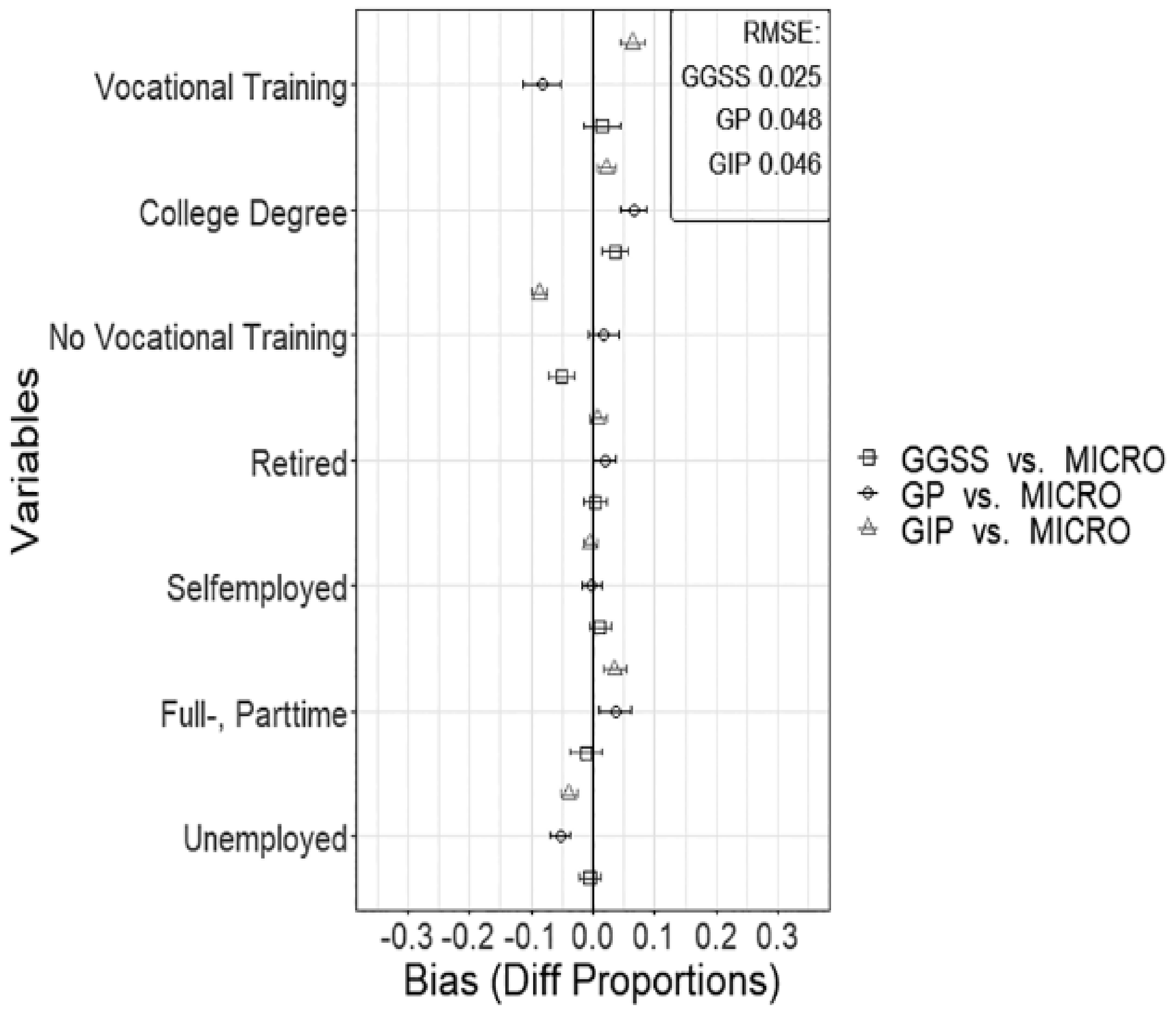
GESIS Leibniz Institute for the Social Sciences

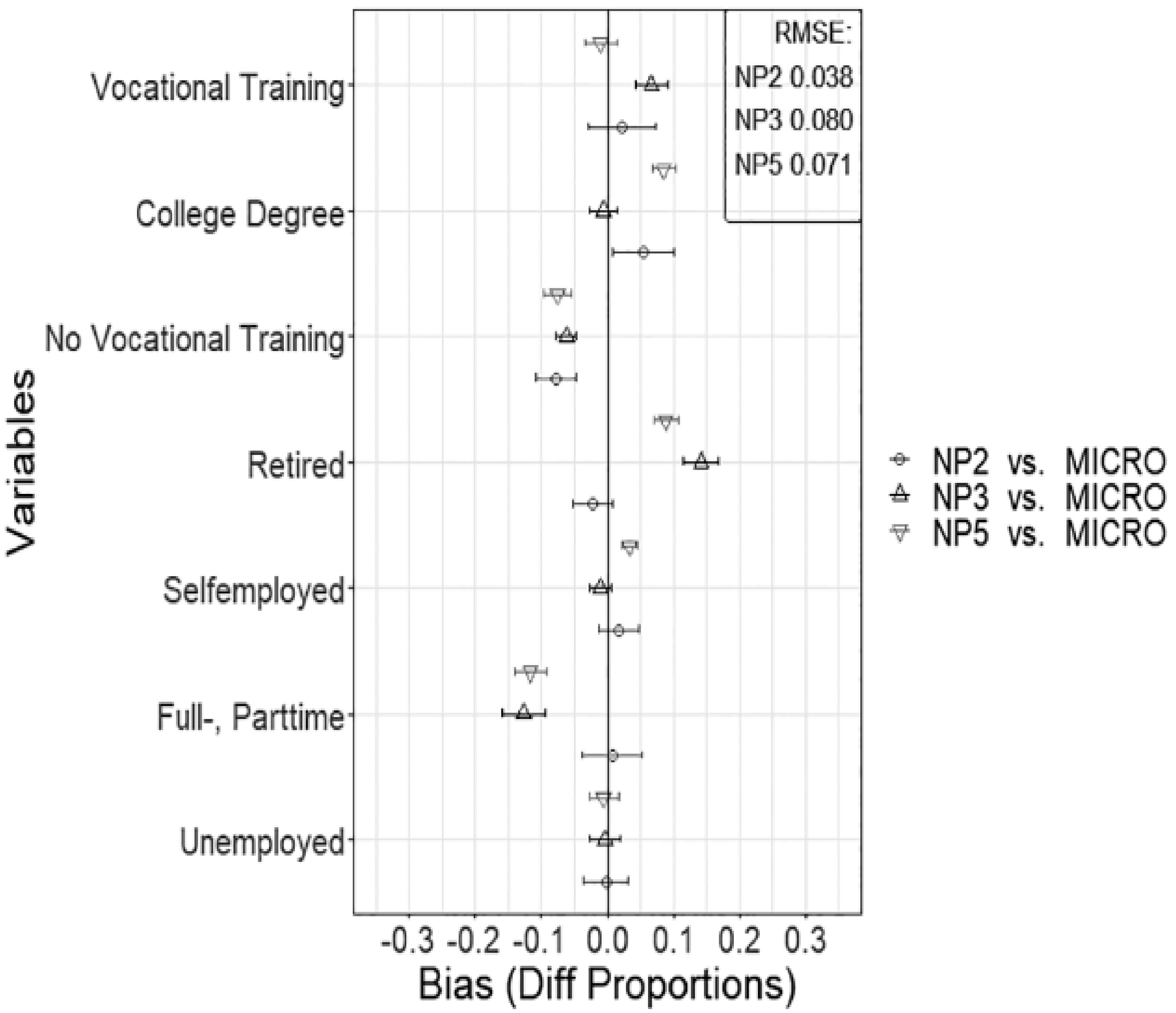
We compared three probability and five nonprobability surveys in terms of selection bias.

- All German General Population Surveys
- Bias in comparison to the German Microcensus
- or the German GSS
- We used raking weights for all surveys

GESIS Leibniz Institute for the Social Sciences

Univariate Comparison – Job-Demographics

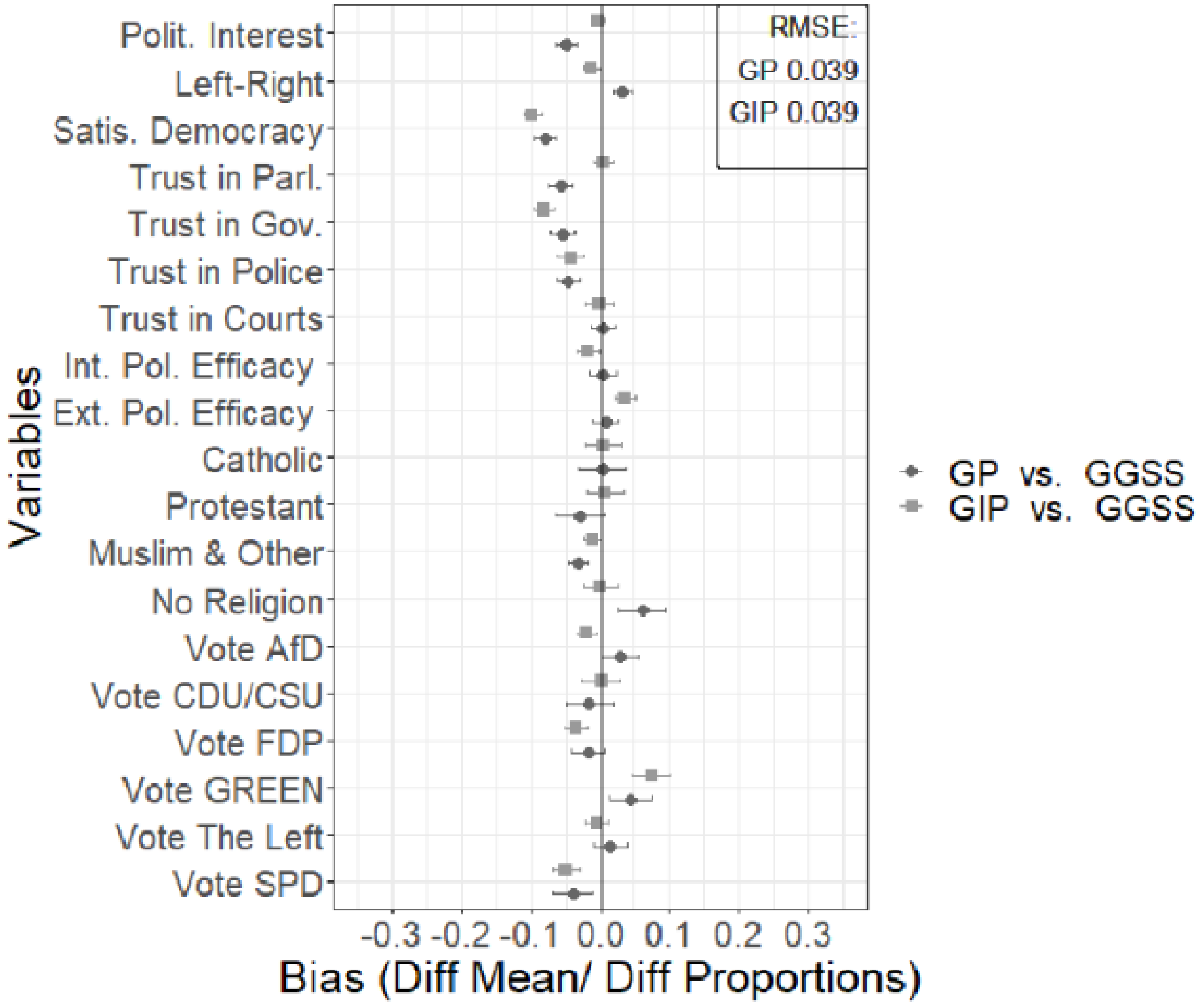


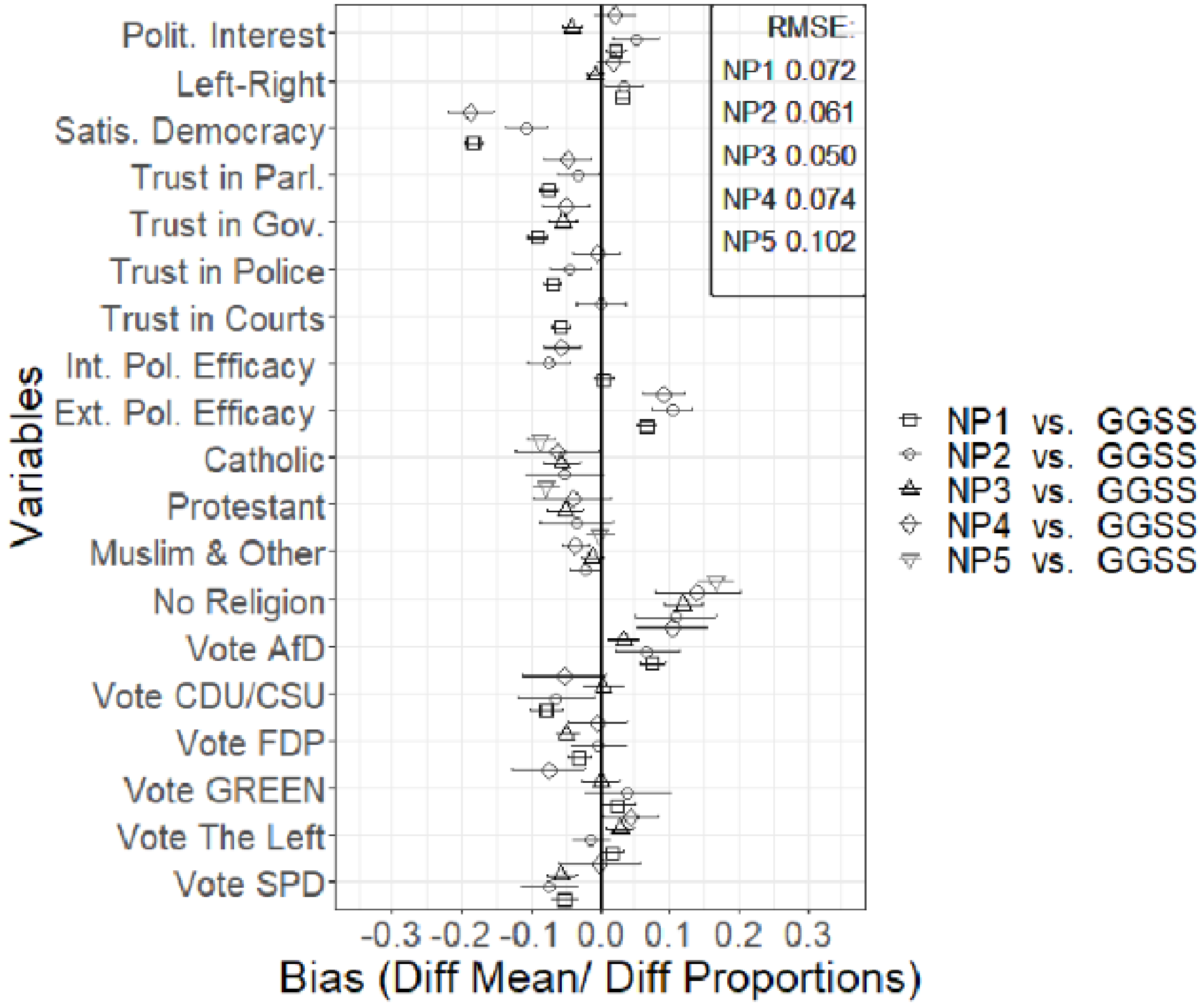


Note: MICRO = German microcensus (Source: Research Data Center [RDC] of the Federal Statistical Office and Statistical Offices of the La¨nder, Mikrozensus 2019; author’s calculations).

Rohr, Björn, Henning Silber, and Barbara Felderer. 2025. “Comparing the Accuracy of Univariate, Bivariate, and Multivariate Estimates across Probability and Nonprobability Surveys with Population Benchmarks.” *Sociological Methodology* 00811750241280963. doi: 10.1177/00811750241280963.

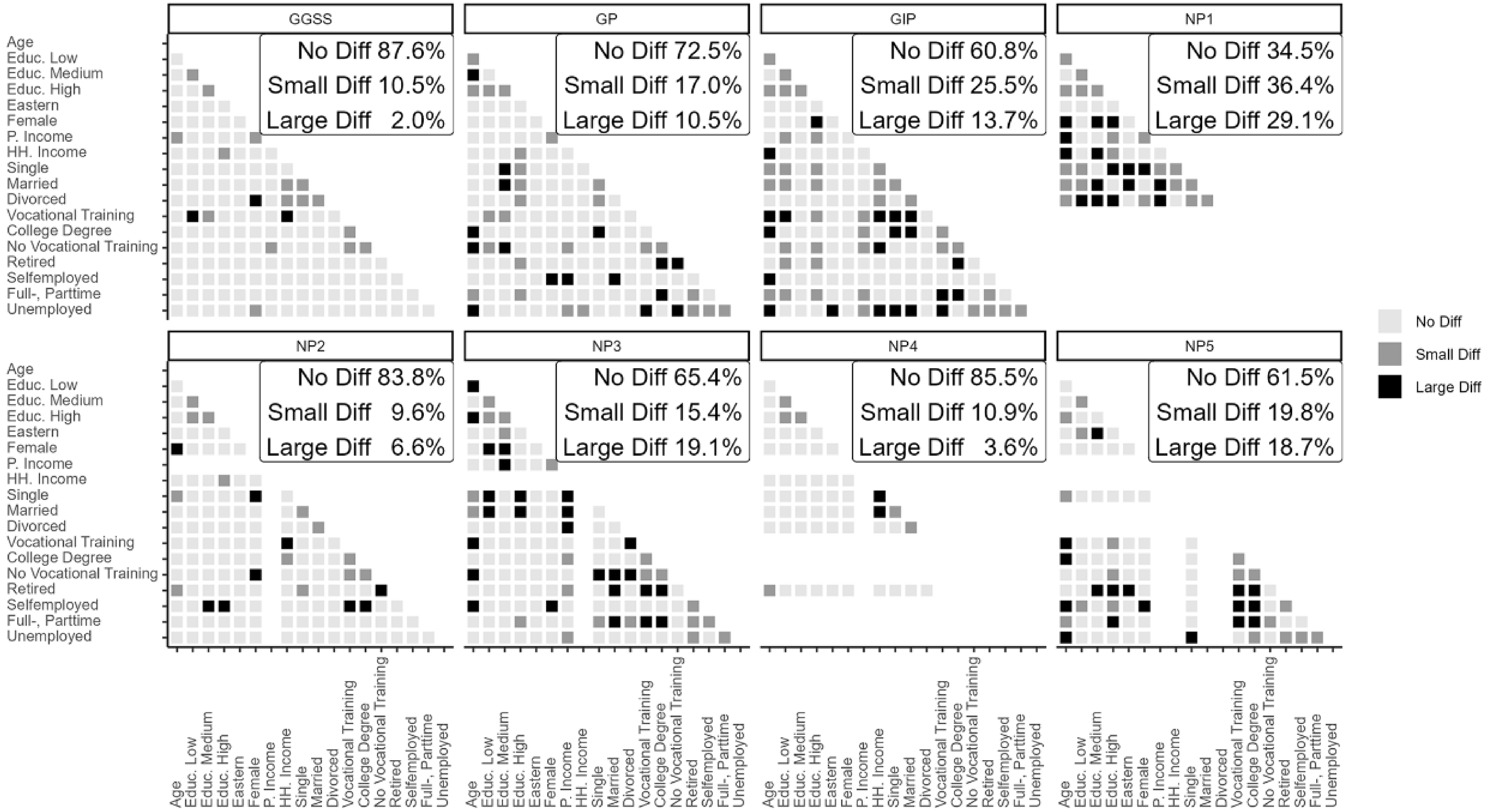
Univariate Comparison – Substantive Variables





Rohr, Björn, Henning Silber, and Barbara Felderer. 2025. "Comparing the Accuracy of Univariate, Bivariate, and Multivariate Estimates across Probability and Nonprobability Surveys with Population Benchmarks." *Sociological Methodology* 00811750241280963. doi: 10.1177/00811750241280963.

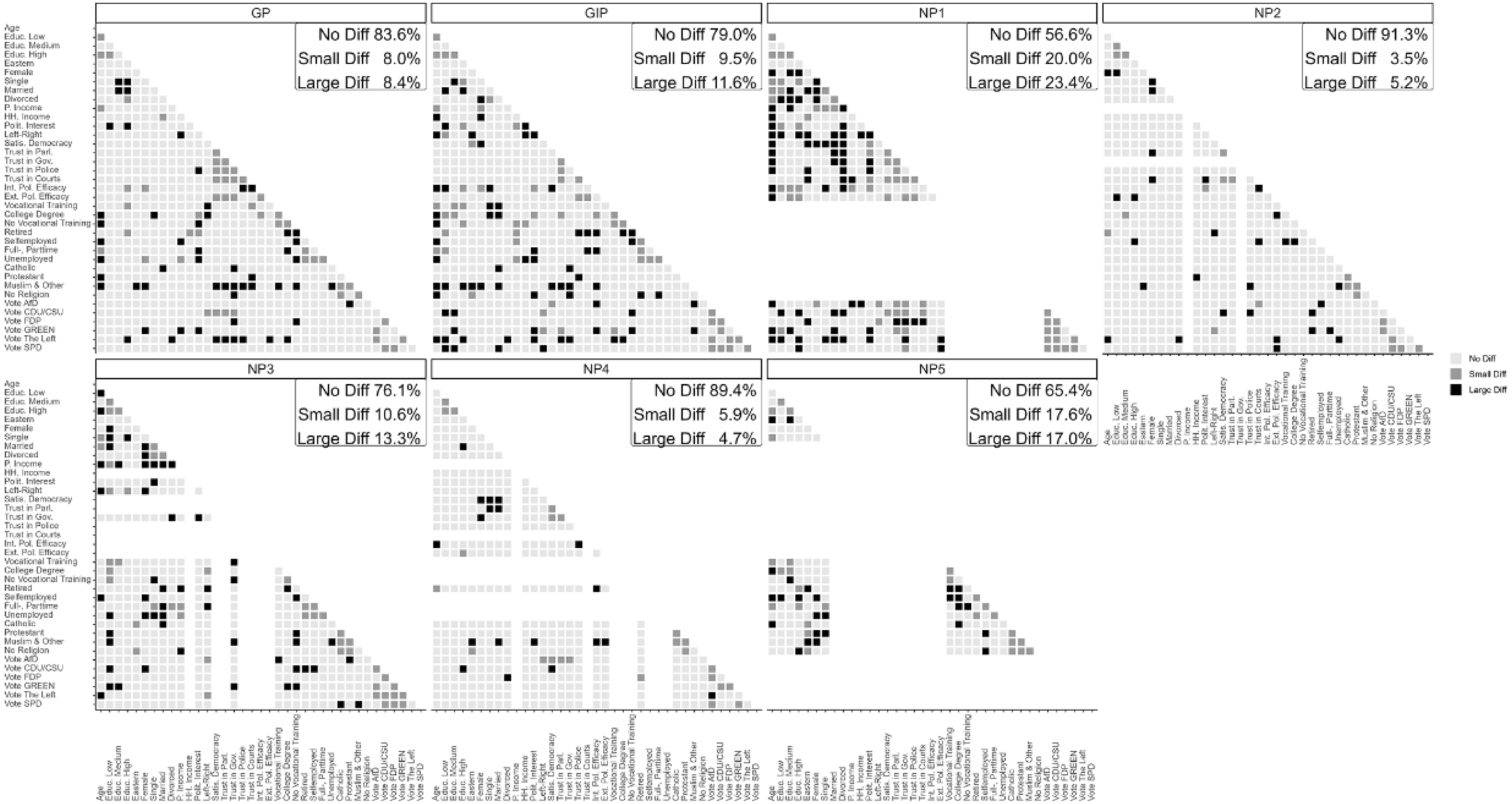
Bivariate Comparison – Job Demographics



Note: Benchmark = German microcensus (Source: Research Data Center [RDC] of the Federal Statistical Office and Statistical Offices of the Länder, Mikrozensus 2019; author's calculations).

