RISQ

Representativity Indicators for Measuring Survey Quality

ITACOSM09

Siena, 10-12 June 2009

SEVENTH FRAMEWORK PROGRAMME

RISQ

REPRESENTATIVITY

INDICATORS

SURVEY

QUALITY

Financed by the 7th Framework Programme of the European Union

SEVENTH FRAMEWORK PROGRAMME

RISQ

National Statistical Institutes of

- Netherlands
- Norway
- Slovenia

and

Universities of

- Leuven
- Southampton

NO RESPONSE

- Source of non random errors
- Originates biased estimates
  - Response rate
  - Contrast between respondents and non respondents

INDICATORS

- Measure the degree to which a survey is representative of the population under investigation

- Supports the comparison of quality of different surveys and facilitates an efficient allocation of data collection resources

- **R-Indicators** – represents the closeness to representative response as a function of fully observed auxiliary information only

- **Partial R-Indicators** – measure the impact of the auxiliary variables on deviations from representative response

- Partial R-Indicators
  - **Unconditional** – measure the contribution of single variables to a lack of representative response
  - **Conditional** – measure the contribution of single variables to a lack of representative response *given* other variables
- Should supplement R-indicators

R-Indicators (Schouten et al 2008)

ρ_i - response propensity

- typically estimated through a logistic model

$$\rho_i = \rho_X(x_i) = E(R_i \mid x_i)$$

$x = (x_1, x_2, \ldots, x_m)'$ is known for all sample units

$$R_i = \begin{cases} 0 \quad if \quad i \quad is \quad non\ respondent \\ 1 \quad if \quad i \quad is \quad respondent \end{cases}$$

R-Indicators

$$R(\rho)=1-2S(\rho)$$

$$S(\rho)=\sqrt{\frac{1}{N-1}\sum_U(\rho_i-\bar{\rho}_U)^2}$$

$$\bar{\rho}_U=\frac{1}{N}\sum_U\rho_i$$

$$0\leq R(\rho)\leq 1$$

The population variance is estimated by a design-weighted sample variance

- ## Partial R-Indicators
  - **Unconditional (variable Z is used to model response propensities)**

$$P_1(Z,\rho_{X,Z})=\sqrt{S_b^2(\rho_{X,Z}\mid Z)}$$

where

$$S_b^2(\rho_{X,Z}\mid Z)=\frac{1}{N-1}\sum_k N_k(\bar{\rho}_{X,Z,k}-\bar{\rho}_{X,Z})^2\cong\sum_k\frac{N_k}{N}(\bar{\rho}_{X,Z,k}-\bar{\rho}_{X,Z})^2$$

Z **is a categorical auxiliary variable with**

$k=1,2,\ldots,K$

Population variances can be estimated by:

$$\hat{S}_b^2\left(\rho_{X,Z} \mid Z\right) = \sum_K \frac{\hat{N}_k}{N}\left(\hat{\bar{\rho}}_{X,Z,K} - \hat{\bar{\rho}}_{X,Z}\right)^2$$

$$\hat{S}_b^2\left(\rho_{X,Z} \mid Z = k\right) = \frac{\hat{N}_k}{N}\left(\hat{\bar{\rho}}_{X,Z,K} - \hat{\bar{\rho}}_{X,Z}\right)^2$$

With $\hat{N}_k = \sum_{i \in s_k} d_i$ being the estimated population size of stratum k .

If variable Z is not used to model response propensities, replace $\rho_{X,Z}$ with ρ_X .

Partial R-Indicators

- **Conditional (the auxiliary variable in study Z must be included in the model)**

$$P_2\left(Z,\rho_{X,Z}\right)=\sqrt{S_w^2\left(\rho_{X,Z}\mid X\right)}$$

where

$$S_w^2\left(\rho_{X,Z}\mid X\right)=\frac{1}{N-1}\sum_{l=1}^{L}\sum_{U_l}\left(\rho_{X,Z}\left(x_i,z_i\right)-\bar{\rho}_{X,Z,l}\right)^2$$

and

$$\hat{S}_w^2\left(\hat{\rho}_{X,Z}\mid X\right)=\frac{1}{N-1}\sum_{l=1}^{L}\sum_{s_l}d_i\left(\hat{\rho}_{X,Z}\left(x_i,z_i\right)-\hat{\bar{\rho}}_{X,Z,l}\right)^2$$

Simulation Study:

- dataset from 1995 Israel Census Sample of Individuals aged 15 and over (size=753.711)
- Probabilities of response were defined according to: child indicator, income group, age group, sex, number of persons in household and locality type

- Using the response indicator as dependent variable, a logistic regression model was fitted on the population with the above explanatory variables