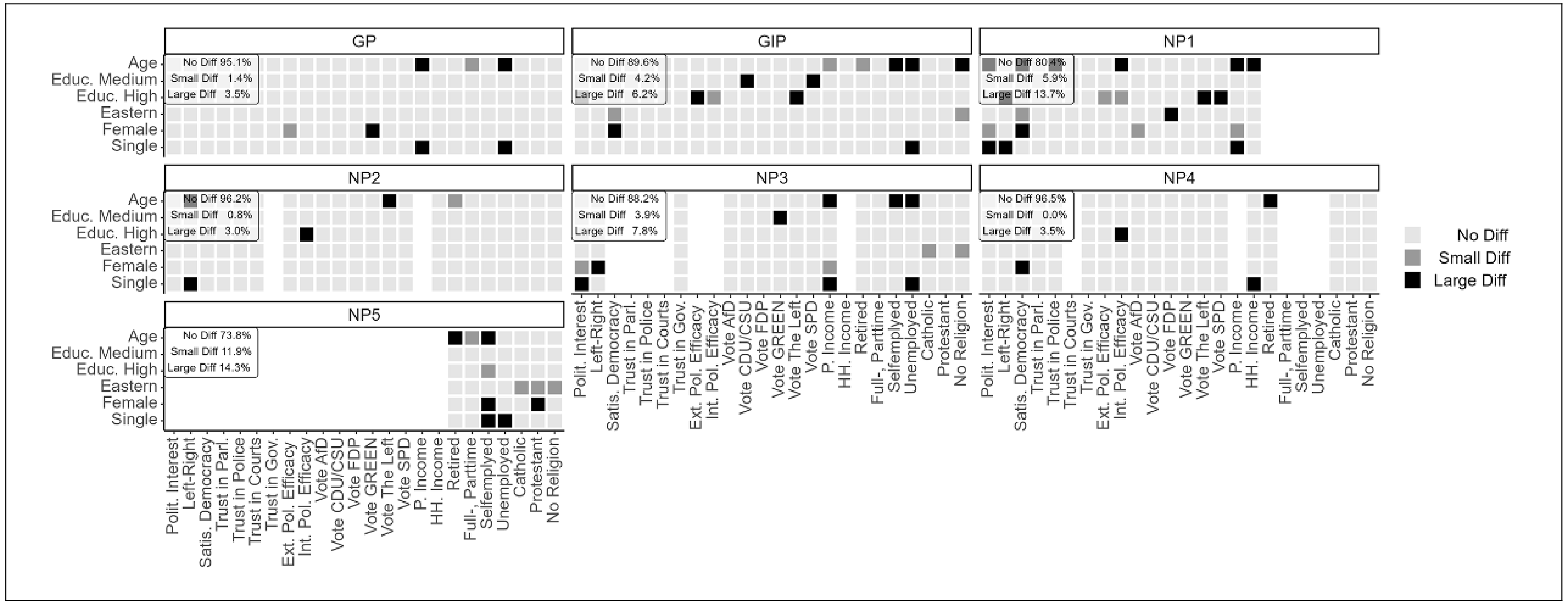
Rohr, Björn, Henning Silber, and Barbara Felderer. 2025. "Comparing the Accuracy of Univariate, Bivariate, and Multivariate Estimates across Probability and Nonprobability Surveys with Population Benchmarks." *Sociological Methodology* 00811750241280963. doi: 10.1177/00811750241280963.

Bivariate Comparison – Substantive Variables



Rohr, Björn, Henning Silber, and Barbara Felderer. 2025. "Comparing the Accuracy of Univariate, Bivariate, and Multivariate Estimates across Probability and Nonprobability Surveys with Population Benchmarks." *Sociological Methodology* 00811750241280963. doi: 10.1177/00811750241280963.

Multivariate Comparison



Rohr, Björn, Henning Silber, and Barbara Felderer. 2025. “Comparing the Accuracy of Univariate, Bivariate, and Multivariate Estimates across Probability and Nonprobability Surveys with Population Benchmarks.” *Sociological Methodology* 00811750241280963. doi: 10.1177/00811750241280963.

When is Selective Participation a Problem?

In the following, we sometimes show how selective nonresponse will cause bias in a survey. Note that the process is similar for any kind of selective participation.

When is Selective Participation a Problem?

- Selective Participation does not necessarily lead to bias in every estimate.
- An estimate of **blood type** based on a survey of students might not be biased, even when **women are overrepresented**.
- An estimate of the **political preferences and attitudes** might be more biased.

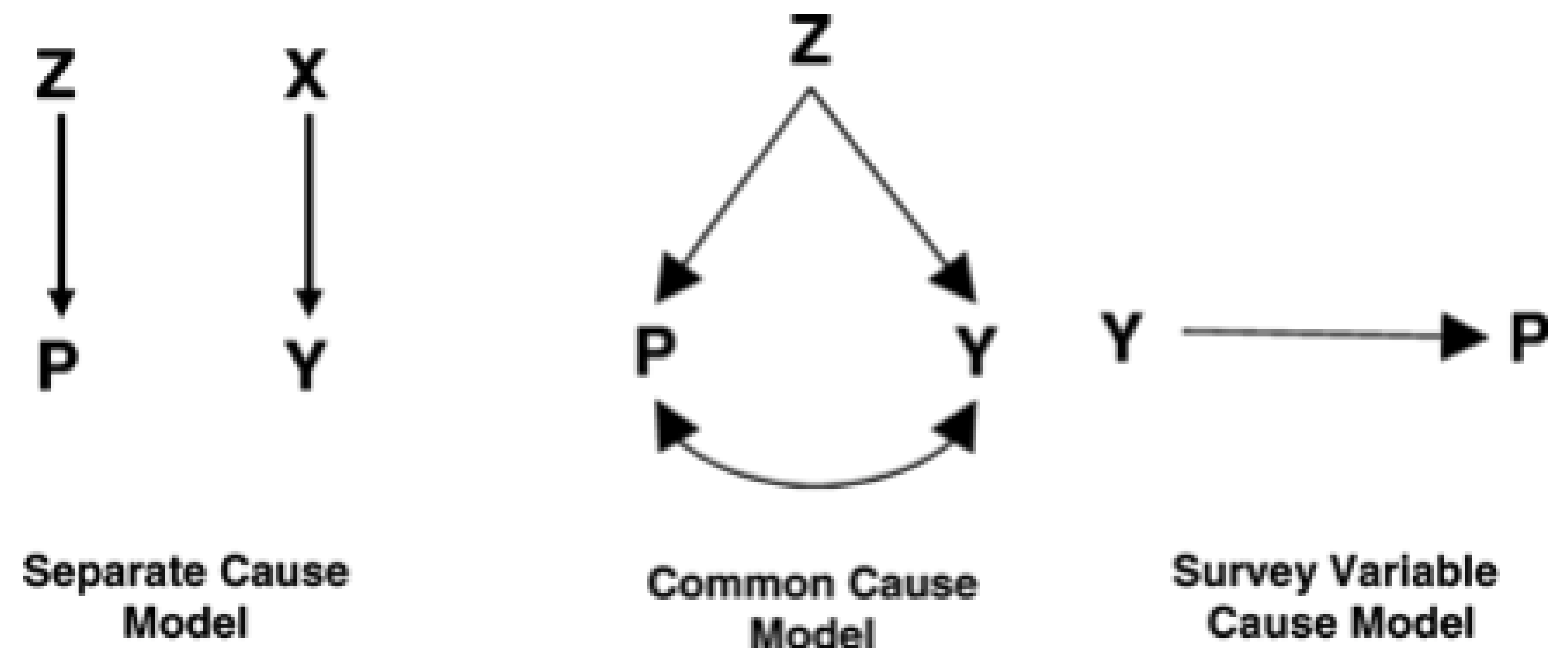
Three Causal Models Linking Response Propensity with Selection Bias

The models involve the following elements:

- Participation propensity: P
- survey variable of interest: Y
- variables that causally affect participation propensities and/or are related to the survey variable of interest: Z, X

Although this model was originally designed with nonresponse (and thus probability sampling) in mind, it works similarly for selective participation in general.

Selective Participation and Selection Bias



Robert M. Groves, Emilia Peytcheva, The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis Public Opinion Quarterly, Volume 72, Issue 2, Summer 2008, Pages 167–189, https://doi.org/10.1093/poq/nfn011

Selective Participation and Selection Bias

Groves (2006) distinguishes between three nonresponse models:

- Separate Cause Model:
  Variables Z are associated with response propensity P but not with the variable of interest Y. Y is only associated with variable X.
- Common Cause Model:
  The same set of variables Z are associated with P and Y.
- Survey Variable Cause Model:
  The variable of interest Y is associated with the response propensity.

GESIS Leibniz Institute for the Social Sciences

Selection Bias

Leibniz Association

Nonresponse Bias

$$Bias(\bar{Y}_R) = \bar{Y}_R - \bar{Y}$$

with

- mean true value of the survey respondents* $\bar{Y}_R$,
- the full population mean $\bar{Y}$.

Selection bias is different from zero if respondents systematically differ from the full population.

*In this context we ignore sampling errors and measurement errors

Nonresponse bias

Some manipulation leads to the *deterministic view* on nonresponse bias:

$$Bias(\bar{Y}_R) = (1 - \mathrm{RR})(\bar{Y}_R - \bar{Y}_{NR})$$

with

- Number of respondents n,
- Sample size N,
- Response rate RR = $\frac{n}{N}$
- population mean of nonrespondents is $\bar{Y}_{NR}$

Nonresponse bias

The *stochastic view* on nonresponse bias (Bethlehem, 1988) assumes that every individual of the sample has some unknown propensity to respond:

$$Bias(\bar{Y}_R) \approx \frac{1}{\bar{\rho}} Cov(y, \rho)$$
$$\approx \frac{1}{\bar{\rho}} Cor(y, \rho)\sigma_y\sigma_\rho$$

- Response propensity ρ which is a function of personal characteristics, characteristics of the situation, survey…
- Mean response propensity $\bar{\rho}$,
- Population correlation and covariance $Cor(y, \rho)$ and $Cov(y, \rho)$
- Population standard deviations $\sigma_\rho$and σ_y

Nonresponse bias

Two factors affect nonresponse bias.

- the (non-) response rate and
- the differences between respondents and nonrespondents.
- No clear relationship exists between the two factors.

No response bias exists if:

- the response rate is 100% or
- respondents do not differ from nonrespondents.

Key Difference between Nonresponse Bias and Selection Bias

Because of: $Bias(\bar{Y}_R)$ = (1 - RR)($\bar{Y}_R$-$\bar{Y}_{NR}$)

- Information from the sample frame can be used to estimate nonresponse bias.
  - For nonprobability surveys, there is no sample frame.
  - The same problem exists for characteristics not available in the sample frame

Therefore, we have to rely on: $Bias(\bar{Y}_R)$ = $\bar{Y}_R$- $\bar{Y}$,

- to calculate nonresponse bias based on population benchmarks
  - as population benchmarks are only available for a few characteristics, we often have to rely on other benchmarks (e.g., high quality probability surveys) and
  - have to assume that their estimates are without bias.

Selection Bias Analysis

Ingredients of Nonresponse Bias Analysis

- The **nonresponse bias formulas** are based on “true” values for the general populations, respondents and nonrespondents, which are usually not known.
- Some of them can be estimated using observed survey and/or auxiliary information (survey benchmarks).
- Instead of determining nonresponse bias for a certain statistic, we rather evaluate the risk of nonresponse bias using several indicators.

Ingredients of Bias Analysis

- For nonprobability surveys, indicators like response propensity do not exist as they need a clearly defined sample of potential respondents.
- Nonresponse bias always depends on the variable of interest.
- We can only measure bias in non-probability surveys for estimates where we already have an accurate benchmark estimate.
- For all other estimates, we can only assume (or hope) that they are unbiased.
- Expert knowledge is needed to judge whether miss-representation of auxiliary variables is associated with nonresponse bias in the estimate of variables of interest.

Synthetic Data Sets

Data are generated to stem from two cross-sectional web surveys with N = 2000. The response rate is simulated to be 25 % in each survey (n = 500) but with differing nonresponse and selection mechanisms.

Socio-demographics are available from the sample frame.

Frame data:

- age in five categories
- gender
- nationality
- response indicator (1=response/0=nonresponse)

data sets „frame1“ and „frame2“.

Synthetic Data Sets

Survey 1 (survey1_prob): randomly sampled individuals from Registration Office.

- Key survey variables: life satisfaction, IT knowledge and IT literacy.
- Nonresponse is affected by the socio-demographic variables age, gender, education, nationality.
- Survey variable „life satisfaction“ is not correlated to socio-demographics.
- Survey variable „IT knowledge“ is correlated to observed socio-demographics.
- Survey variable „IT literacy“ is correlated to observed socio-demographics and unobserved education.
- Additional Variable: internet usage

Synthetic Data Sets

Survey 2 (survey2_nonprob):

- Key survey variable: Internet usage.
- Nonresponse is not affected by the socio-demographics but by frequency of Internet usage (not known from the frame).
- Additional variables: life satisfaction, IT knowledge and IT literacy.

Synthetic Data Sets

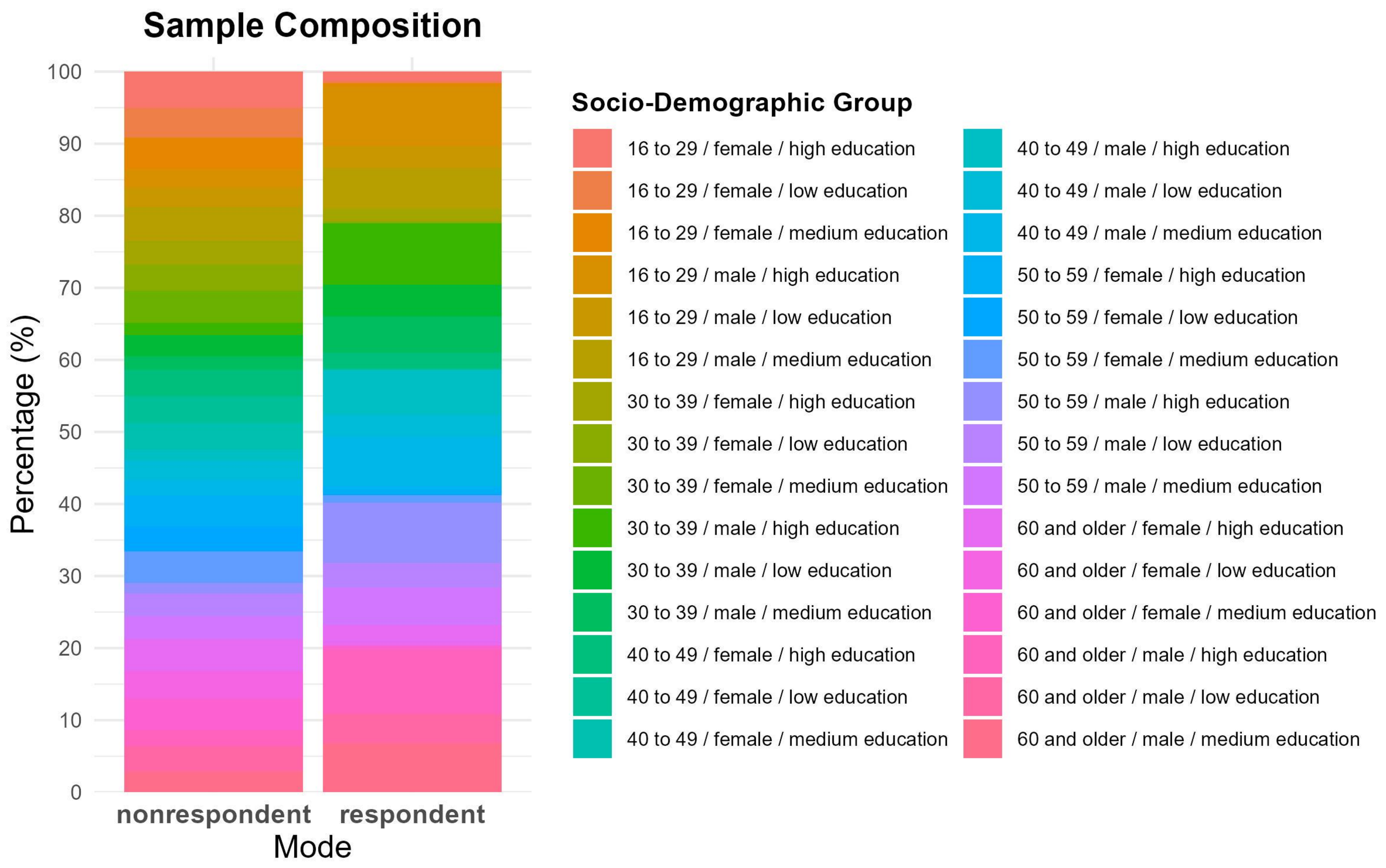
Unavailable personal or survey information for N=2000 nonrespondents and respondents are collected in a separate data set for illustration purpose:

Unobserved survey 1 and unobserved survey 2 include:

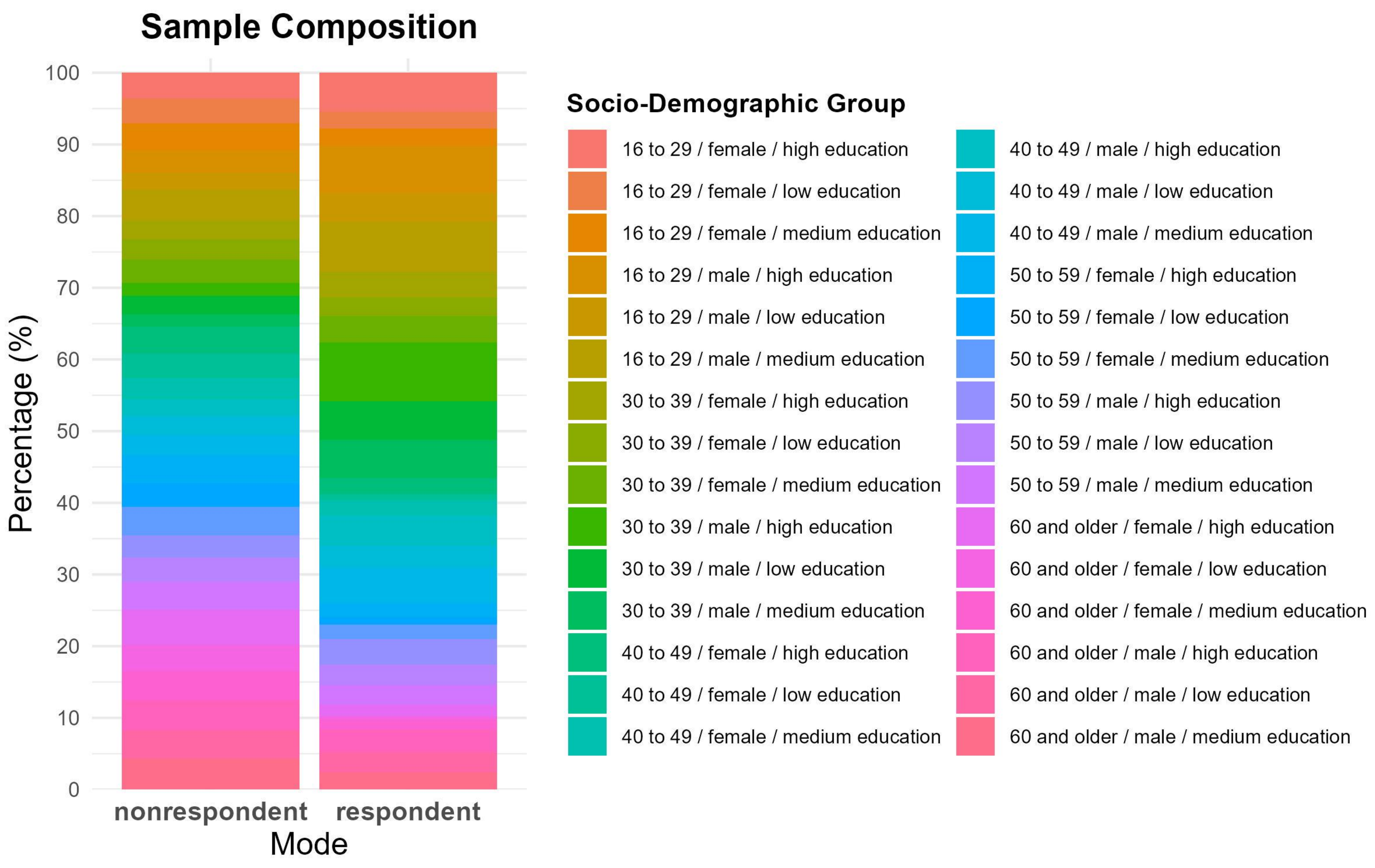
- education in three categories and
- life satisfaction, IT knowledge, IT literacy and Internet usage.

Data sets „unobserved1“ and „unobserved2“.

Sample Composition of probability-based Survey 1



Sample Composition of non-probabilistic Survey 2



Exercises

GESIS Leibniz Institute for the Social Sciences

Exercise 1: Nonresponse Model

Leibniz Association

Exercise 1 Nonresponse Model

Think about a probability Survey (Survey 1):

Which X variables can you use from the frame?

Which Y variables do you plan to analyze?

Run a logistic regression model (nonresponse model) to investigate which variables are associated with survey nonresponse and to what extent.

Interpret the findings. What do you learn from it?

Exercise 1: Nonresponse Model – Frame Information

```
svy1.response <- glm(response ~ age_16to29+age_30to39+age_40to49+
                     age_50to59+age_60p+german+female,
                     family = "binomial", data = frame1)

summary(svy1.response)
```

First, we ran a model based on information available in the sample frame

- The variable indicating response to the survey is the dependent variable.
- Demographics are the independent variables.

Exercise 1: Nonresponse Model – Frame Information

```
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.67194    0.36139  -7.394 1.43e-13 ***
age_16to291 -0.33753    0.17985  -1.877   0.0606 .  
age_30to391  0.04607    0.18396   0.250   0.8022    
age_40to491  0.09579    0.18799   0.510   0.6104    
age_50to591 -0.09862    0.18318  -0.538   0.5903    
age_60p1          NA         NA      NA       NA    
german1      2.67667    0.34936   7.662 1.84e-14 ***
female1     -2.53448    0.14978 -16.921  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Age does not significantly affect survey response
- German citizens have **significantly** higher likelihood to respond than foreigners
- Females a **significantly** lower likelihood than men

Exercise 1: Nonresponse Model – Frame Information

```
svy1.response.unobs <- glm(response ~ age_16to29+age_30to39+age_40to49+
                             age_50to59+age_60p+german+female+
                             edu_low+edu_mid+edu_high+usage+
                             satisfaction+itknowledge+
                             itliteracy, family = "binomial", data = full_prob)
summary(svy1.response.unobs)
```

Second, we investigate whether the unobserved variables are associated with survey response and add them to the model

- This is only possible for our simulated data, as normally we do not know non-frame information (e.g., education, satisfaction, it-knowledge, etc., for nonrespondents)

Exercise 1: Nonresponse Model – Frame Information

```
Coefficients: (2 not defined because of singularities)
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)   -2.594898   0.492031  -5.274 1.34e-07 ***
age_16to291   -0.800888   0.544002  -1.472    0.141    
age_30to391   -0.191004   0.378533  -0.505    0.614    
age_40to491   -0.007674   0.275865  -0.028    0.978    
age_50to591   -0.113320   0.202451  -0.560    0.576    
age_60p1             NA         NA      NA       NA    
german1        2.934405   0.357715   8.203 2.34e-16 ***
female1       -2.869562   0.186970 -15.348  < 2e-16 ***
edu_low1      -1.645555   0.191078  -8.612  < 2e-16 ***
edu_mid1      -1.026239   0.156696  -6.549 5.78e-11 ***
edu_high1            NA         NA      NA       NA    
usage          0.002414   0.020542   0.118    0.906    
satisfaction   0.099897   0.060772   1.644    0.100    
itknowledge    0.046484   0.062447   0.744    0.457    
itliteracy     0.002882   0.013024   0.221    0.825    
```

- Education also **significantly** affects response.
- Our substantial variables (internet usage, satisfaction, it-knowledge, it-literacy) are **not significantly related** to response

GESIS Leibniz Institute for the Social Sciences

Exercise 2: R-Indicator

Leibniz Association

Exercise 2 R-Indicator

- The R-indicator is a measure to evaluate how representative a survey is with regard to several characteristics.
- The R-indicator is based on predicted response propensities
- It was developed by Natalie Shlomo, Chris Skinner and Barry Schouten (2012)
- The indicator is between 0 and 1, where 1 indicates full representativeness of the survey, while 0 indicates maximum miss-representation.
- Partial R-Indicators can give insights into how individual variables contribute to bias. (Not covered here)

Shlomo, N., Skinner, C., & Schouten, B. (2012). Estimation of an indicator of the representativeness of survey response. Journal of Statistical Planning and Inference, 142(1), 201–211. https://doi.org/10.1016/j.jspi.2011.07.008

Exercise 2 R-Indicator

Use the R_indicator() function of sampcompR to compute the R-indicator for your survey using frame Information.

What happens to the R-indicator if you drop an influential X-variable from the nonresponse model?

GESIS Leibniz Institute for the Social Sciences

Exercise 2 R-Indicator

```
R_indicator(dfs=c("frame1"),
            variables=c("age_30to39","age_40to49","age_50to59",
                        "age_60p","female","german"),
            response_identificators = c("response"),
            get_r2 = T)
```

```
$frame1
 R-Indicator           SE    Pseudo R2
0.5757485009 0.0003711326 0.2308003458
```

- The R indicator (~0.576) indicates that the survey suffers from some miss-representation.
- The (mcfadden)- Pseudo R2 of ~0.231 shows that the chosen variables are suited to explain a good amount of variance in the response model, although this certainly can be improved.

GESIS Leibniz Institute for the Social Sciences

Exercise 2 R-Indicator

```
R_indicator(dfs=c("frame1"),
            variables=c("age_30to39","age_40to49","age_50to59",
                        "age_60p"),
            response_identificators = c("response"),
            get_r2 = T)
```

```
$frame1
 R-Indicator           SE    Pseudo R2
9.556446e-01 1.141078e-05 2.373082e-03
```

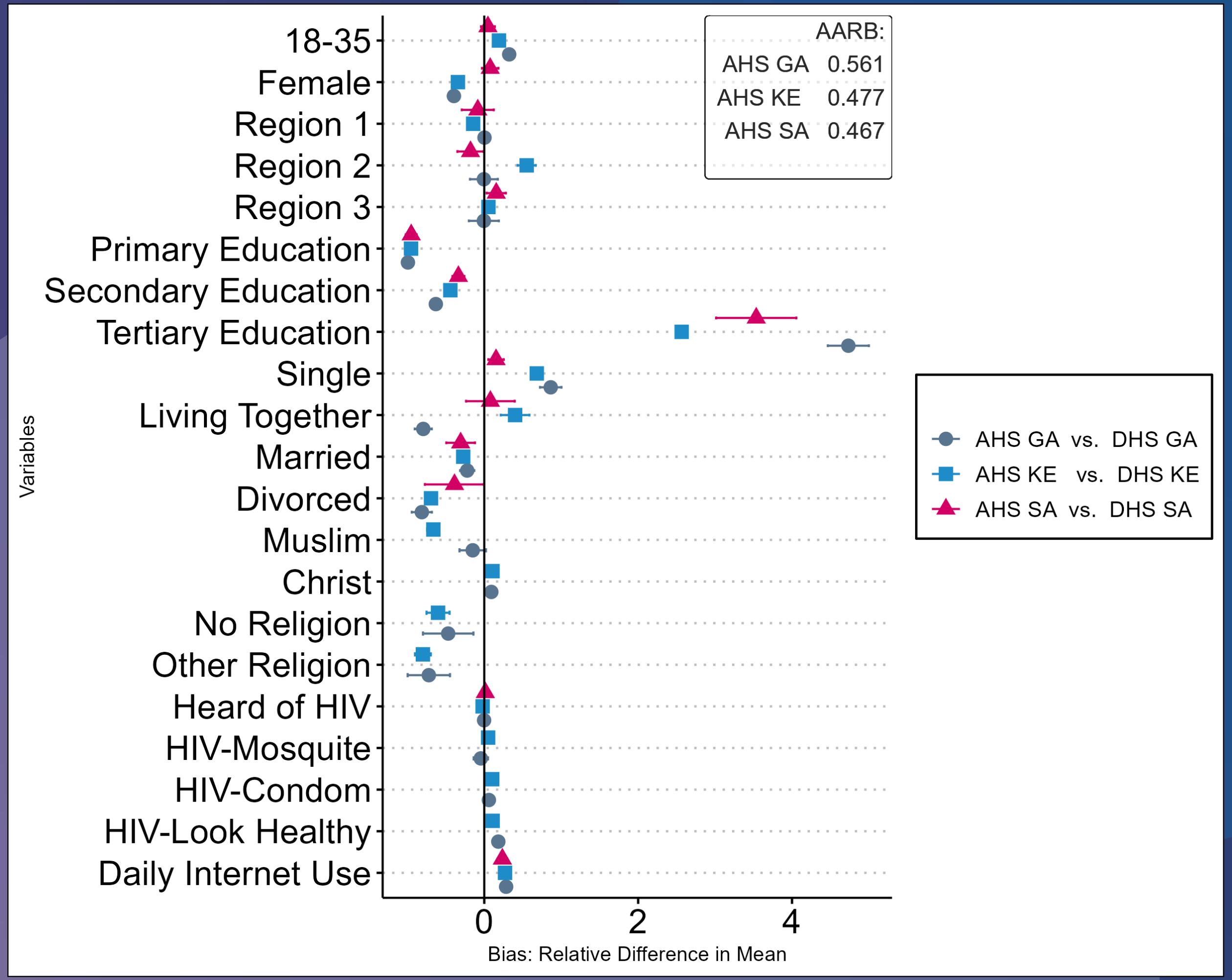
- When only age variables are used in the model, the R indicator shows nearly perfect representativeness (0.956)
- The results can only be interpreted as representativeness of age groups in the survey as compared to the frame
- The very low R-indicator of ~0.002 shows that the model is not well suited to explain the nonresponse.

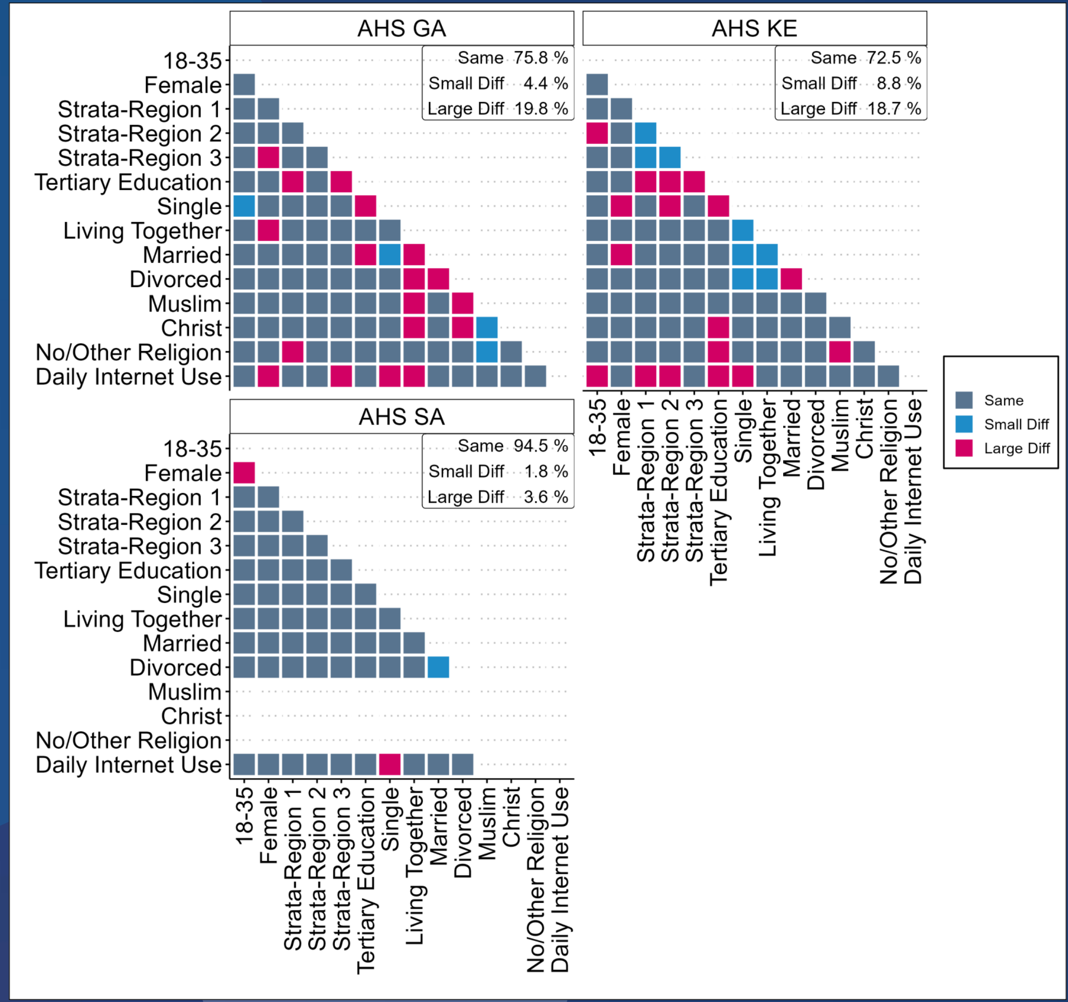
Exercise 3: Bias in univariate estimates

GESIS Leibniz Institute for the Social Sciences

What is SampcompR?

Leibniz Association





SampcompR

- SampcompR is an R-Package to estimate differences between samples (i.e., surveys and a benchmarks (survey)).

- Bias can be estimated on a univariate, bivariate, and multivariate level.

- Bias can be visualized in a plot or outputted as a table.

GESIS Leibniz Institute for the Social Sciences

Selected Functions of SampcompR

Univariate Functions

- uni_compare()
- plot_uni_compare()
- uni_compare_table()

Bivariate Functions

- biv_compare()
- plot_biv_compare()
- biv_compare_table()

Multivariate Functions

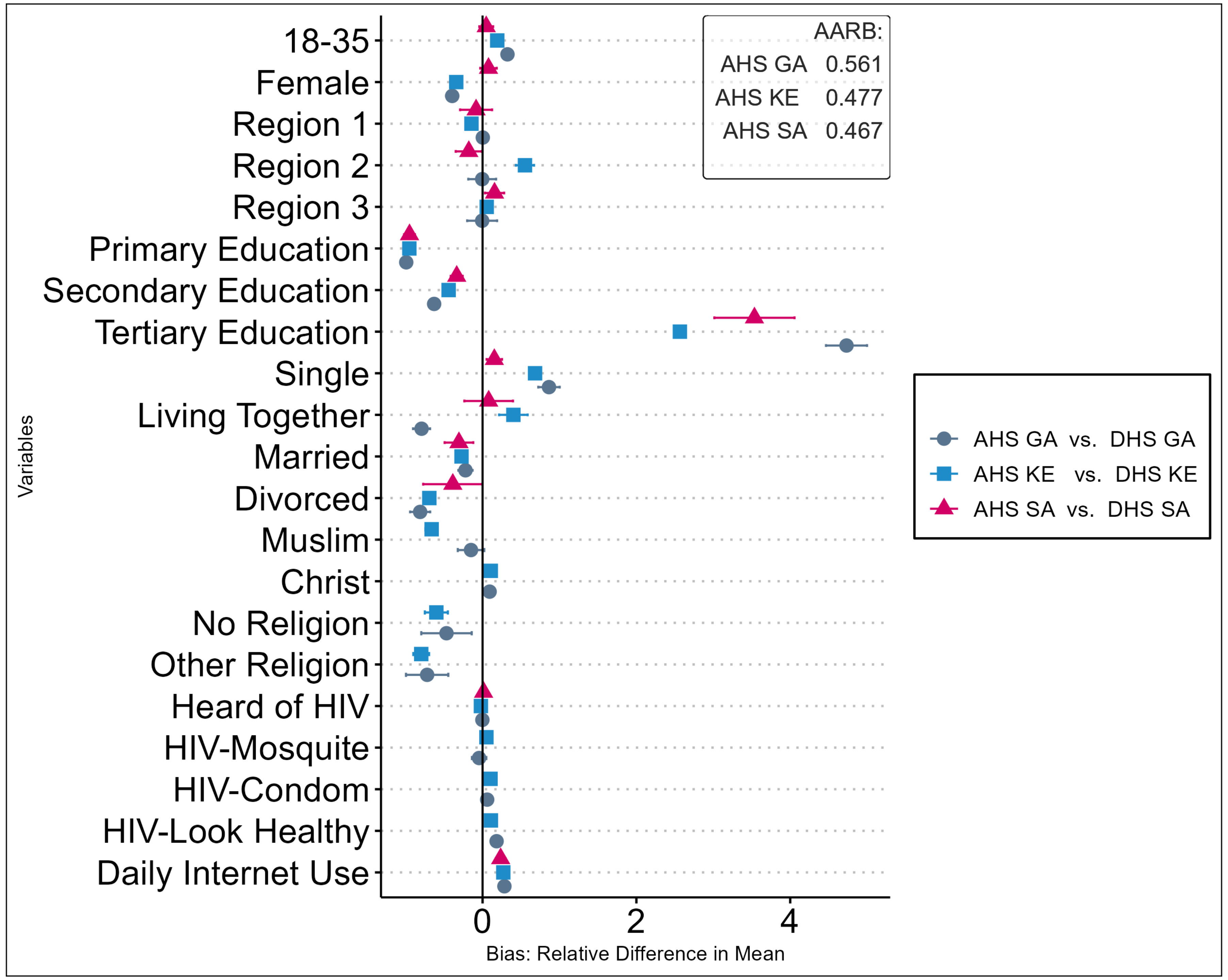
- multi_compare()
- plot_multi_compare()
- multi_compare_table()

Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```



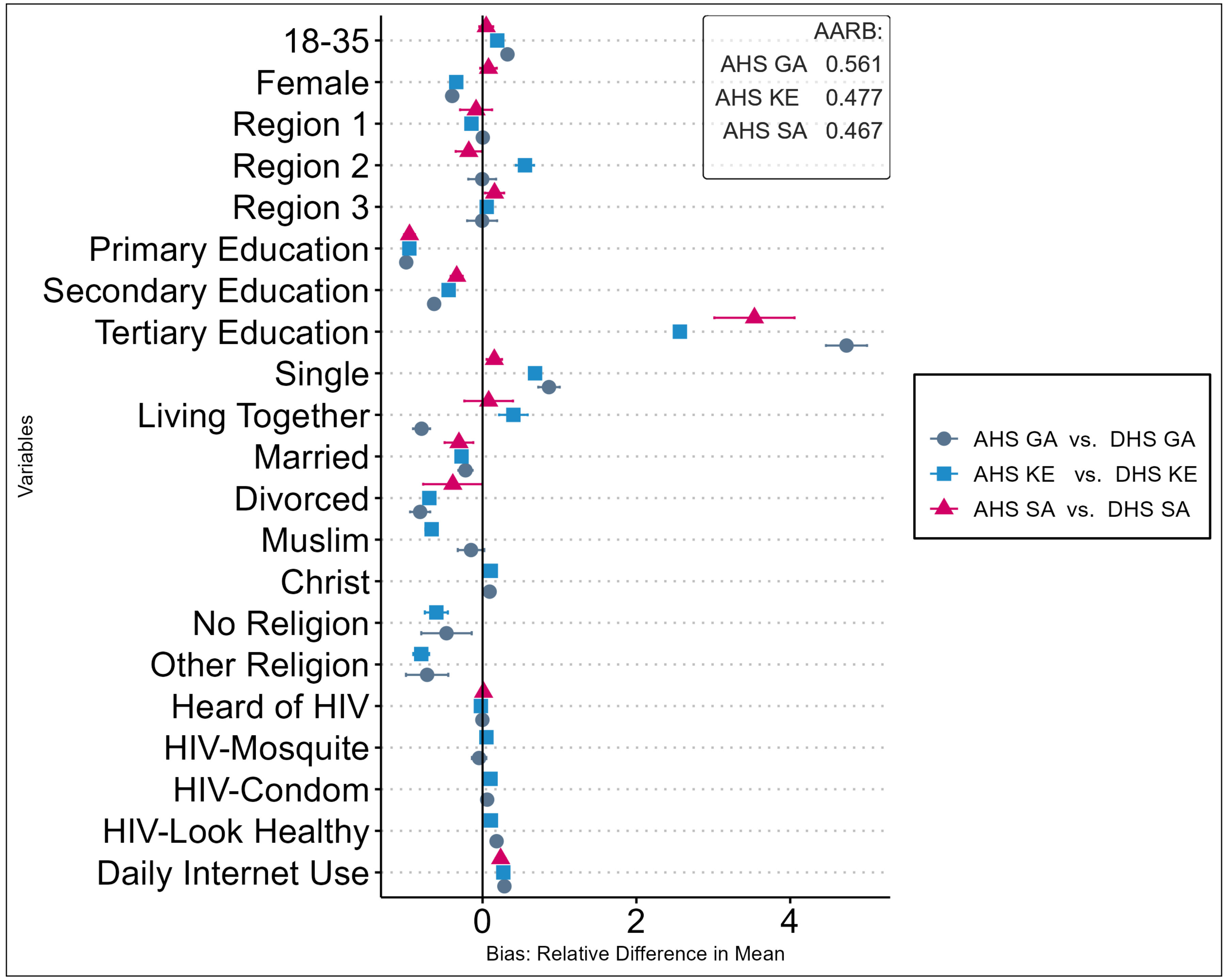
Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)

uni_plot<-plot_uni_compare(bias_AHS)
```

Three **surveys** are compared against three **benchmark surveys**



Example Univariate Comparison

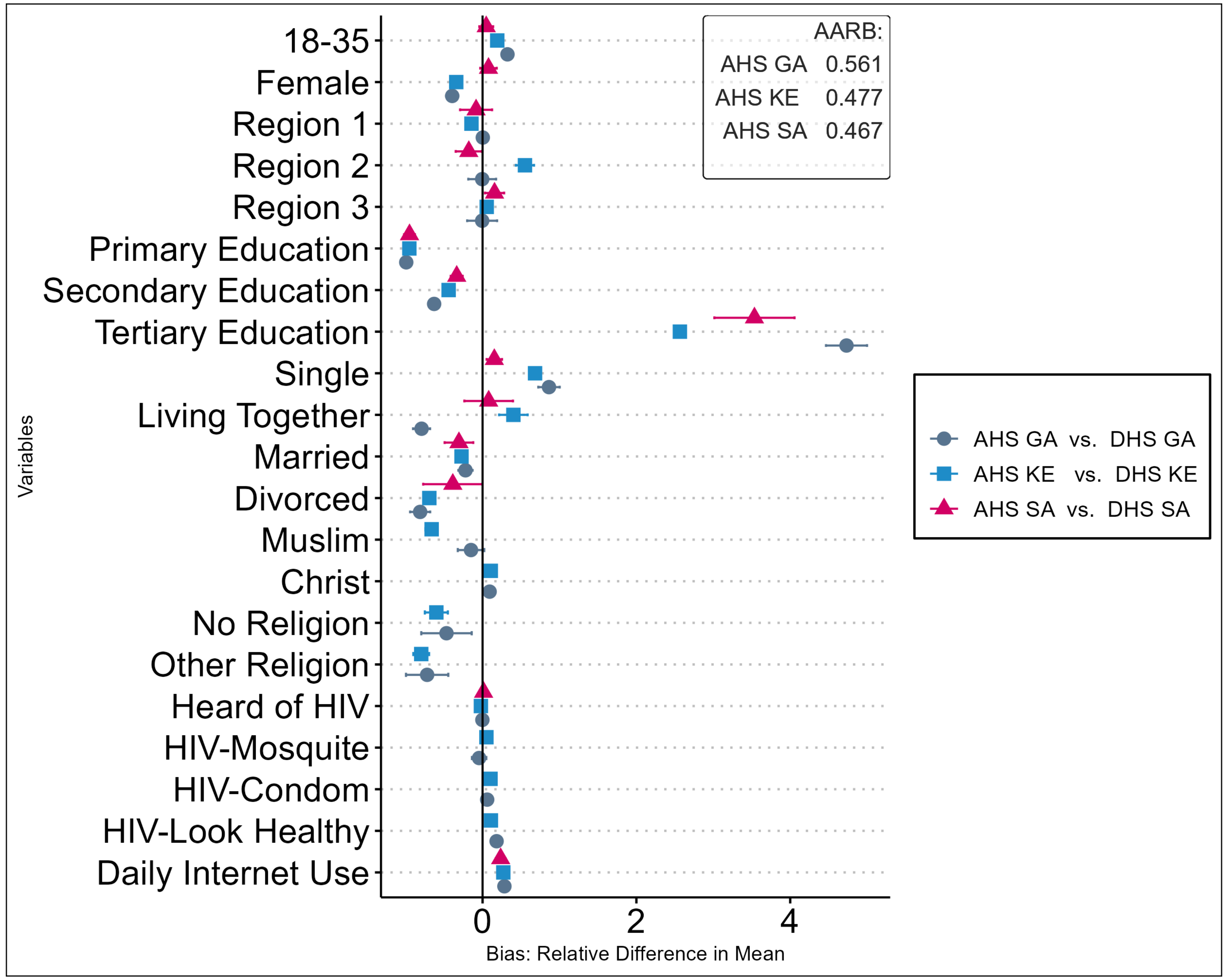
Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)


uni_plot<-plot_uni_compare(bias_AHS)
```

Bootstrap confidence intervals are calculated with 2000 bootstraps

Alternative: No bootstrap confidence intervals (nboots =0)



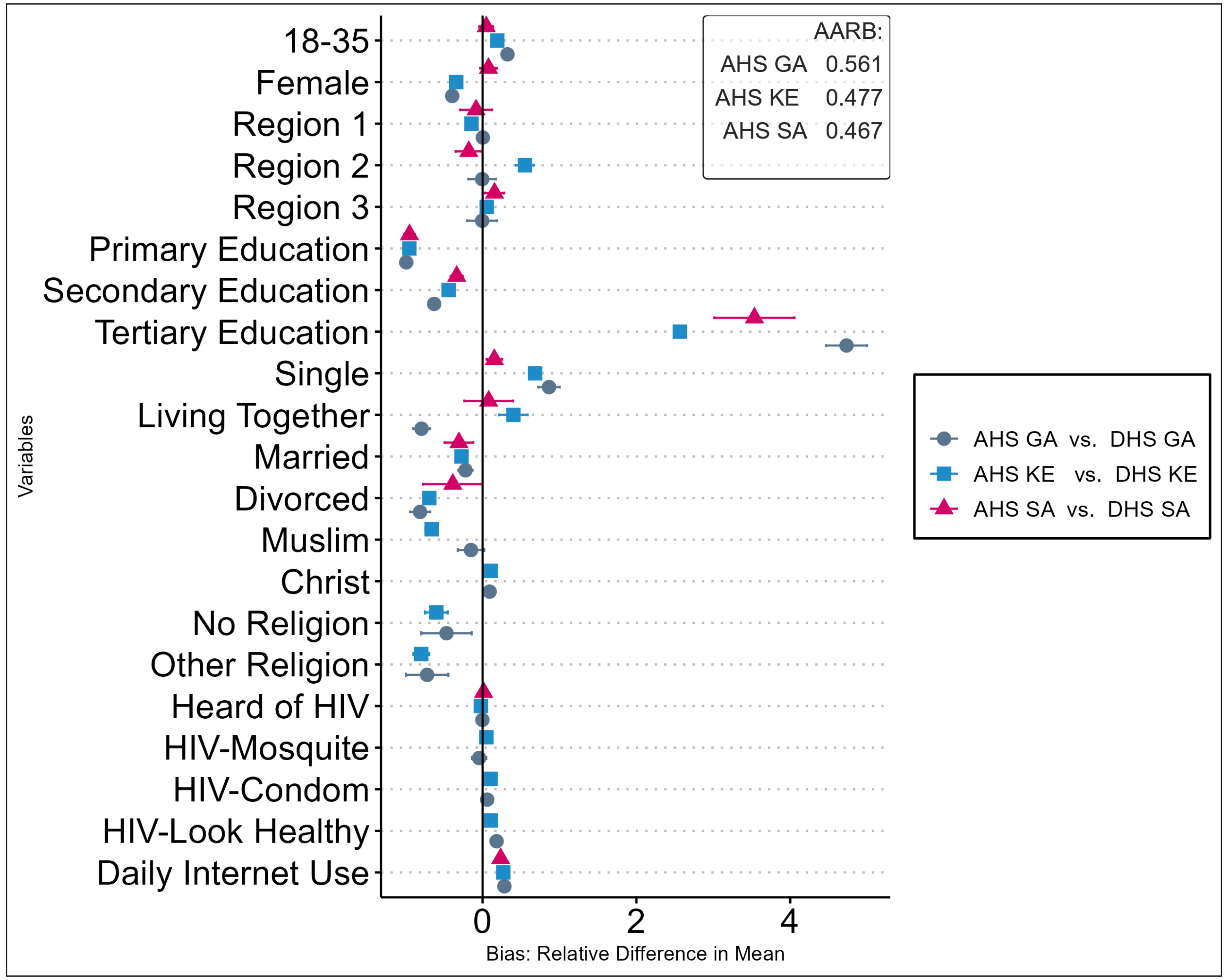
Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
variables<-c("young", "female", "strata_reg1", "strata_reg2", "strata_reg3",
             "primary_educ2", "secondary_educ","tertiary_educ", "single",
             "living_together", "married", "divorced","muslim", "christ",
             "no_rel", "other_rel","hiv_heard","hiv_mosquito_corr",
             "hiv_condom_corr","hiv_healthy_corr","internet_daily")
                      varlabels = varlabels,
                      colors = colors)

uni_plot<-plot_uni_compare(bias_AHS)
```

funct = rel_mean → Relative bias in mean is calculated
Alternatives: Absolute Relative bias in mean, (absolute) bias in mean



Example Univariate Comparison

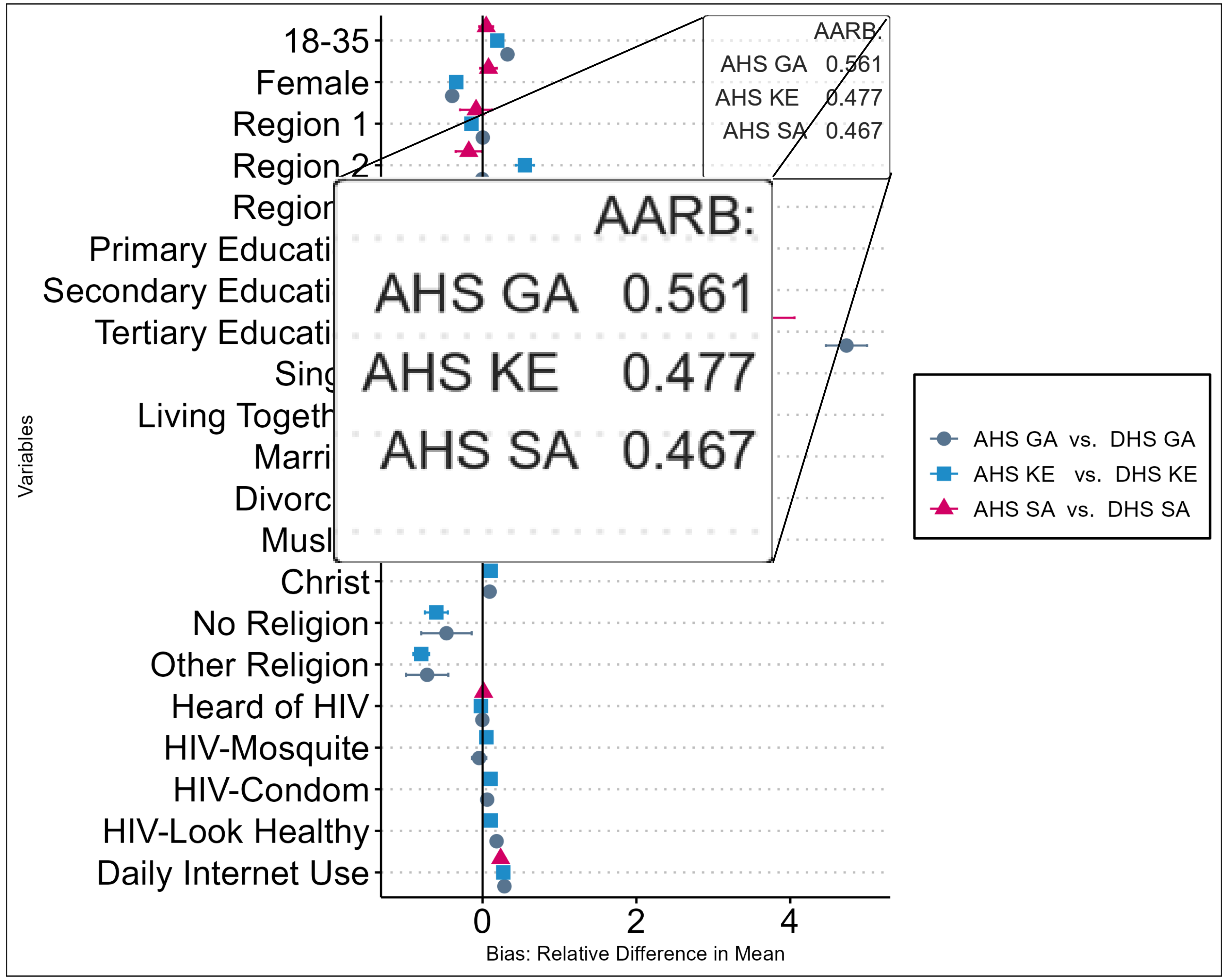
Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)

uni_plot<-plot_uni_compare(bias_AHS)
```

To summarize the bias in each survey, Average Absolute Relative Bias (AARB) is calculated
Alternatives: RMSE, MSE

The confidence intervals are Bonferroni corrected.
Alternatives: No correction



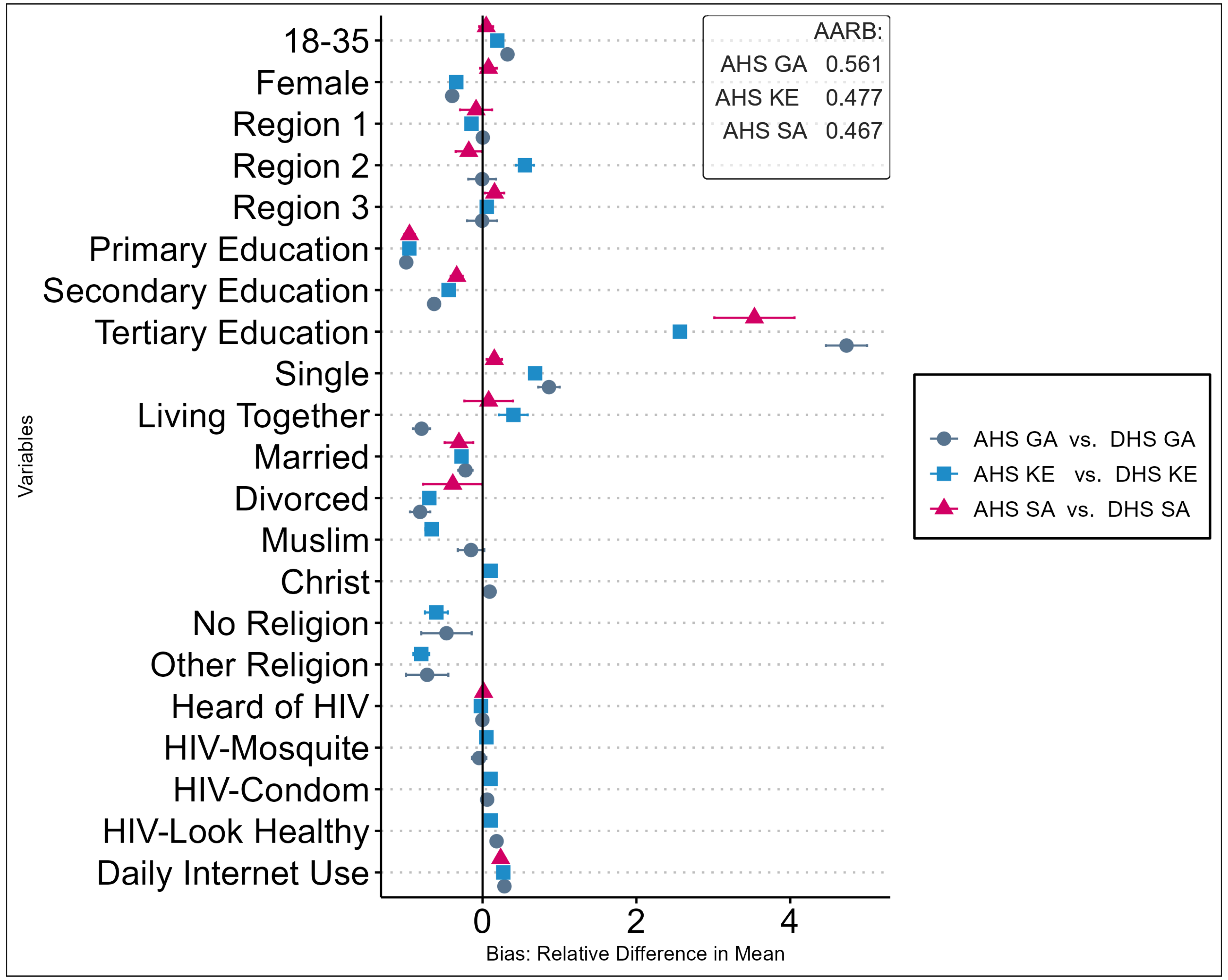
Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)

uni_plot<-plot_uni_compare(bias_AHS)
```

Analysis accounts for sampling design (e.g., weights and strata)
Alternatives: No weights, different weights in every benchmark.



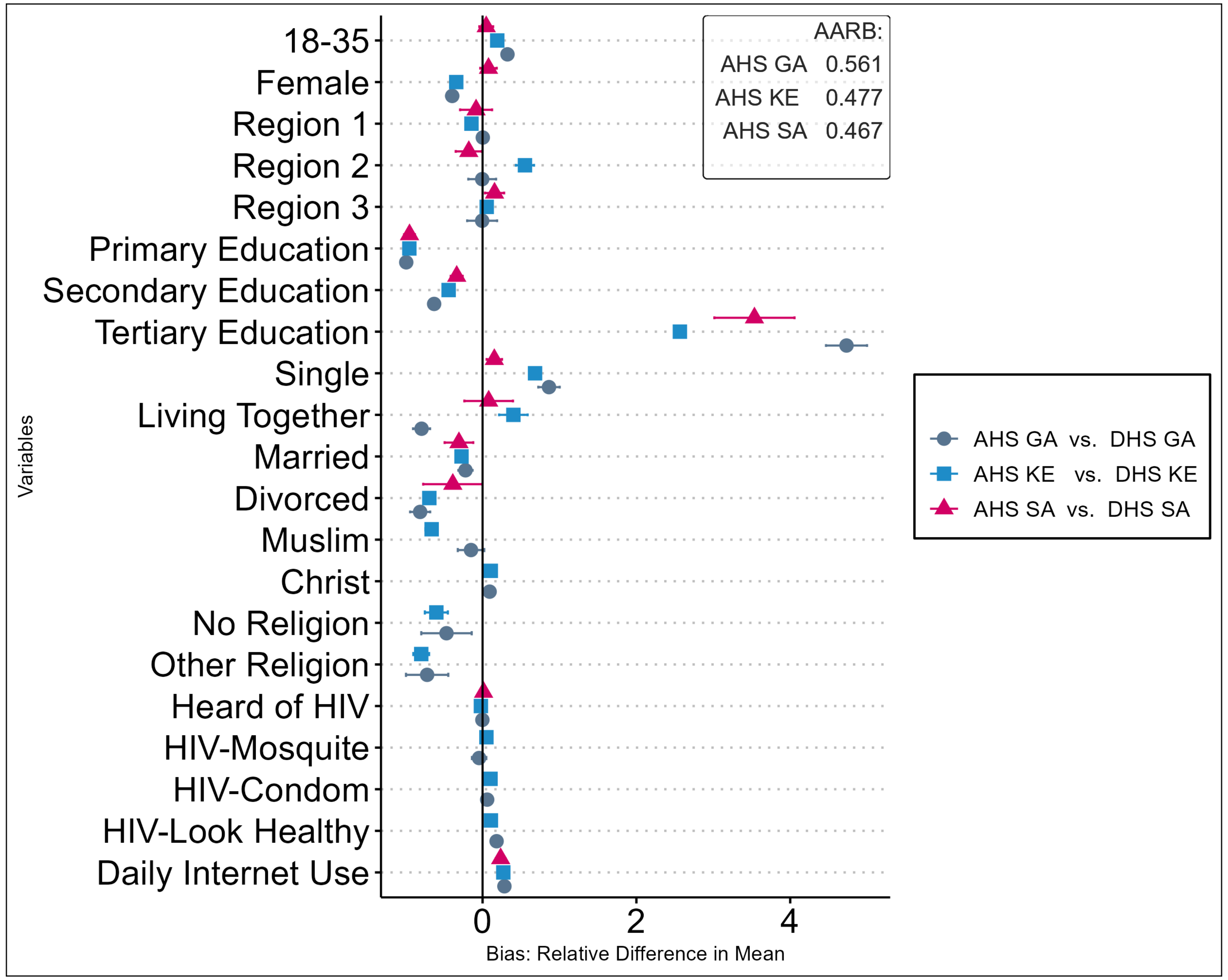
Example Univariate Comparison

Univariate Plot

```
bias_AHS<-uni_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                      benchmarks = c("DHS_Ga","DHS_Ke", "DHS_Sa"),
                      nboots = 2000,
                      funct = "rel_mean",
                      variables = variables,
                      summetric = "avg2",
                      conf_adjustment=T,
                      name_dfs = c("AHS GA",
                                   "AHS KE ",
                                   "AHS SA"),
                      name_benchmarks = c("DHS GA",
                                          "DHS KE",
                                          "DHS SA"),
                      weight_bench = c("weight"),
                      id_bench = c("cluster"),
                      strata_bench = c("strata"),
                      varlabels = varlabels,
                      colors = colors)

uni_plot<-plot_uni_compare(bias_AHS)
```

We put the results of uni_compare in the plot function to get a plot.



Example Univariate Comparison

Univariate Table

```r
univar_table <- uni_compare_table(bias_AHS, ci_line=F)


univar_kable_ahs <- univar_table |>
  kable(caption = "Bias: Difference in Proportions Between AHS and DHS",
        align = c("l","c","c"),
        booktabs = TRUE) |>
  row_spec(0, bold=TRUE) |>
  footnote(
    general_title = "Note.",
    general = "Bias; difference in proportions for dichotome
    variables; confidence intervals (in\n paranthesis) where
    measured by bootstrapping with 2000 iterations; N varries
    between\n variables and surveys due to missing information.",
    footnote_as_chunk = T
  ) |>
  kable_classic(full_width =F, position="center", font_size=14) |>
  row_spec(21, extra_css = "border-bottom: 1.5px solid;")

univar_kable_ahs
```

Bias: Difference in Proportions Between AHS and DHS

Variables	**AHS GA**	**AHS KE**	**AHS SA**
18-35	0.324 (0.267, 0.381)	0.191 (0.168, 0.214)	0.047 (-0.042, 0.137)
Female	-0.395 (-0.474, -0.316)	-0.343 (-0.378, -0.308)	0.077 (-0.030, 0.184)
Region 1	0.003 (-0.062, 0.068)	-0.145 (-0.177, -0.113)	-0.084 (-0.292, 0.123)
Region 2	-0.005 (-0.186, 0.176)	0.551 (0.434, 0.669)	-0.179 (-0.349, -0.009)
Region 3	-0.005 (-0.200, 0.189)	0.054 (-0.011, 0.119)	0.156 (0.031, 0.280)
Primary Education	-0.993 (-1.007, -0.978)	-0.952 (-0.965, -0.939)	-0.950 (-1.020, -0.880)
Secondary Education	-0.630 (-0.691, -0.568)	-0.441 (-0.483, -0.400)	-0.336 (-0.413, -0.259)
Tertiary Education	4.732 (4.465, 4.999)	2.566 (2.489, 2.642)	3.535 (3.014, 4.056)
Single	0.864 (0.724, 1.003)	0.682 (0.618, 0.747)	0.153 (0.053, 0.253)
Living Together	-0.793 (-0.904, -0.683)	0.401 (0.215, 0.586)	0.079 (-0.237, 0.395)
Married	-0.222 (-0.314, -0.130)	-0.274 (-0.308, -0.239)	-0.307 (-0.493, -0.121)
Divorced	-0.812 (-0.942, -0.681)	-0.692 (-0.754, -0.631)	-0.387 (-0.772, -0.002)
Muslim	-0.150 (-0.322, 0.022)	-0.662 (-0.742, -0.583)	NA
Christ	0.092 (0.043, 0.141)	0.107 (0.098, 0.116)	NA
No Religion	-0.469 (-0.797, -0.141)	-0.601 (-0.750, -0.451)	NA
Other Religion	-0.721 (-0.996, -0.446)	-0.798 (-0.900, -0.695)	NA
Heard of HIV	-0.003 (-0.020, 0.015)	-0.022 (-0.028, -0.016)	0.014 (-0.004, 0.031)
HIV-Mosquite	-0.045 (-0.138, 0.048)	0.050 (0.040, 0.060)	NA
HIV-Condom	0.060 (>0.000, 0.119)	0.103 (0.093, 0.113)	NA
HIV-Look Healthy	0.182 (0.118, 0.247)	0.108 (0.100, 0.117)	NA
Daily Internet Use	0.284 (0.253, 0.314)	0.270 (0.254, 0.285)	0.235 (0.172, 0.298)
Average Error	0.561	0.477	0.467
RANK	3	2	1
N	111 - 494	1890 - 2552	275

Note. Bias; difference in proportions for dichotome variables; confidence intervals (in paranthesis) where measured by bootstrapping with 2000 iterations; N varries between variables and surveys due to missing information.

Exercise 3

Install sampcompR from CRAN (install.packages(sampcompR))

Use sampcompR to estimate bias in univariate estimates for survey 1 and survey 2 (or for your own data if you like).

Plot the results and output them as a table.

How would you interpret the results?

Exercise 3 – Comparison against Population Values

```
## Generate Results
diff_pop_mic <-  uni_compare(nboots = 0 funct= "d_prop",
                             plot_title = "comparison of survey 1 respondents and nonrespondents",
                             dfs = c("prob_resp", "nonprob_part"),
                             summetric = "avg2",
                             benchmarks=c("benchmark"),
                             variables=c("age_16to29", "age_30to39" ,"age_40to49" ,
                                         "age_50to59", "age_60p", "german",
                                         "female"))

## Plot Results
plot_diff_pop_mic<-plot_uni_compare(diff_pop_mic)
plot_diff_pop_mic
```

```
age_16to29 age_30to39 age_40to49 age_50to59    age_60p    edu_low    edu_mid    edu_hi
      0.21       0.17       0.17       0.22       0.23       0.31       0.34       0.
    female     german
      0.49       0.87
```

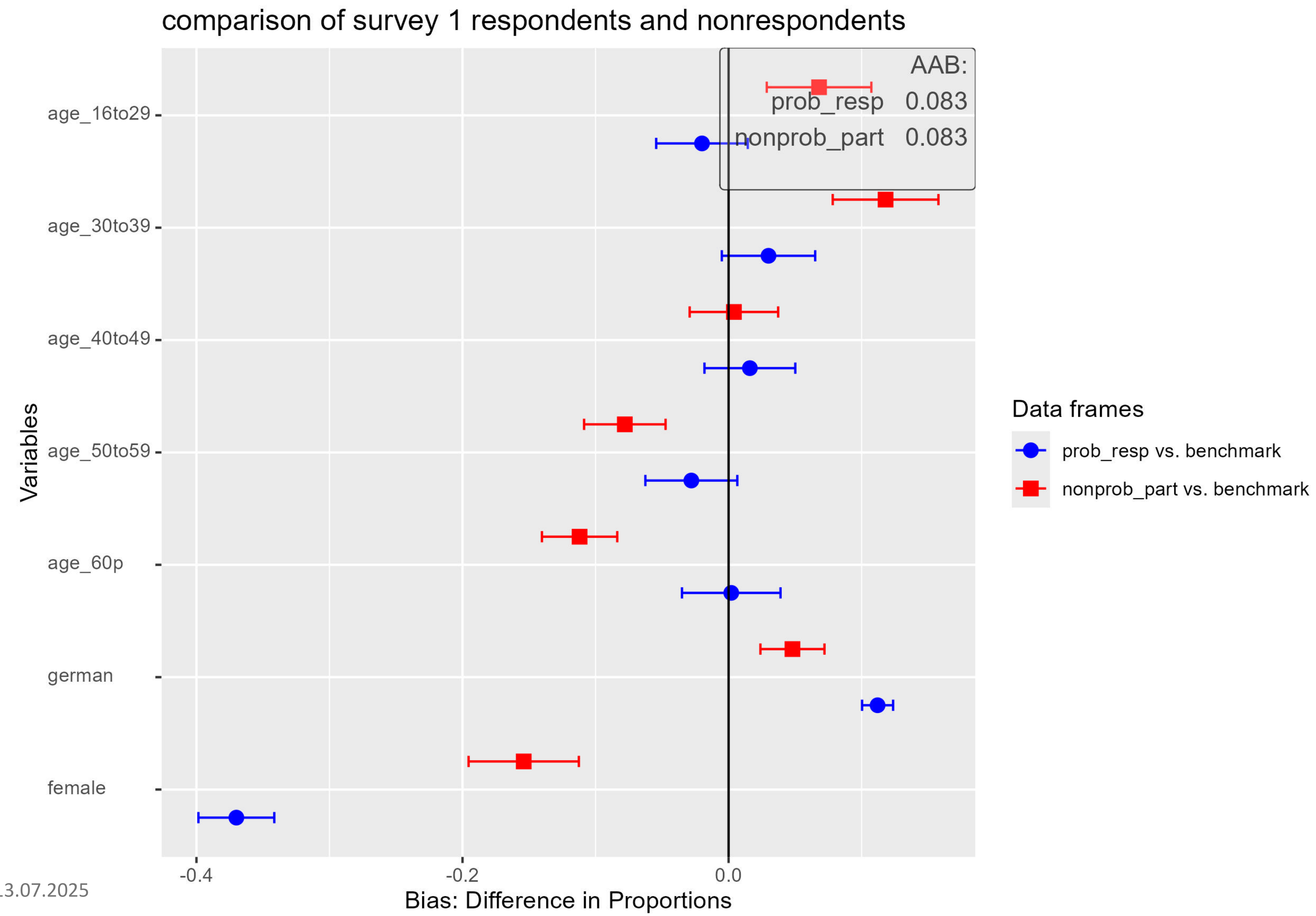
First, we compare nonrepsonse bias between demographics of the probability and nonprobability survey using population benchmarks. The population values must be given as a named vector.

Nonresponse bias is estimated as difference in proportions for dummy variables.

$$\widehat{\boldsymbol{P}}_{\boldsymbol{surv}} - \boldsymbol{P}_{\boldsymbol{bench}},$$

where $\widehat{\boldsymbol{P}}_{\boldsymbol{surv}}$ is the estimated survey proportion of respondents in the respective category (e.g., female), and $\boldsymbol{P}_{\boldsymbol{bench}}$ is the respective population value.

Exercise 3 – Comparison against Population Values



- The estimates of probability survey and nonprobability survey are similarly biased du to nonresponse on average.*
- For the probability survey gender and citizenship are most biased.
- For the non-probability survey age is also rather biased.

* Remember that this is only simulated data and could be very different in reality.

Exercise 3 – Comparison against Benchmark Survey Estimates

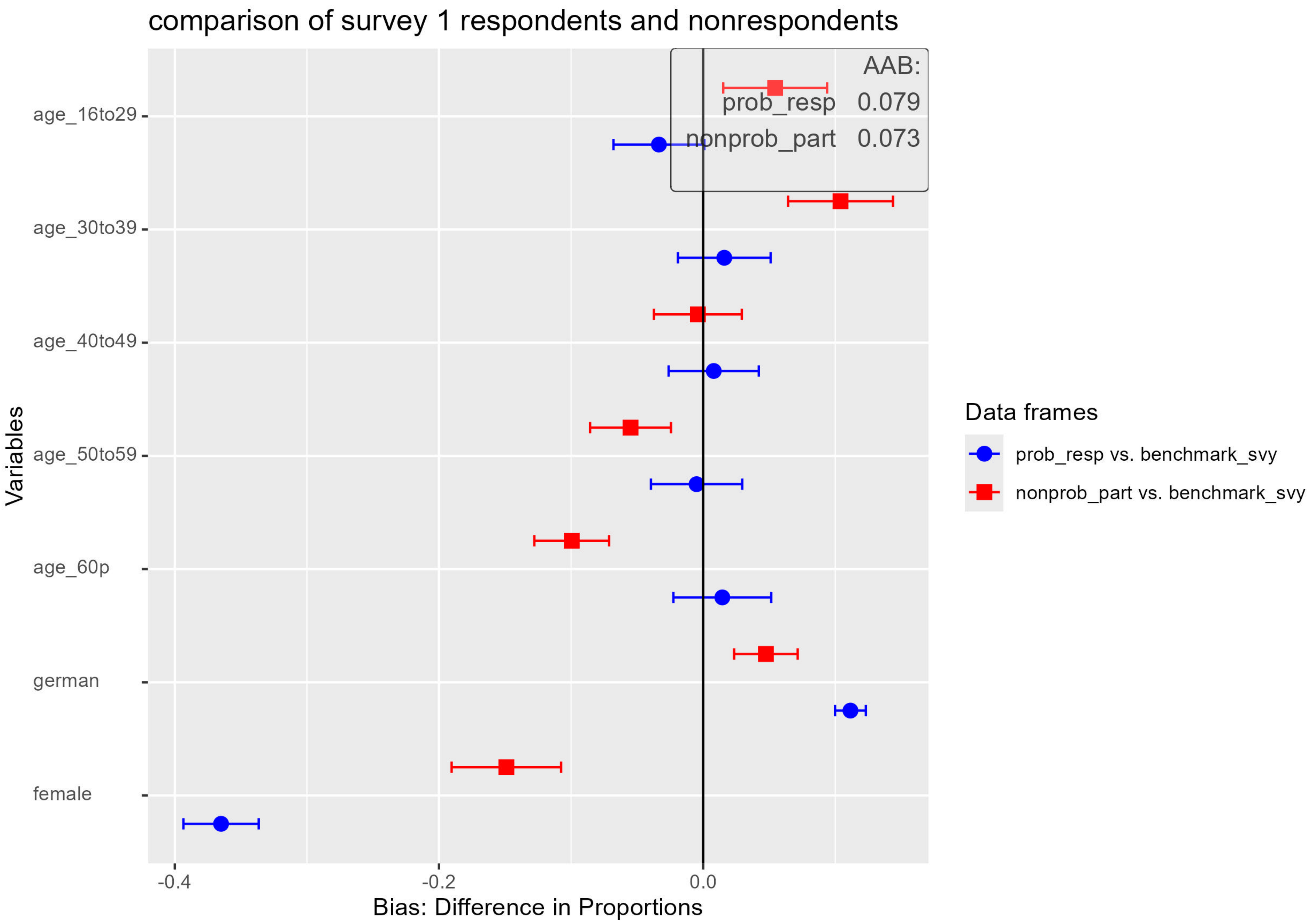
```
diff_pop_svy <-  uni_compare(nboots = 0,funct= "d_prop",
                             plot_title = "comparison of survey 1 respondents and nonrespondents",
                             dfs = c("prob_resp", "nonprob_part"),
                             summetric = "avg2",
                             benchmarks=c("benchmark_svy"),
                             variables=c("age_16to29", "age_30to39" ,"age_40to49" ,
                                         "age_50to59", "age_60p", "german",
                                         "female"))
```

```
uni_diff_table<-uni_compare_table(diff_pop_svy)
uni_diff_table
```

uni_compare_table can be used to output the results as table.

- As an alternative to population values, use other high-quality surveys as a benchmark.
- Those surveys should arguably be of very high quality.
- Example surveys could be a microcensus or the GSS.
- For instance, in this example the benchmark survey is another survey with no nonresponse.

Exercise 3 – Comparison against Benchmark Survey Estimates



- The results are similar to the results using benchmarks.
- The benchmark survey suffers from random sampling errors affecting confidence intervals.

GESIS Leibniz Institute for the Social Sciences

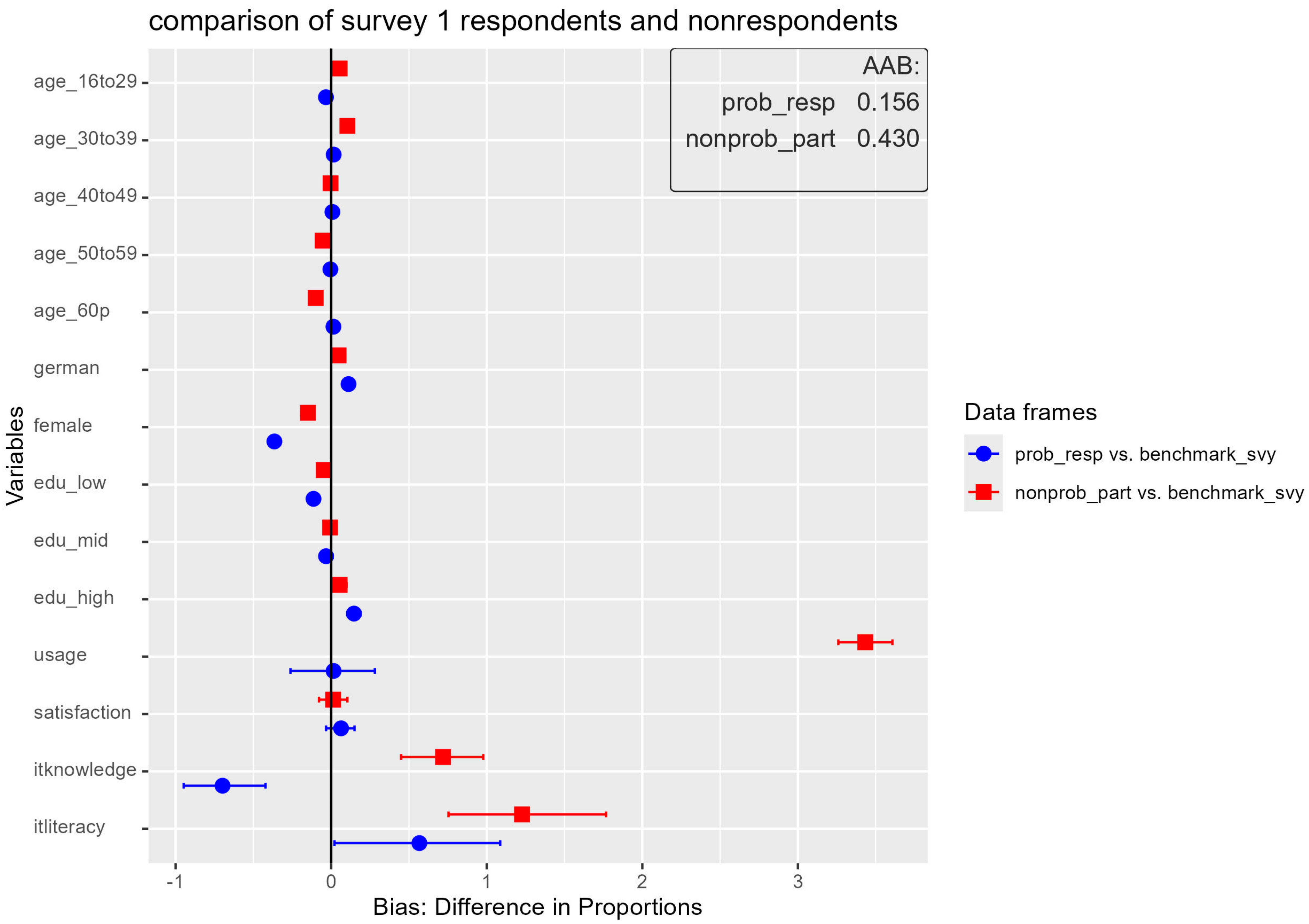
Exercise 3 – Comparison against Benchmark Survey Estimates with bootstrap CIs

```
diff_pop_boot <-  uni_compare(nboots = 1000,funct= "d_prop",
                              plot_title = "comparison of survey 1 respondents and nonrespondents",
                              dfs = c("prob_resp", "nonprob_part"),
                              summetric = "avg2",
                              benchmarks=c("benchmark_svy"),
                              variables=c("age_16to29", "age_30to39" ,"age_40to49" ,
                                          "age_50to59", "age_60p", "german" ,
                                          "female","edu_low","edu_mid","edu_high",
                                          "usage","satisfaction","itknowledge","itliteracy"))
```

- Standard confidence intervals are not well suited for nonprobability surveys (McPhee et al 2022).
- One recommended alternative is bootstrap confidence intervals.
- Bootstrap Cis are more computationally heavy to estimate but can be done with sampcompR.
- This is done by setting the nboots parameter to the number of bootstrap intervals required. It is often recommended to use more than 1000 intervals.

McPhee, Cameron, Frances Barlas, Nancy Brigham, Jill Darling, David Dutwin, Chris Jackson, Mickey Jackson, Ashley Kirzinger, Roderick Little, Emily Lorenz, Jenny Marlar, Andrew Mercer, Paul J. Scanlon, Steffen Weiss, and Laura Wronski. 2022. *Data Quality Metrics for Online Samples: Considerations for Study Design and Analysis*. *AAPOR Task Force Report*. Available at https://aapor.org/wp-content/uploads/2023/02/Task-Force-Report-FINAL.pdf.

Exercise 3 – Comparison against Benchmark Survey Estimates with bootstrap CIs



- Bootstrapping lead to the same nonresponse bias estimates , but confidence intervals might be different
- In this plot we also added non-frame variables to the comparison.
- Nonresponse bias for the continuous substantive variables is estimated as difference in means.
- This makes interpretation of AAB difficult, as the substantive variables have a higher impact.
- Use relative bias instead.

GESIS Leibniz Institute for the Social Sciences

Exercise 3 – Comparison against Benchmark Survey Estimates with relative differences

```
uni_diff_table<-uni_compare_table(diff_pop_svy)
uni_diff_table
```

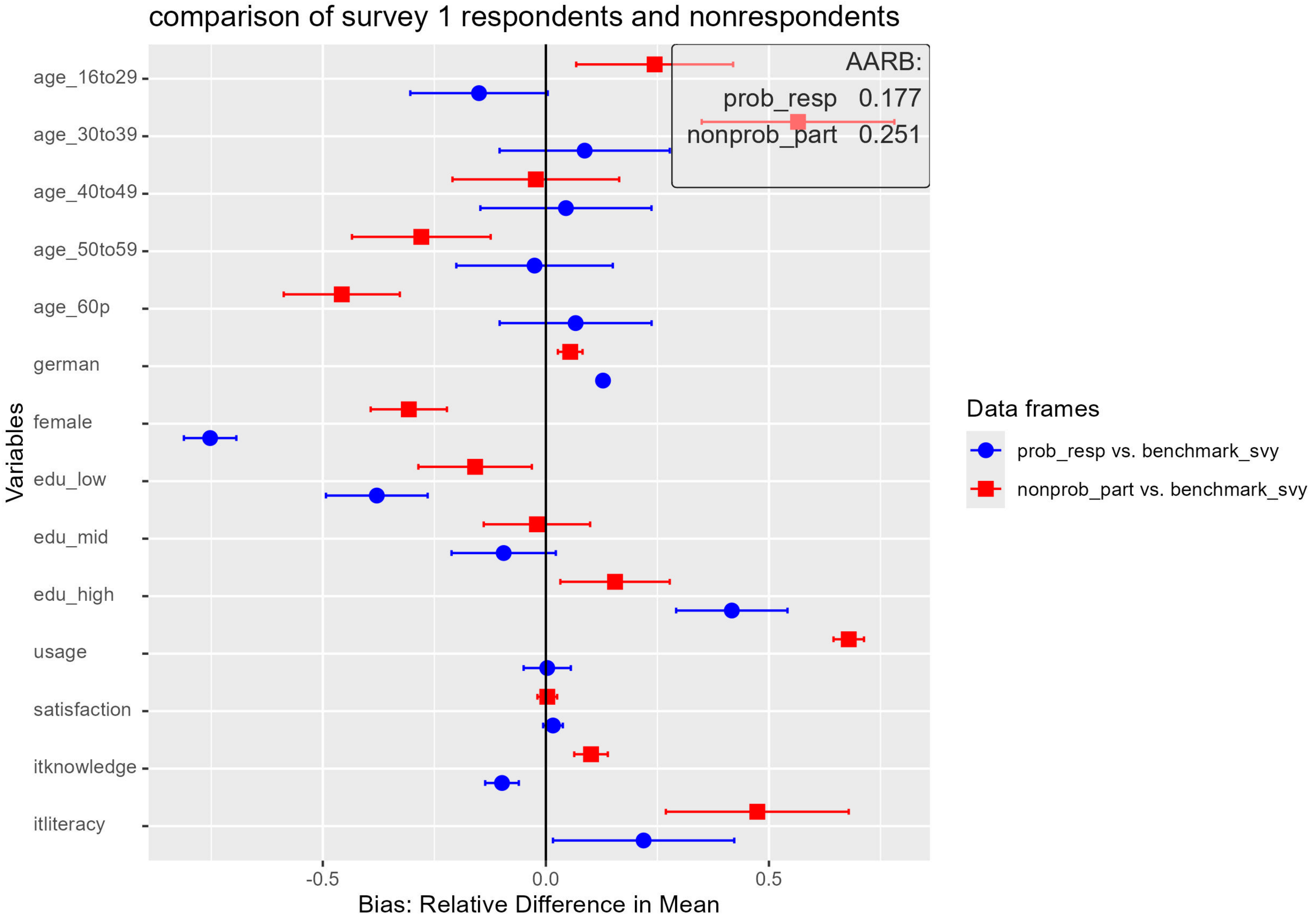
```
diff_pop_rel <-  uni_compare(nboots = 0,funct= "rel_mean",
                             plot_title = "comparison of survey 1 respondents and nonrespondents",
                             dfs = c("prob_resp", "nonprob_part"),
                             summetric = "avg2",
                             benchmarks=c("benchmark_svy"),
                             variables=c("age_16to29", "age_30to39" ,"age_40to49" ,
                                         "age_50to59", "age_60p", "german" ,
                                         "female","edu_low","edu_mid","edu_high",
                                         "usage","satisfaction","itknowledge","itliteracy"))
```

- Relative difference in means is used as function.
- This leads to a relative difference in proportions for dummy variables and a relative difference in mean for continuous variables. It also affects the default name of the X-axis in the plot.

$$\text{Rel Bias} = \frac{\overline{\widehat{Variable}}_{surv} - \overline{\widehat{Variable}}_{bench}}{\overline{\widehat{Variable}}_{bench}}$$

Where, $\overline{\widehat{Variable}}_{surv}$ is the estimated mean of a variable in the survey and $\overline{\widehat{Variable}}_{bench}$ is the estimated mean of a variable in a benchmark survey.

Exercise 3 – Comparison against Benchmark Survey Estimates with bootstrap CIs



- In this plot nonresponse biases of different variables are comparable with each other.
- All variables have a similar impact on the average bias.
- The average noresponse bias is now called AARB which means Absolute Average Relative Bias
- We also see that IT-literacy is biased although it showed an unsignificant coefficient in the response model.
- This is because in the response model bias was also controlled for education, which is correlated to it-litearcy.

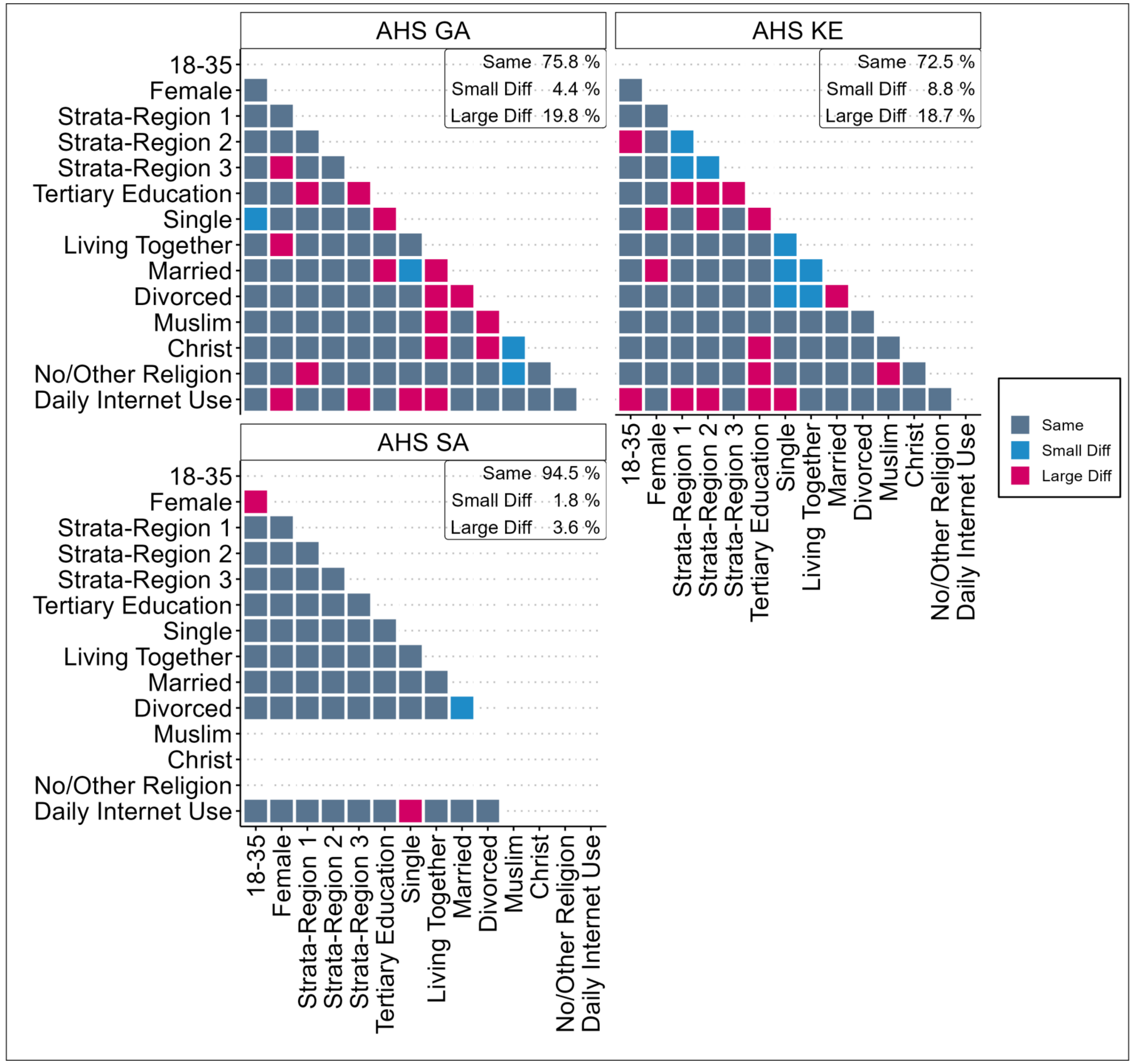
Exercise 4: Bias in bivariate estimates

Example Bivariate Comparison

Bivariate Plot

```
biv_AHS_weighted_DHS <- biv_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                                    benchmarks = c("DHS_Ga",
                                                   "DHS_Ke",
                                                   "DHS_Sa"),
                                    variables = variables2,
                                    plots_label = c("AHS GA",
                                                    "AHS KE",
                                                    "AHS SA"),
                                    p_adjust="bonferroni",
                                    weight_bench = c("weight"),
                                    id_bench = c("cluster"),
                                    strata_bench = c("strata"),
                                    varlabels = varlabels2, nboots =2000,
                                    colors = colors)

biv_plot<-plot_biv_compare(biv_AHS_weighted_DHS,ncol_facet = 2)
```



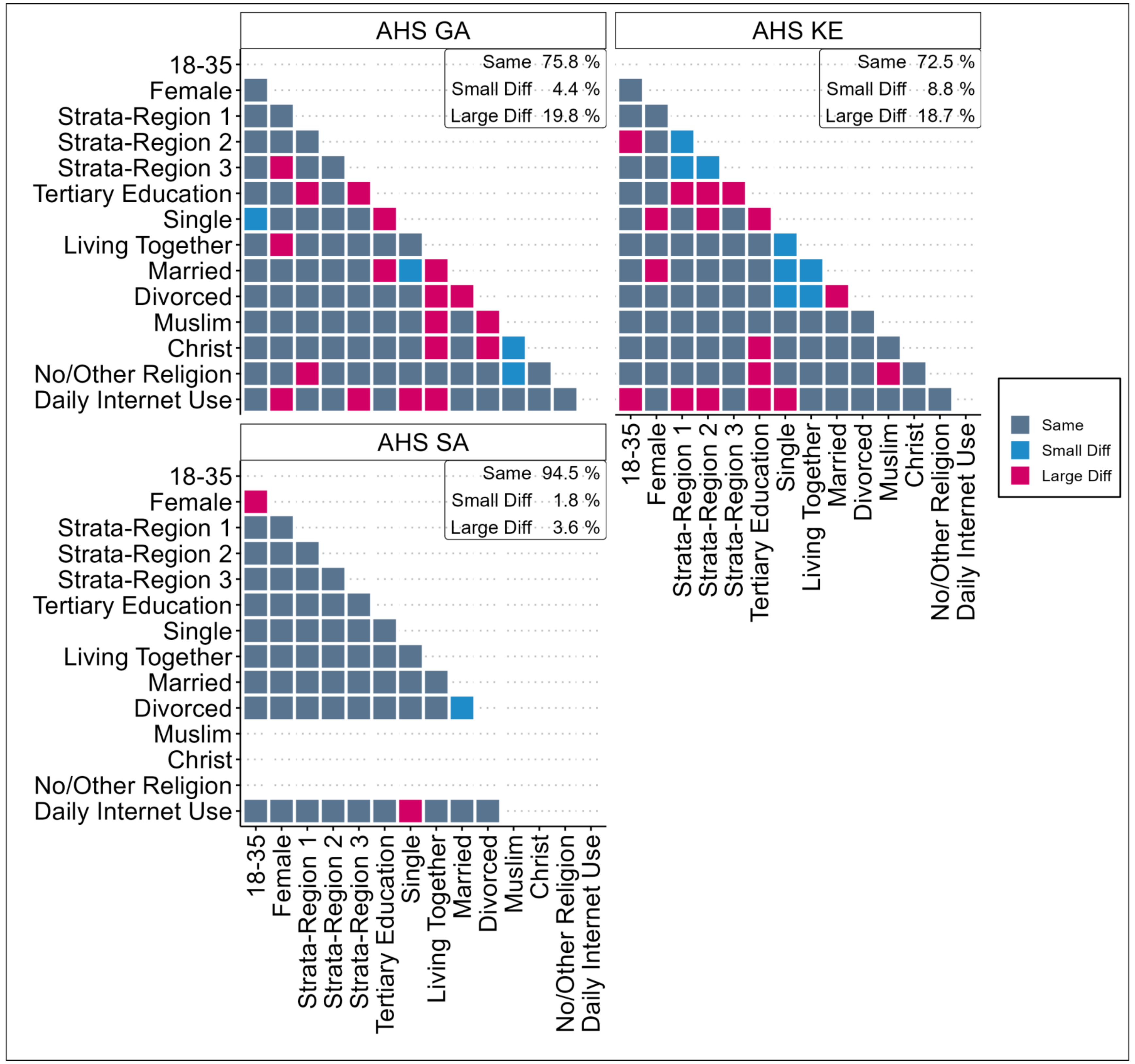
Example Bivariate Comparison

Bivariate Plot

```
biv_AHS_weighted_DHS <- biv_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                                    benchmarks = c("DHS_Ga",
                                                   "DHS_Ke",
                                                   "DHS_Sa"),
                                    variables = variables2,
                                    plots_label = c("AHS GA",
                                                    "AHS KE",
                                                    "AHS SA"),
                                    p_adjust="bonferroni",
                                    weight_bench = c("weight"),
                                    id_bench = c("cluster"),
                                    strata_bench = c("strata"),
                                    varlabels = varlabels2, nboots =2000,
                                    colors = colors)

biv_plot<-plot_biv_compare(biv_AHS_weighted_DHS,ncol_facet = 2)
```

Similar inputs to the function as before.



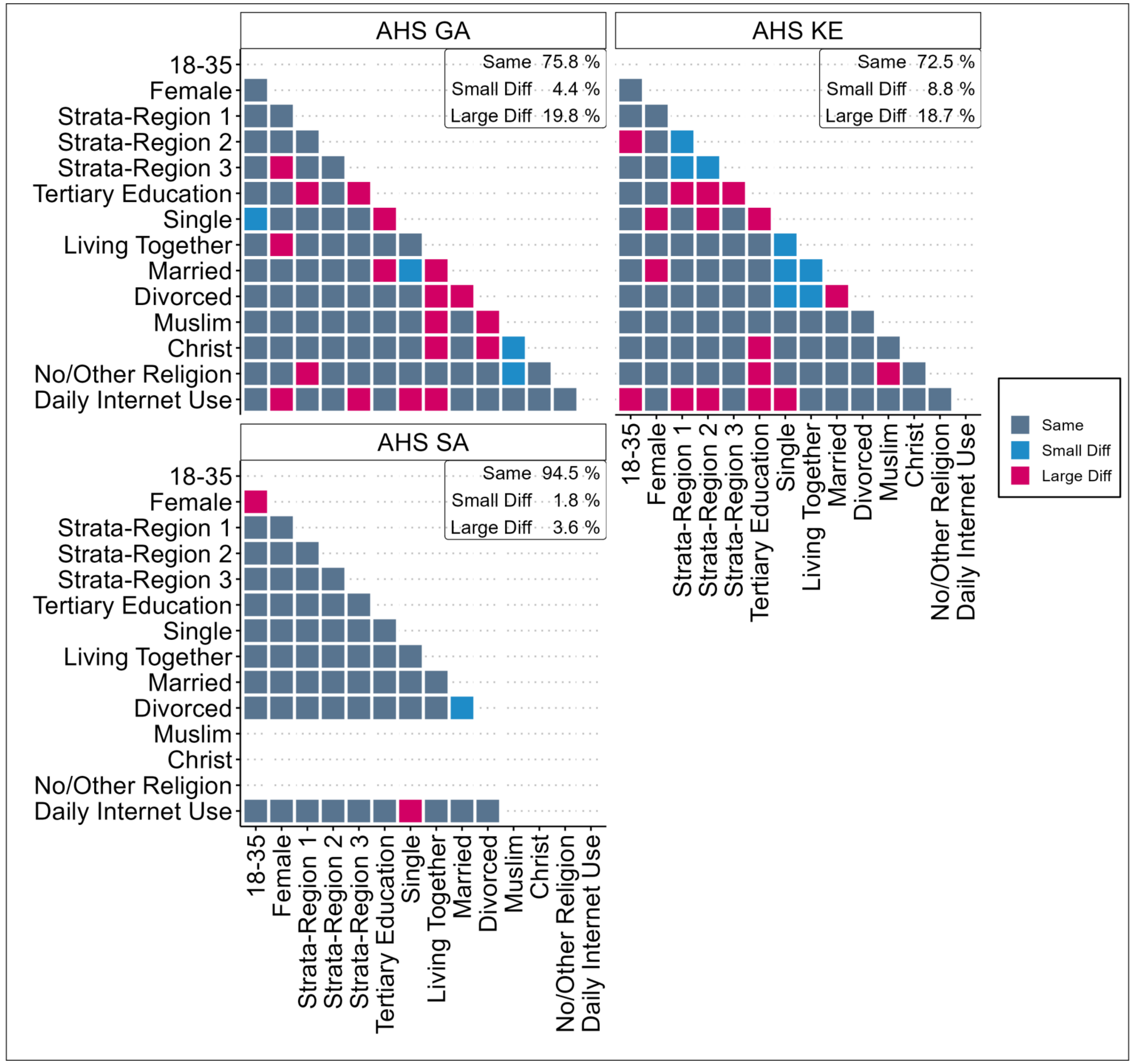
Example Bivariate Comparison

Bivariate Plot

```
biv_AHS_weighted_DHS <- biv_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                                    benchmarks = c("DHS_Ga",
                                                   "DHS_Ke",
                                                   "DHS_Sa"),
                                    variables = variables2,
                                    plots_label = c("AHS GA",
                                                    "AHS KE",
                                                    "AHS SA"),
                                    p_adjust="bonferroni",
                                    weight_bench = c("weight"),
                                    id_bench = c("cluster"),
                                    strata_bench = c("strata"),
                                    varlabels = varlabels2, nboots =2000,
                                    colors = colors)

biv_plot<-plot_biv_compare(biv_AHS_weighted_DHS,ncol_facet = 2)
```

1. Pearson’s R is calculated for every pair of variables in surveys and benchmarks.
2. Pearson’s Rs are compared against each other to evaluate the proportion of significantly different correlations.



Example Bivariate Comparison

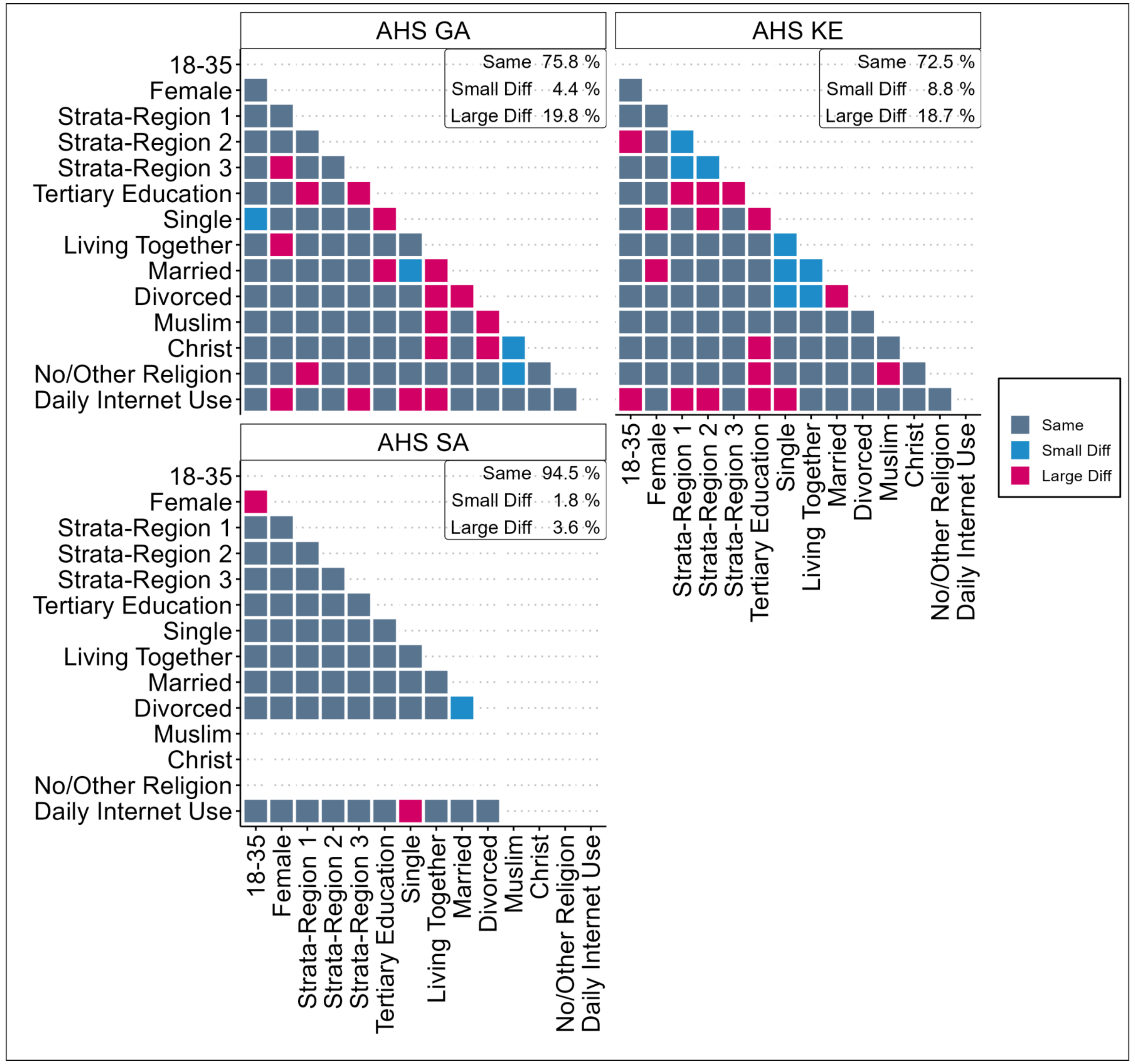
Bivariate Plot

```
biv_AHS_weighted_DHS <- biv_compare(dfs=c("AHS_Ga","AHS_Ke","AHS_Sa"),
                                    benchmarks = c("DHS_Ga",
                                                   "DHS_Ke",
                                                   "DHS_Sa"),
                                    variables = variables2,
                                    plots_label = c("AHS GA",
                                                    "AHS KE",
                                                    "AHS SA"),
                                    p_adjust="bonferroni",
                                    weight_bench = c("weight"),
                                    id_bench = c("cluster"),
                                    strata_bench = c("strata"),
                                    varlabels = varlabels2, nboots =2000,
                                    colors = colors)

biv_plot<-plot_biv_compare(biv_AHS_weighted_DHS,ncol_facet = 2)
```

Same = Pearson’s R is not significantly* different between surveys or no significant correlation found in survey or benchmark.
Small Diff = Pearson’s rRis significantly* different between survey and benchmark & one of them is significantly* different from 0.
Large Diff = “Small Diff” + one Pearson’s R is double the size of the other, or they differ in direction.

*p< 0.05



Exercise 4

Use sampcompR to estimate bias in bivariate estimates for survey 1 and survey 2 (or for your own data if you like).

Plot the results.

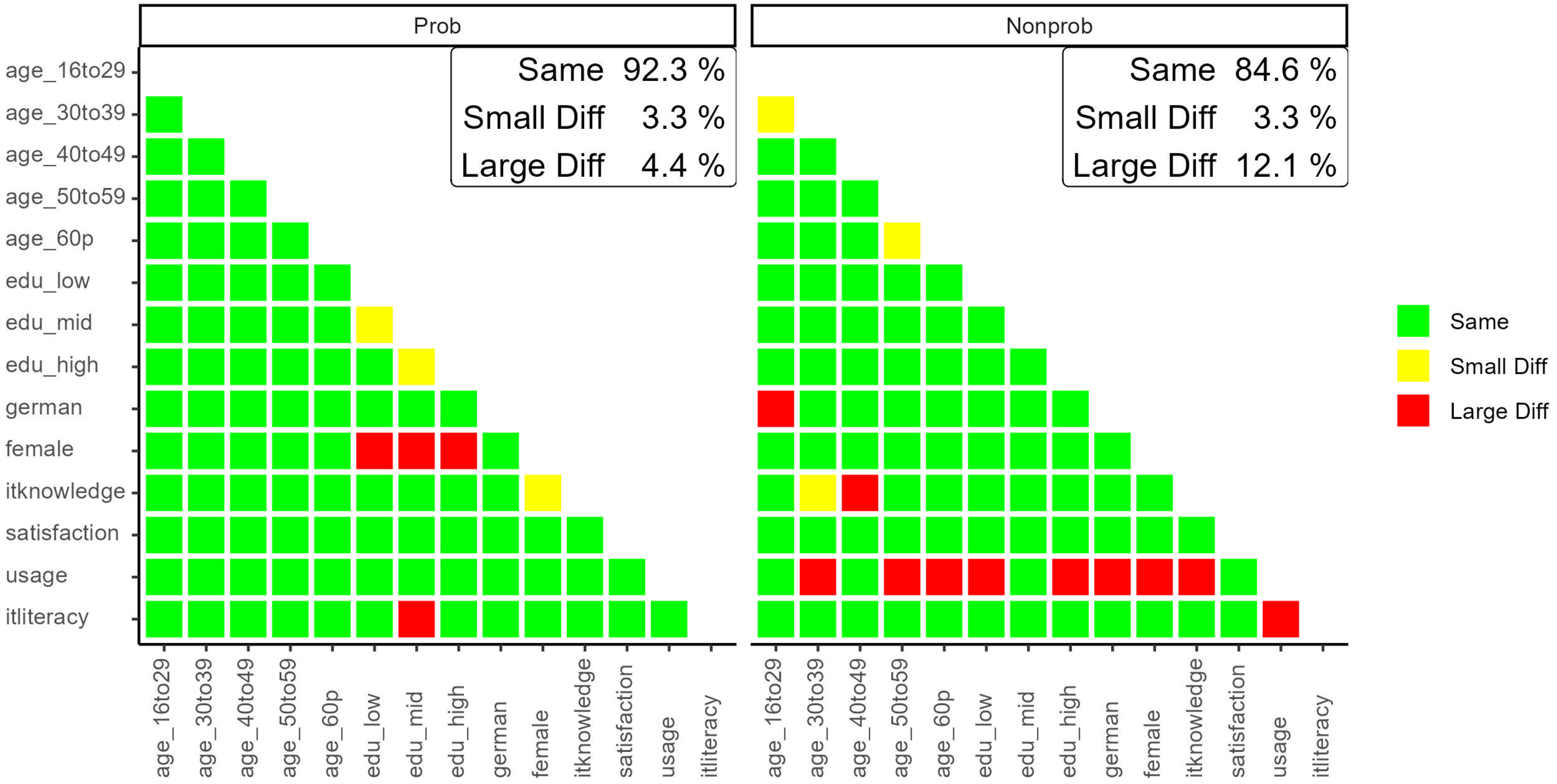
How would you interpret the results?

Exercise 4 – Bivariate Results

```
biv.diff <- biv_compare(nboots = 0, dfs = c("prob_resp", "nonprob_part"),
                        plot_title = "comparison of survey respondents and full sample",
           benchmarks=c("benchmark_svy"),
           diff_perc_size = 2,
           plots_label = c("Prob", "Nonprob"),
           variables=c("age_16to29", "age_30to39" ,"age_40to49" ,"age_50to59",
                       "age_60p","edu_low", "edu_mid", "edu_high","german" ,
                       "female", "itknowledge",
                       "satisfaction", "usage","itliteracy"))
```

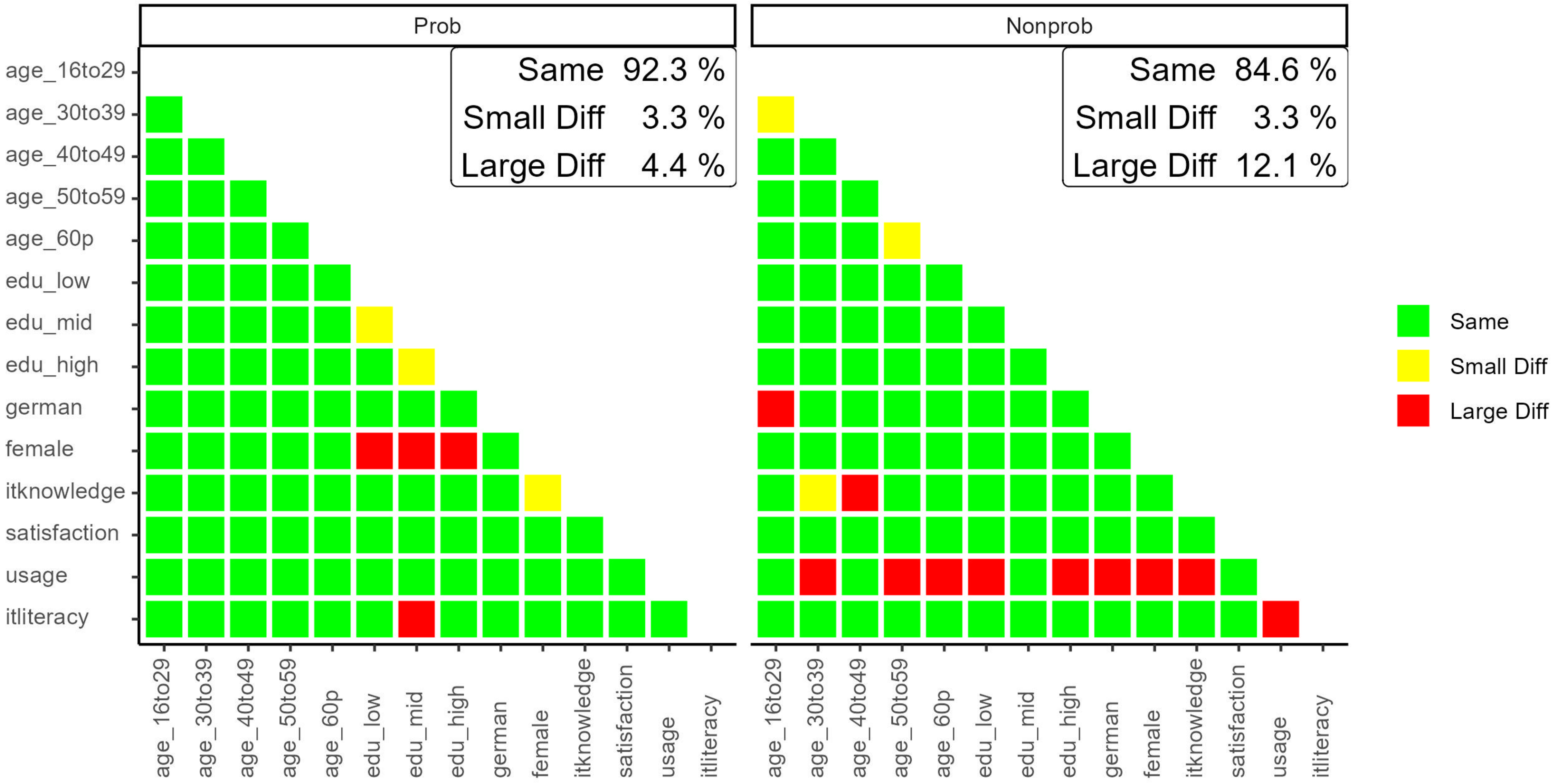
- For biv_compare the inputs are very similar to univariate inputs.
- For all variables pairwise correlations are estimated and correlations are compared between surveys.
- By default, correlations are Pearson's *R* correlations

Exercise 4 – Bivariate Results



- Same (green) means that there is no significant difference.
- Small Diff (yellow) means that there is a significant difference, but the difference is small, and the correlations show the same direction.
- Large Diff (red) means that the benchmark correlation estimates are half or double the surveys estimates, or that one is negative while the other is positive.

Exercise 4 – Bivariate Results



- In this example we see only few biased correlations for the probability survey.
- It-literacy shows a biased correlation to medium education, as well as the correlation of gender and education.
- For the nonprobability survey, internet usage is strongly biased for nearly any correlation.
- It-knowledge also shows biased correlations with gender (prob survey) and age (nonprob survey).

GESIS Leibniz Institute for the Social Sciences

Exercise 5: Bias in multivariate estimates

Leibniz Association

Example Multivariate Comparison

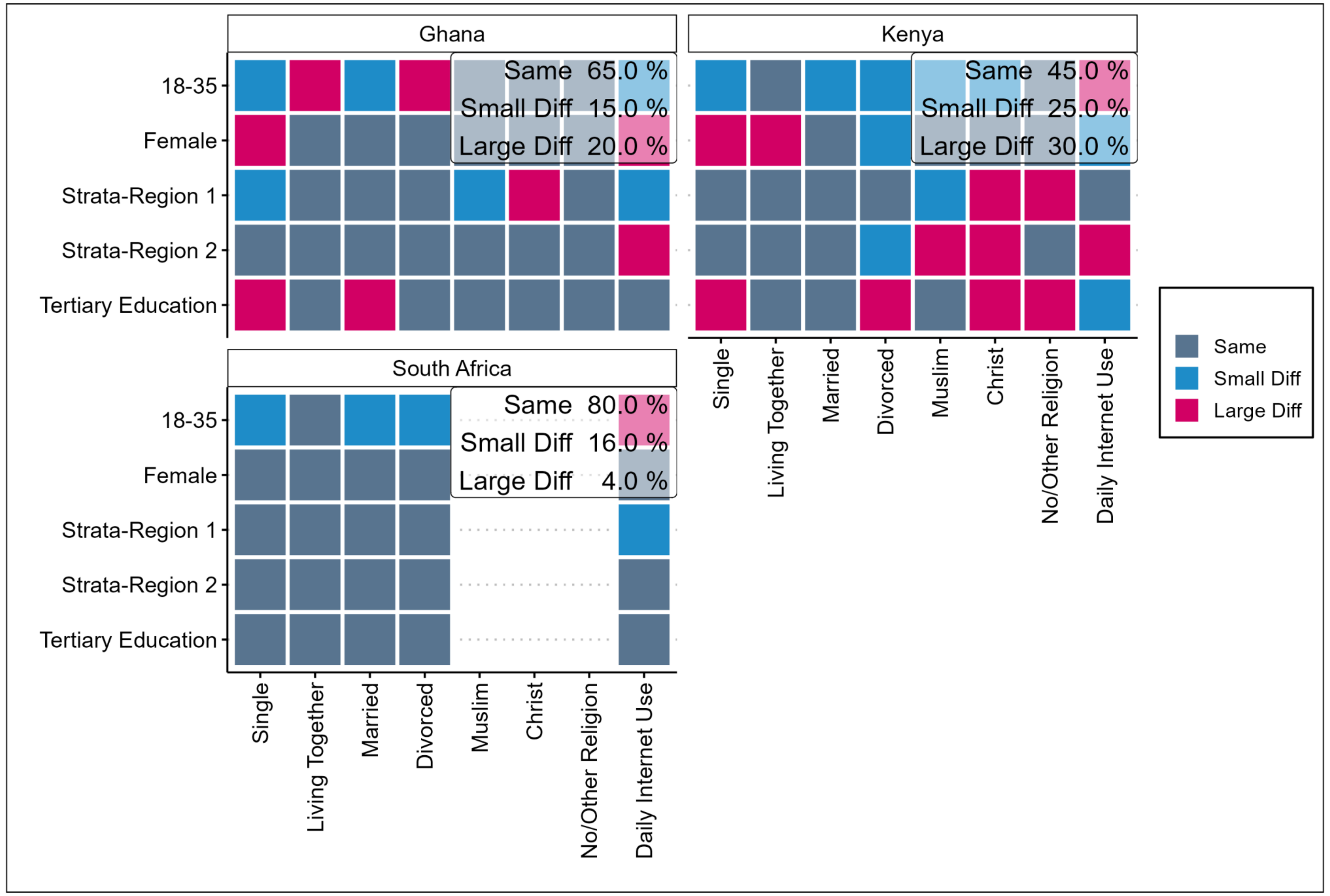
Example Multivariate Plot

```
multi_data_GA<-multi_compare(df = AHS_GA, benchmark = DHS_Ga,
                             dependent = dependent ,independent = independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_KE<-multi_compare(df = AHS_KE, benchmark = DHS_Ke,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_SA<-multi_compare(df = AHS_SA, benchmark = DHS_Sa,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")


multi_plots_Africa2<-
  plot_multi_compare(multi_compare_objects = c("multi_data_GA",
                                               "multi_data_KE",
                                               "multi_data_SA"),
                                        perc_diff_transparance = 0.5,
                                        label_x = xlab,
                                        label_y = ylab,
                                        missings_x = F,
                                        colors = colors)
```



Example Multivariate Comparison

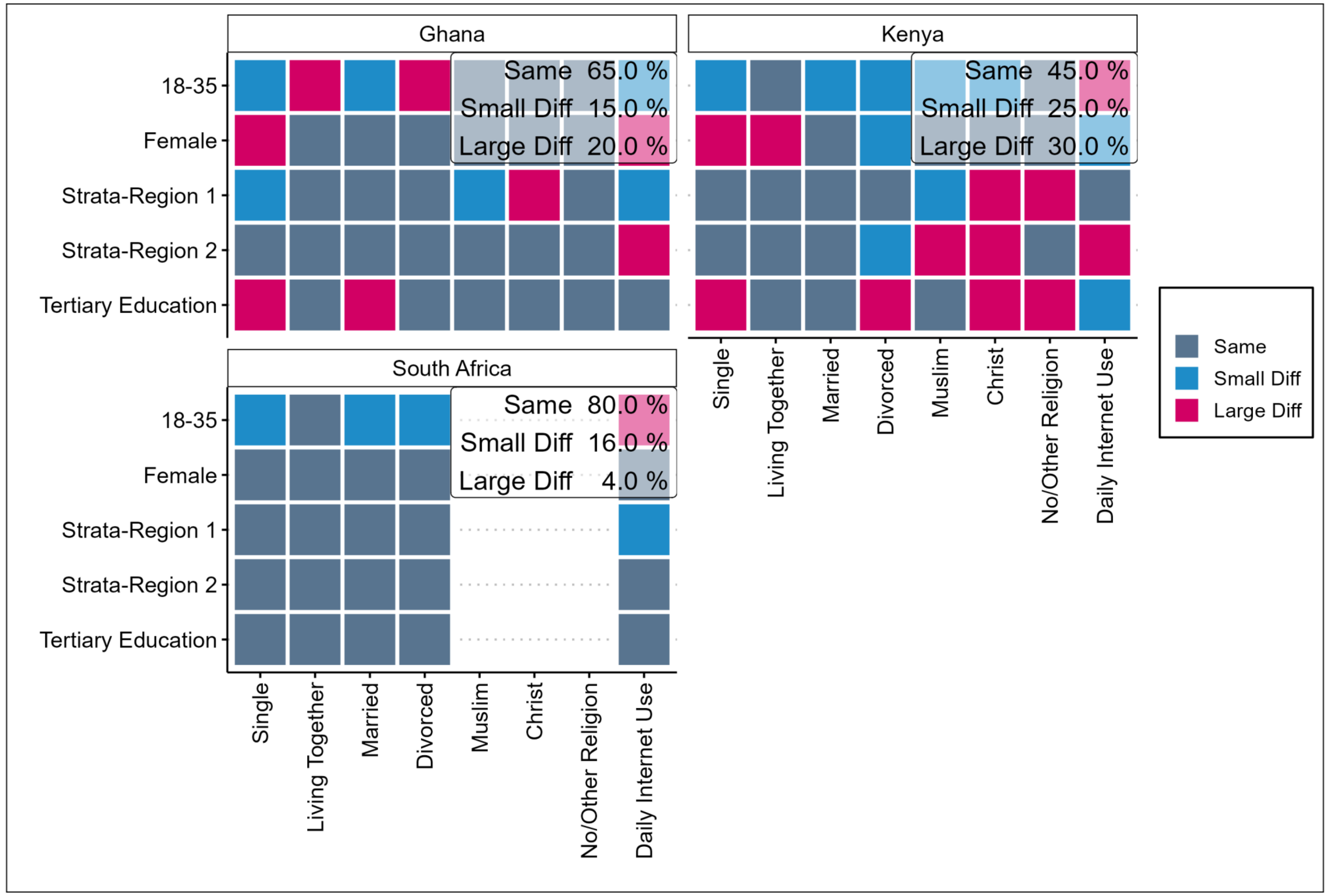
Example Multivariate Plot

```
multi_data_GA<-multi_compare(df = AHS_GA, benchmark = DHS_Ga,
                             dependent = dependent ,independent = independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_KE<-multi_compare(df = AHS_KE, benchmark = DHS_Ke,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")

multi_data_SA<-multi_compare(df = AHS_SA, benchmark = DHS_Sa,
                             dependent =dependent ,independent =independent,
                             family = "logit", p_adjust="bonferroni",nboots = 2000,
                             weight_bench = "weight", id_bench = "cluster",strata_bench="strata")


multi_plots_Africa2<-
  plot_multi_compare(multi_compare_objects = c("multi_data_GA",
                                               "multi_data_KE",
                                               "multi_data_SA"),
                     perc_diff_transparance = 0.5,
                     label_x = xlab,
                     label_y = ylab,
                     missings_x = F,
                     colors = colors)
```



- ❖In the Plot on the right, every model is a column, and every square is a coefficient.
- ❖The color again indicates the difference between benchmark and survey.

Exercise 5

Use sampcompR to estimate bias in multivariate estimates for Survey 1 and Survey 2 (or for your own data if you like).

Plot the results.

How would you interpret he results?

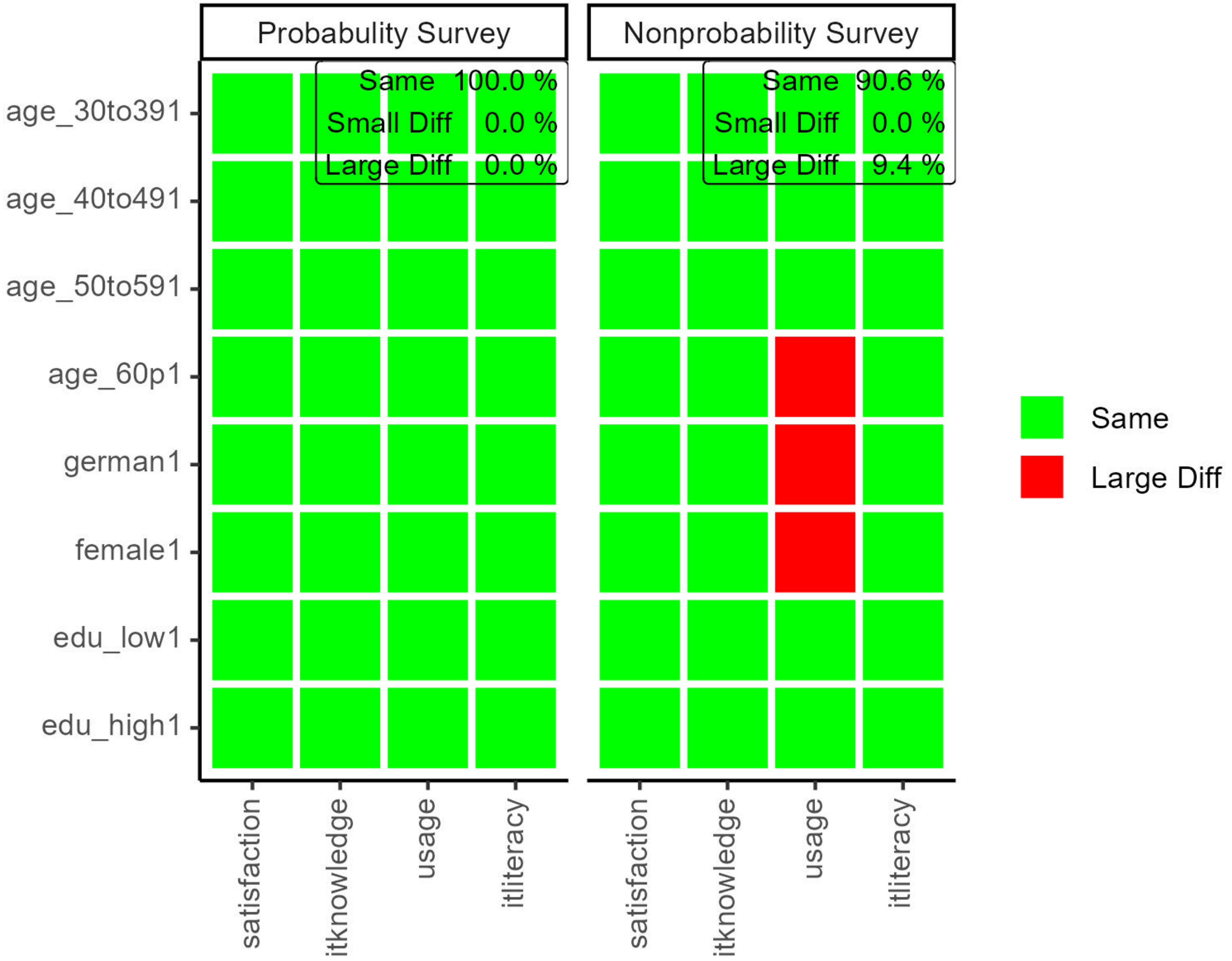
Exercise 5

```
## multivariate comparison for dataset 1
multi.diff.1 <-  multi_compare(df="prob_resp", bench="benchmark_svy",
              independent = c("age_30to39" ,"age_40to49" , "age_50to59",
                              "age_60p" ,"german","female",
                              "edu_low", "edu_high"),
              dependent=c("satisfaction", "itknowledge",
                          "usage","itliteracy"),
              family="ols")

## multivariate comparison for dataset 2
multi.diff.2 <- multi_compare(df="nonprob_part",
                              bench="benchmark_svy",
                              independent = c("age_30to39","age_40to49","age_50to59",
                                              "age_60p","german","female",
                                              "edu_low", "edu_high"),
                                             dependent=c("satisfaction", "itknowledge",
                                                         "usage","itliteracy"),
                                             family="ols")
```

- Do the comparison separate for both surveys.
- It will also print an info which variables are biased.
- Every glm() regression is possible
- If you ever encounter a situation where you need different models for different variables you can later combine multi_compare objects with multi_compare_merge() and plot them together

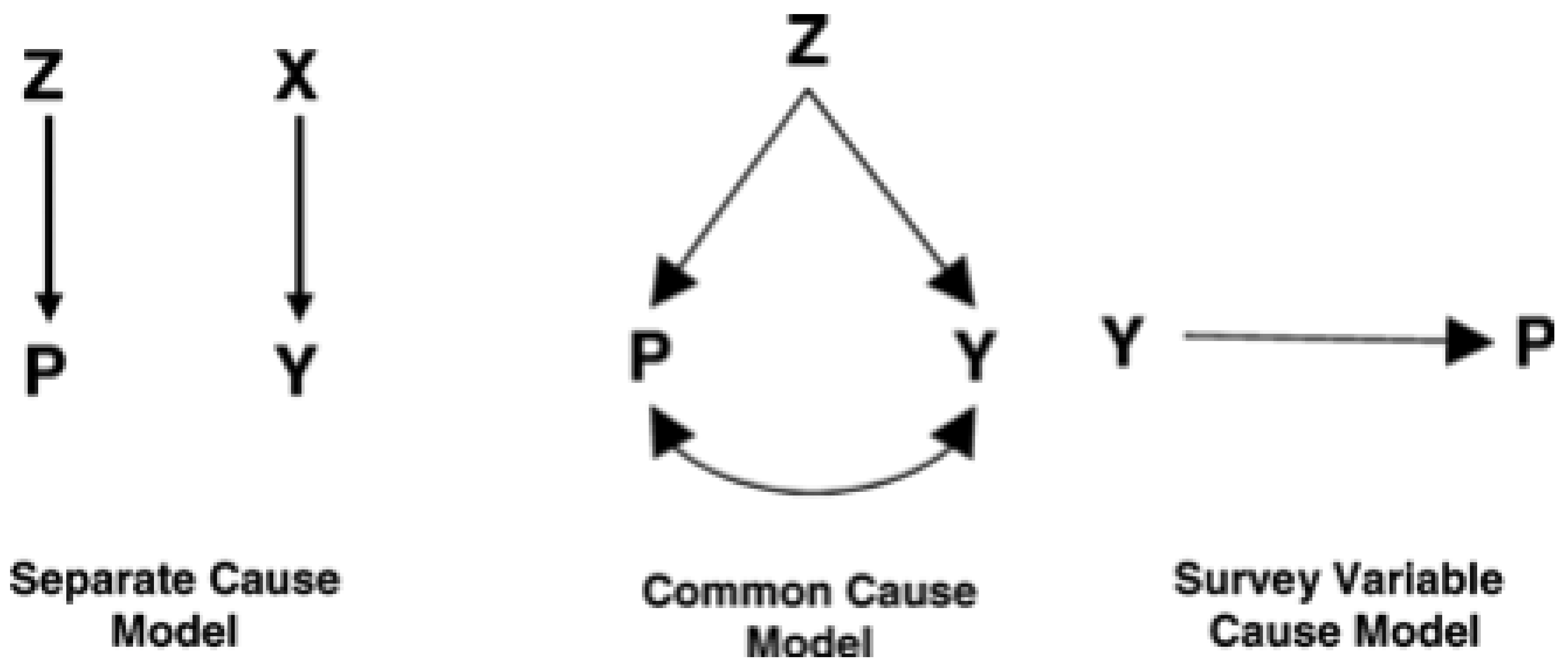
Exercise 5



- Both surveys show nearly no bias in multivariate estimates
- One exception is the model on internet use for the nonproability survey.
- Here the coefficients of the oldest age group, German citizenship and gender are highly biased.

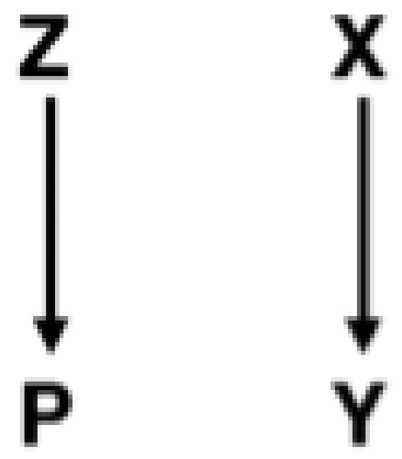
How does weighting correct bias?

When is weighting effective?



Robert M. Groves, Emilia Peytcheva, The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis Public Opinion Quarterly, Volume 72, Issue 2, Summer 2008, Pages 167–189, https://doi.org/10.1093/poq/nfn011

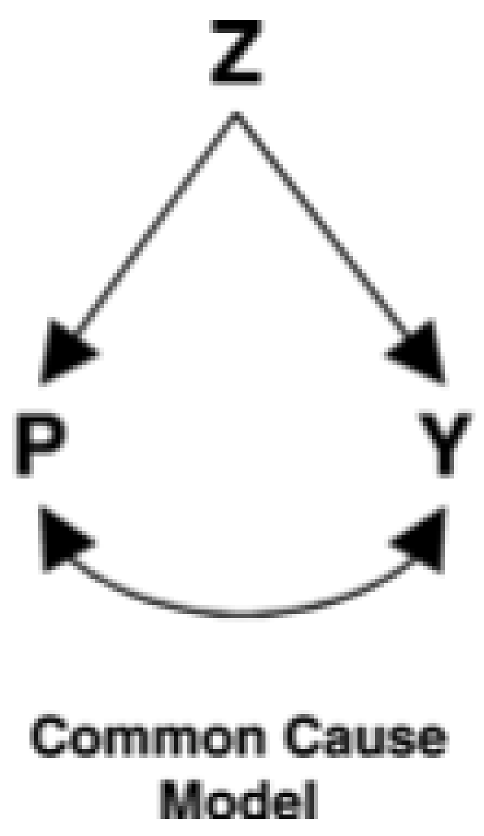
Separate Cause Model



Separate Cause Model

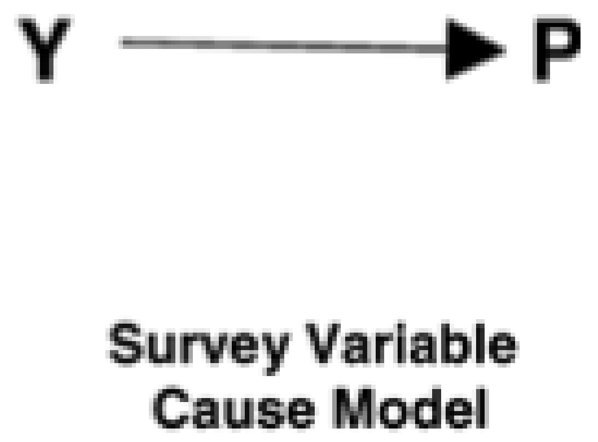
For this model, weighting is not necessary because estimates of Y are unbiased.

Separate Cause Model



- Weights should be included in the estimation process.
- Only fully removes bias if all **Z** variables are known and available (e.g., from a benchmark survey or official statistics)

Separate Cause Model

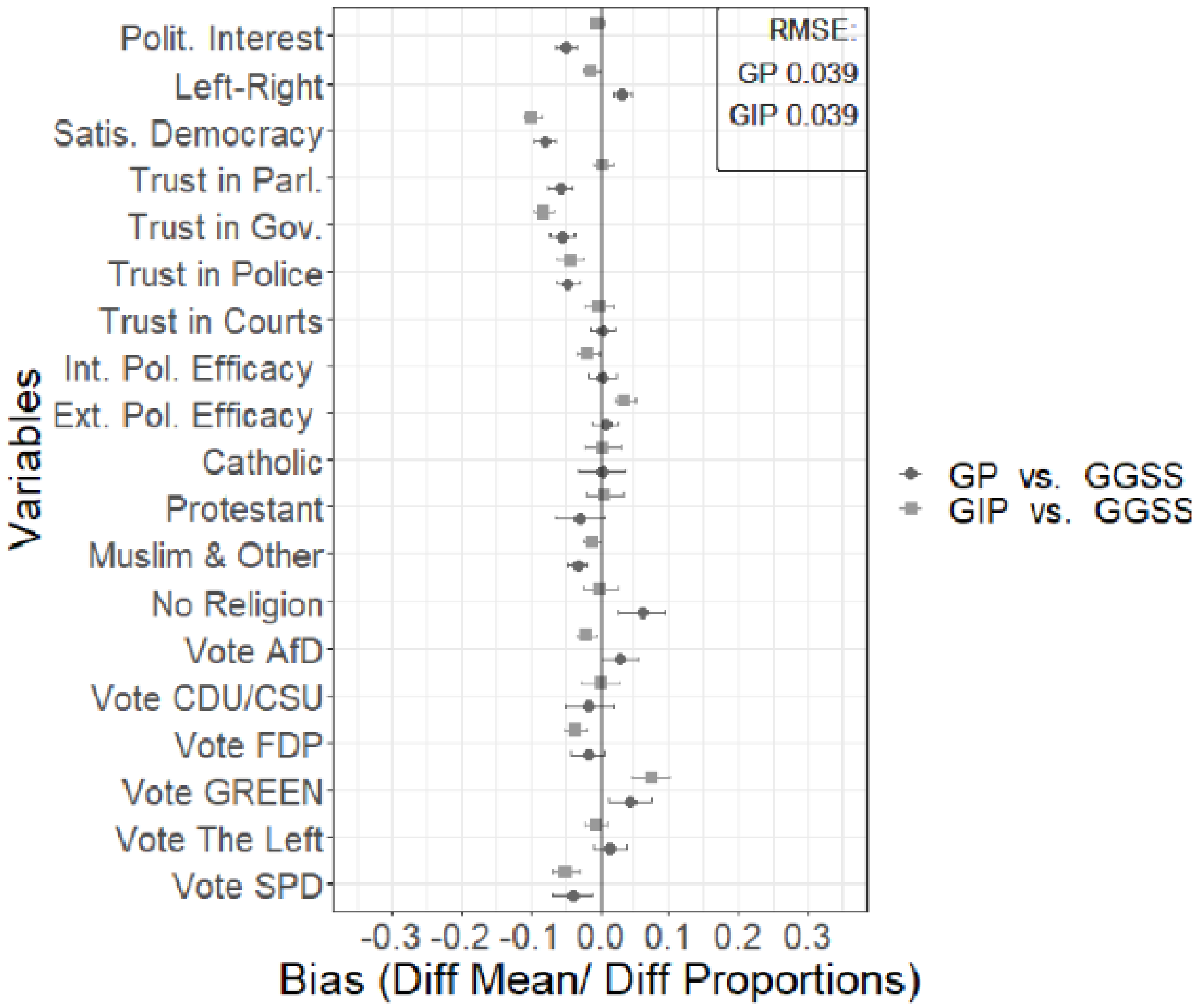


- Propensity of participation is affected by the variable of interest.
- High-quality estimates for the variable of interest are not known; otherwise, we would not have to measure it in our current survey.
- **<u>Weighting can not remove bias due to Y variables</u>**

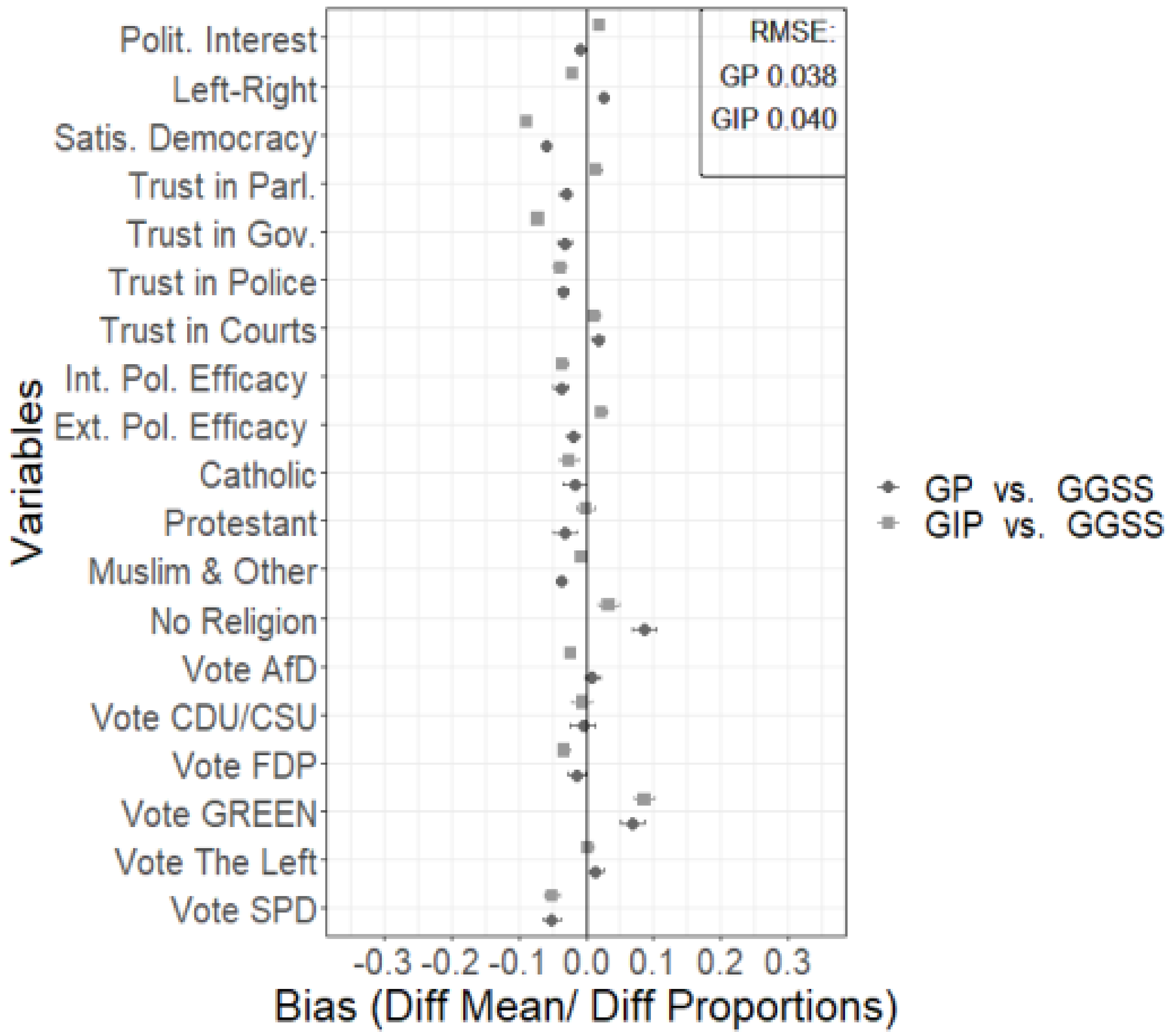
Common procedures include:

1. **Propensity weighting**
   1. Propensity scores are used to estimate weights
   2. Individual level information needed
2. **Poststratification**
   1. Joint distributions are used to estimate weights
   2. Aggregate information needed
   3. There are more complex developments for this method called Multilevel Regression and Poststratification (MRP)
3. **Raking**
   1. Marginal distributions are used to estimate Weights
   2. Further developments for applying MRP for Marginal Distribution information.

Weighting does not always reduce bias



with raking adjustment weights



without adjustment weights

GIP is more biased after weighting

GESIS Leibniz Institute for the Social Sciences

Exercise 6: Nonresponse weights

Leibniz Association

Exercise 6 – Nonresponse weights

Use the response models from Exercise 1 to calculate nonresponse weights as inverse of the fitted values.

Investigate the weights regarding their summary and standard deviation.

Use them in sampcompR to univestigate their impact on univariate bias. (take a look in the documentation)

Exercise 6 – Gaining Nonresponse Weights

```
## Nonresponse weight based on frame information
svy1.nr.weight.frame <- 1/svy1.response$fitted.values
summary(svy1.nr.weight.frame)
sd(svy1.nr.weight.frame)

## Nonresponse weight based also on unobserved variables
svy1.nr.weight.unobs <- 1/svy1.response.unobs$fitted.values
summary(svy1.nr.weight.unobs)
sd(svy1.nr.weight.unobs)

prob_resp<- prob_resp |>
  mutate(wgt1 = svy1.nr.weight.frame[survey1_prob$response==1],
         wgt2 = svy1.nr.weight.unobs[survey1_prob$response==1],
         id = 1:500)
```

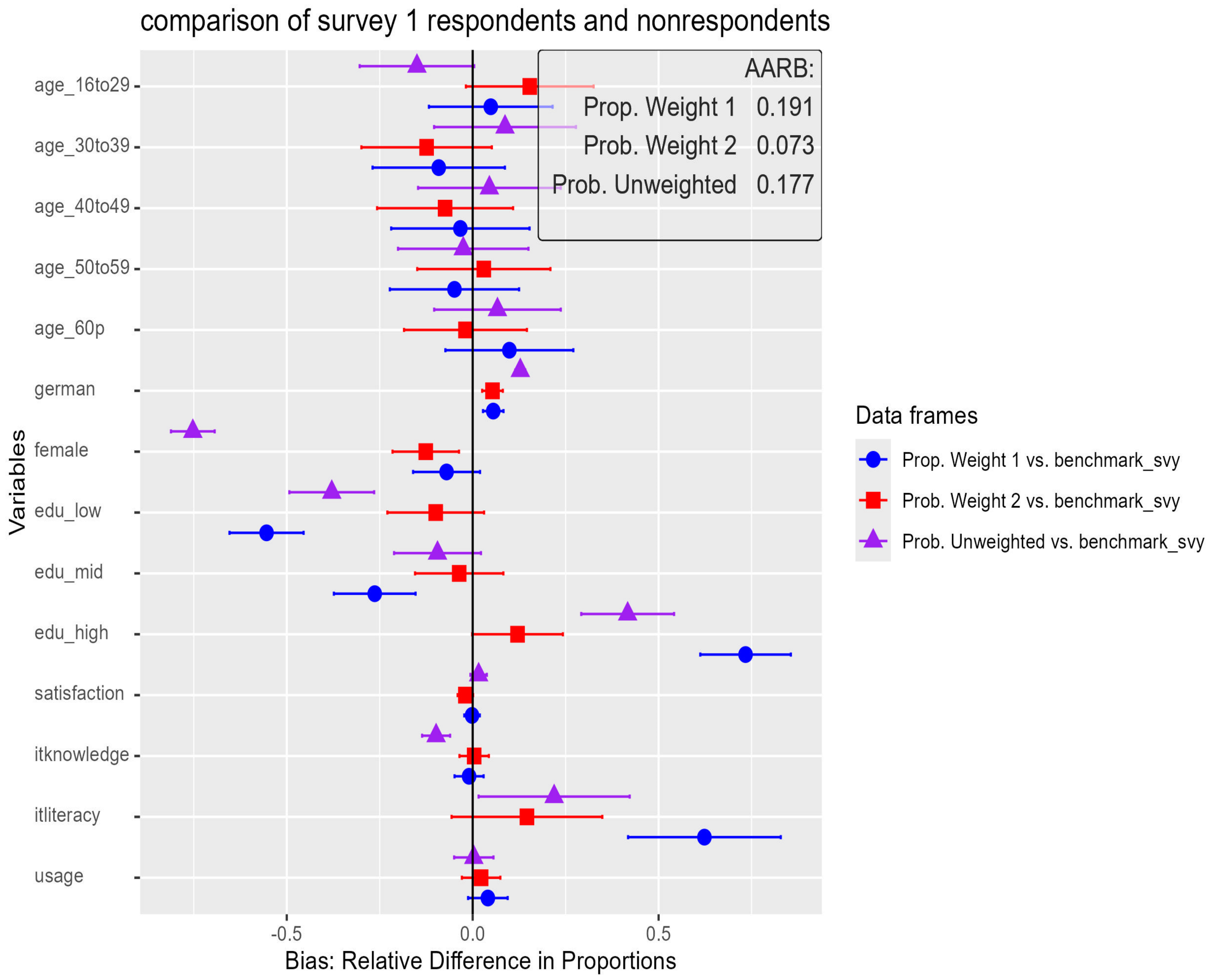
- Nonresponse weights can be obtained from the response models
- They are the inverse of the response propensities

Exercise 6 – Gaining Nonresponse Weights

```
diff_pop_weighted <-  uni_compare(nboots = 0,funct= "rel_prop",
                              plot_title = "comparison of survey 1 respondents and
nonrespondents",
                              dfs = c("prob_resp", "prob_resp",
                                      "prob_resp"),
                              summetric = "avg2",
                              benchmarks=c("benchmark_svy"),
                              weight = c("wgt1","wgt2",NA),
                              id = c("ID","ID",NA),
                              variables=c("age_16to29", "age_30to39" ,"age_40to49" ,
                                          "age_50to59", "age_60p", "german",
                                          "female","edu_low","edu_mid","edu_high",
                                          "satisfaction", "itknowledge",
                                          "itliteracy","usage"),
                              name_dfs = c("Prop. Weight 1",
                                           "Prob. Weight 2",
                                           "Prob. Unweighted"))
```

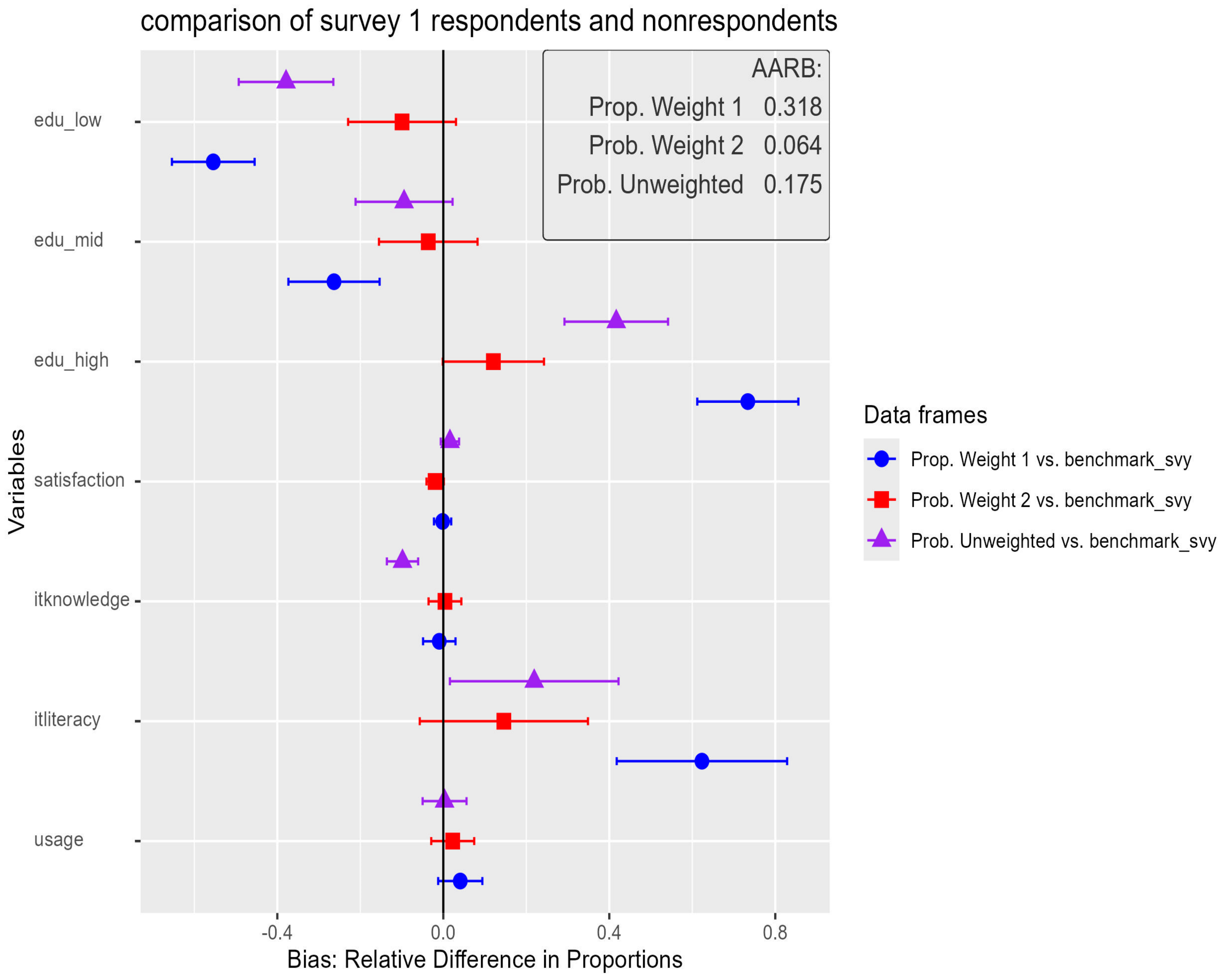
- Weights can be inserted here
- An id is also needed then

Exercise 6 – Gaining Nonresponse Weights



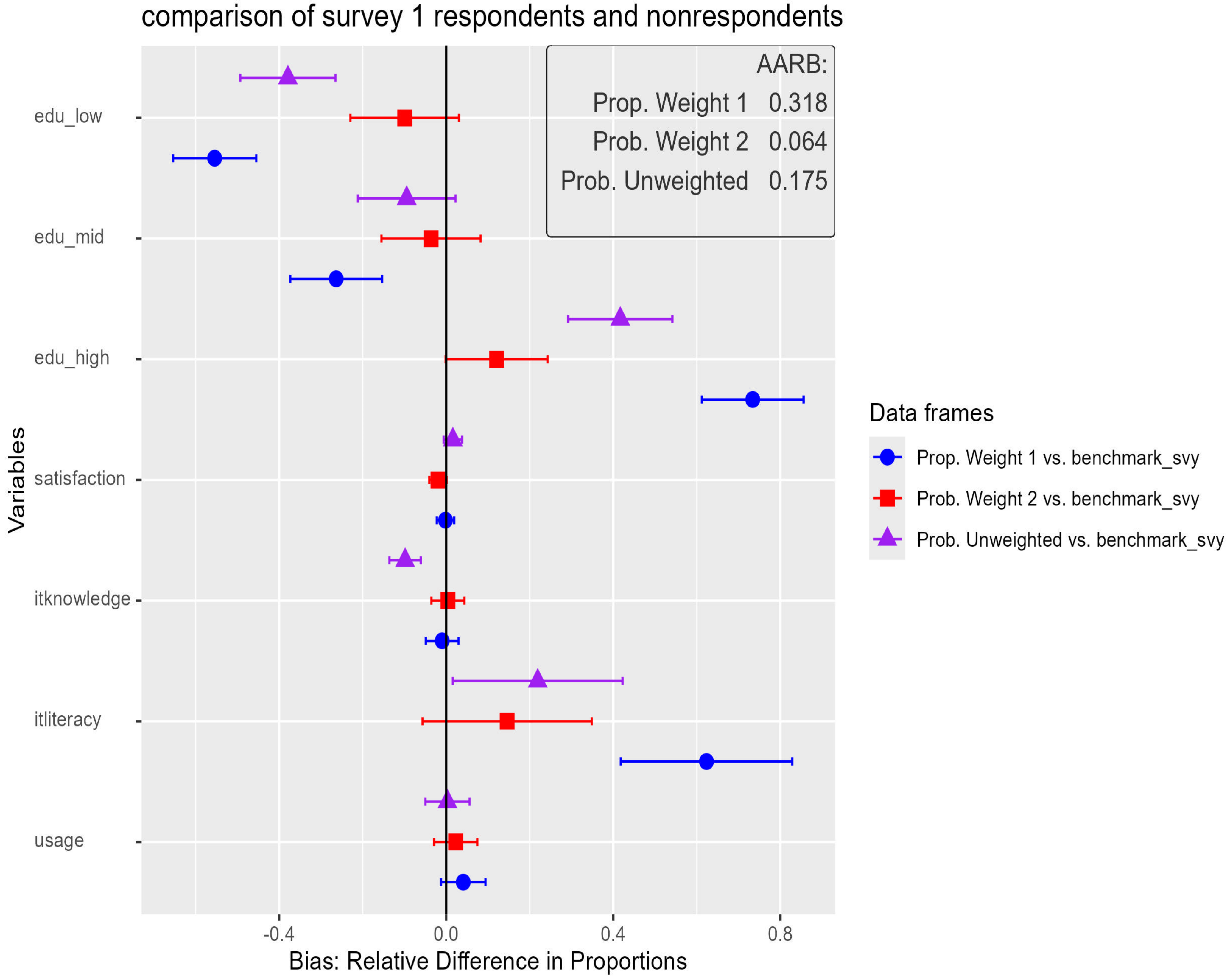
- Each weight reduced bias in frame-demographic variables as they were part of all nonresponse models
- However, for education and the substantial variables that were not part of the frame, bias even increased when using the first weight.
- In our data those variables were only weakly correlated to the weight 1.
- Some might even have a correlation opposite to their correlation to nonresponse.
- This can even lead to increased bias.

Exercise 6 – Generating Nonresponse Weights



- When studying bias after weighting, we would recommend generating a plot that only includes variables of interest that were not used for weighting.
- Here, it is even easier to see that the nonresponse weight based on the frame is not suitable for weighting.

Exercise 6 – Gaining Nonresponse Weights



To effectively reduce nonresponse bias, variables included in the weighting process should be

1. Correlated with survey response
2. Correlated with the variables of interest.

The higher those two correlations are the better a weight will work in reducing bias.

For real unsimulated probability surveys, demographics like age, gender are often used due to availability.

GESIS Leibniz Institute for the Social Sciences

Correlation of y and nonresponse weights

Leibniz Association

Correlation of y and nonresponse weights

The stochastic view on nonresponses bias shows that nonresponse bias is larger, the larger the correlation between the response propensity and Y.

$$Bias(\bar{Y}_R) \approx \frac{1}{\bar{\rho}} Cor(y, \rho)\sigma_y\sigma_\rho$$

In the absence of benchmark data on Y, the correlation of Y and nonresponse weights (= inverse response propensity) can give insight on nonresponse bias in Y.

Exercise 7 - Correlation of y and nonresponse weights

Determine the correlation of your nonresponse weights and your variables.

Are there differences between your individual models?

Exercise 7 - Correlation of y and nonresponse weights

```
full_prob$nr.weight.frame <- svy1.nr.weight.frame
full_prob$nr.weight.unobs <- svy1.nr.weight.unobs


cor.frame <- c(cor(as.numeric(as.character(full_prob$german)),
                   full_prob$nr.weight.frame),
               cor(as.numeric(as.character(full_prob$female)),
                   full_prob$nr.weight.frame),
               cor(full_prob$satisfaction, full_prob$nr.weight.frame),
               cor(full_prob$itknowledge, full_prob$nr.weight.frame),
               cor(full_prob$usage , full_prob$nr.weight.frame),
               cor(full_prob$itliteracy , full_prob$nr.weight.frame))


cor.unobs <- c(cor(as.numeric(as.character(full_prob$german)),
                   full_prob$nr.weight.frame),
               cor(as.numeric(as.character(full_prob$female)),
                   full_prob$nr.weight.frame),
               cor(full_prob$satisfaction, full_prob$nr.weight.unobs),
               cor(full_prob$itknowledge, full_prob$nr.weight.unobs),
               cor(full_prob$usage , full_prob$nr.weight.unobs),
               cor(full_prob$itliteracy , full_prob$nr.weight.unobs))
```

Calculate the correlation between the nonresponse weights and the variables of interest

Exercise 7 - Correlation of y and nonresponse weights

```
            german female satisfaction itknowledge  usage itliteracy
cor.frame  -0.6778 0.3617      -0.0320      0.0973 0.0109     0.0473
cor.unobs  -0.6778 0.3617      -0.0595      0.0823 0.0025    -0.0511
```

- Nonresponse weights are highly correlated with demographics
- Nonresponse weights are only weakly correlated to substantive variables.
- IT-literacy is positively correlated to the first weight (where the model only included frame information), while it is negatively correlated to response.
- This is the reason why the bias was strongly increased when using weight 1 in the previous exercise.

GESIS Leibniz Institute for the Social Sciences

Excursus – Raking weights in sampcompR

Leibniz Association

Excursus – Raking weights in sampcompR

We prepared an example of how you can include raking weights directly calculated by sampcompR in your survey comparison.

Take some time and try it out.

Interpret the results.

Excursus – Raking weights in sampcompR

1. Decide which variables you want to use for weighting (auxiliary variables)

```
target_vars<-c("age_16to29", "age_30to39" ,"age_40to49",
               "age_50to59", "age_60p", "german" ,"female")
```

We chose the same demographic variables as in the first nonresponse model

Excursus – Raking weights in sampcompR

2. Extract benchmark information from your auxiliary variables. They should be a list with marginal distributions for every auxiliary variable
   - Every element of the list should be named after one auxiliary variable
   - Auxiliary variables need to be labeled like in the survey dataset.

Excursus – Raking weights in sampcompR

2. Get raking margins of auxiliary variables:

```
targetlist<-function (target,df,weight=NULL,strata=NULL,id=NULL) {

  if(is.null(weight)==F) weight<-df[,weight]
  if(is.null(id)==F) id<-df[,id]
  if(is.null(id)==T) id<-c(1:nrow(df))
  if(is.null(strata)==F) strata<-df[,strata]
  target<-paste0("~",target)

  design <- survey::svydesign(ids = id, weights = weight, strata = strata, data=df)

  prop.table(survey::svytable(as.formula(target),design))
}
```

```
targets_svy<-map(target_vars,~targetlist(.,benchmark_svy,
                                          id = NULL,
                                          weight = NULL,
                                          strata = NULL))
names(targets_svy)<-target_vars
```

GESIS Leibniz Institute for the Social Sciences

Excursus – Raking weights in sampcompR

2. Get raking margins of auxiliary variables:

- The function from the previous slide can be used to get the targets.
- It also allows for weights and survey design elements from the auxiliary data frame.
- It only works if the auxiliary data is given in a data frame (e.g., based of another survey)

Excursus – Raking weights in sampcompR

3. If you want to compare more than one survey at the same time in sampcompR you have to provide a list with one element of target variables and targets for each survey.

```
target_list<-list(targets_svy, targets_svy)

target_vars_list<-list(target_vars,target_vars)
```

As we want to compare the probability and the nonprobability survey, we create these two lists.

Excursus – Raking weights in sampcompR

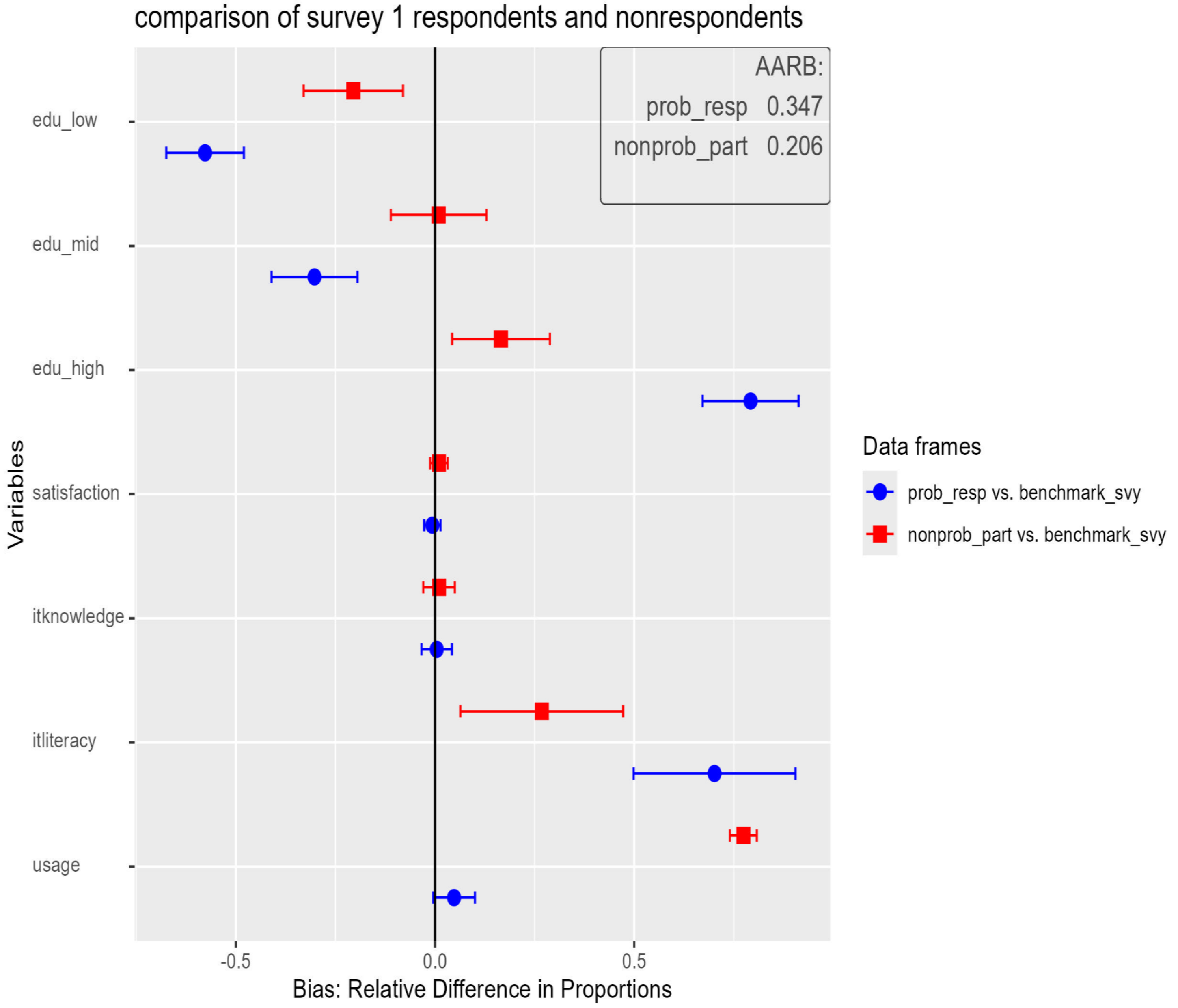
```
diff_pop_rake <-  uni_compare(nboots = 0,funct= "rel_prop",
                              plot_title = "comparison of survey 1 respondents and nonrespondents",
                              dfs = c("prob_resp", "nonprob_part"),
                              summetric = "avg2",
                              benchmarks=c("benchmark_svy"),
                              variables=c("edu_low","edu_mid","edu_high",
                                          "satisfaction", "itknowledge",
                                          "itliteracy","usage"),
                              adjustment_weighting = "raking",
                              adjustment_vars = target_vars_list,
                              raking_targets = target_list)

plot_diff_pop_rake<-plot_uni_compare(diff_pop_rake)
plot_diff_pop_rake
```

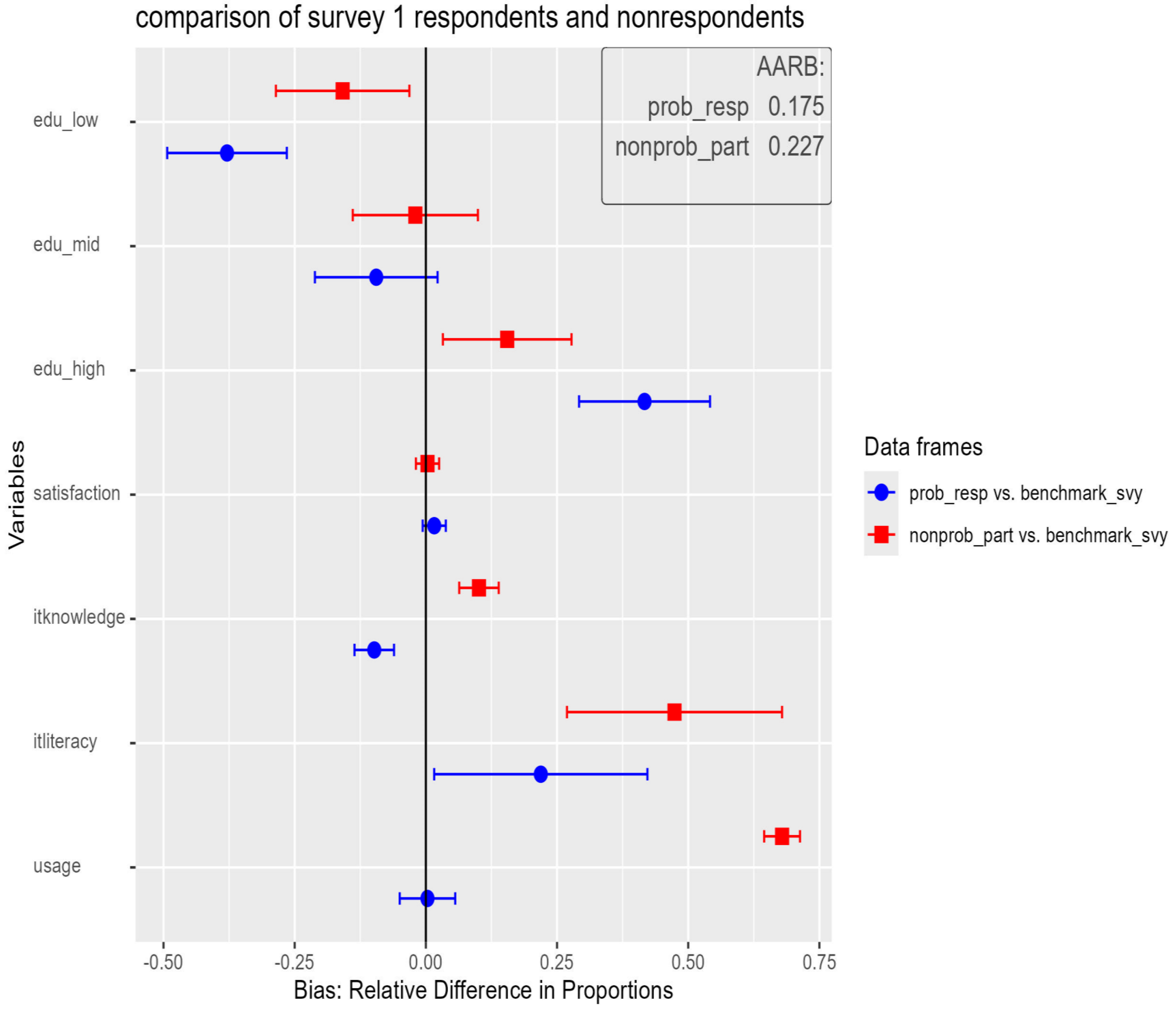
4. Put raking targets and the list of target variables in the uni_compare function and choose “raking” as a weighting method. (works also for biv_compare and multi_compare)

Excursus – Raking weights in sampcompR

Raked



Unweighted



As with the nonresponse weights, bias for the probability survey increased, as weighting did not include all needed variables

Bias in the nonprobability survey slightly decreased.

When (not) to use nonprobability surveys

When are nonprobability surveys fit for purpose

- Qualitative Studies
  - As no inference is intended, nonprobability surveys are suitable
- Hard to Reach Populations
  - Probability surveys can’t be conducted as no frame exists for certain populations
  - Probability surveys are extremely expensive for the population as one needs immense sample sizes to obtain a sufficient number of members of the rare population.
- Preliminary Studies
  - Pretests of a survey questionnaire
  - Pilot-studies to generate hypotheses, or investigate if a new topic should be further analysed with more reliable data

When might nonprobability surveys be fit for purpose

- Web tracking Analyses
  - Respondents who agree to web-tracking are a very selective group
- Experimental Designs
  - Experimental designs guarantee internal validity
  - Results are also externally valid if treatment effects are homogenous (i.e., treatment works the same for every treated person).
  - Up to date, there is no paper directly comparing experiments between probability and nonprobability surveys
  - The replication crisis indicates that this homogeneity assumption often does not hold.

When are nonprobability surveys no valid option

- Probability surveys are easily available to study the topic
  - The topic is well studied with probability surveys
  - Conducting a probability survey would be rather inexpensive
  - Data already exists that can be used for a secondary analysis
- The topic is of high interest to the public
  - It is likely that the press will report on the topic
  - One even plans to submit a press release
- Political or Social decisions should be derived from the study
  - Political or Social decisions that might impact population members must be as accurate as possible
- **Whenever the goal is to infer to the general population**

Data Archives

Check data archives of large-scale survey projects for secondary data use. For example:

- Gesis archive including ALLBUS, GLES, FReDA, Gesis Panel, German Internet panel and more. Data @ GESIS
- Cross-cultural data from the European Social Survey via the ESS Data Portal. ESS Data Portal | European Social Survey
- PIAAC and other research data on adult education. Data and studies on adult education
- SEOP research data center. DIW Berlin: Research Data Center SOEP
- NEPS data and more. LIfBi

What should be done when nonprobability surveys are used?

- Be as transparent as possible when describing survey recruitment
- Do not call the survey representative (even after weighting or using quota)
- Avoid univariate statistics if possible (multivariate models are found to be more robust)
- Think about potential biasing sources
- Conduct selection bias analysis for variables where benchmarks are available
- Be careful when interpreting and discussing results

GESIS Leibniz Institute for the Social Sciences

Final Remarks

Leibniz Association

Final Remarks

sampcompR

More information on nonprobability surveys can be found in the nonprobability Gesis survey guideline:

Rohr, Björn; Felderer, Barbara; Silber, Henning; Daikeler, Jessica; Roßmann, Joss; Schröder, Jette (2024) When are non-probability surveys fit for my purpose? Mannheim, GESIS – LeibnizInstitute for the Social Sciences (GESIS – Survey Guidelines). DOI: 10.15465/gesis-sg_en_050

More information on nonresponse bias can be found in the nonresponse bias Gesis survey guideline:

Barbara Felderer (2024). Nonresponse Bias Analysis. Mannheim, GESIS - Leibniz Institute for the Social Sciences (GESIS- Survey Guidelines). DOI: 10.15465/gesis-sg_en_047

Literature the course is based on

Felderer, Barbara, and Jessica M. E. Herzing. 2023. "What about the Less IT Literate? A Comparison of Different Postal Recruitment Strategies to an Online Panel of the General Population." *Field Methods* 35(3):219–35. doi: 10.1177/1525822X221132940.

Felderer, Barbara, Antje Kirchner, and Frauke Kreuter. 2019. "The Effect of Survey Mode on Data Quality: Disentangling Nonresponse and Measurement Error Bias." *Journal of Official Statistics* 35(1):93–115. doi: 10.2478/jos-2019-0005.

Rohr, Björn, Henning Silber, and Barbara Felderer. 2025. "Comparing the Accuracy of Univariate, Bivariate, and Multivariate Estimates across Probability and Non-Probability Surveys with Population Benchmarks." doi: 10.1177/00811750241280963

Rohr, Björn, Barbara Felderer, Henning Silber, Brady West, Steffen Pötzschke, and Jan Priebe. 2025. "Sampling for a Cross-National Survey in Six African Countries Using Social Media Advertisements: Comparison of Different Targeting Strategies, Types of Estimates, and Selectivity against Population Benchmarks." doi: 10.31235/osf.io/g9h47_v1

Questions?

Do you have any Questions?

Do you want to discuss your current plans on conducting a probability or nonprobability survey?

Thank you for your participation

Contact:

Barbara Felderer (Barbara.Felderer@gesis.org)
Björn Rohr (Bjoern.Rohr@gesis.org)

Leibniz
Association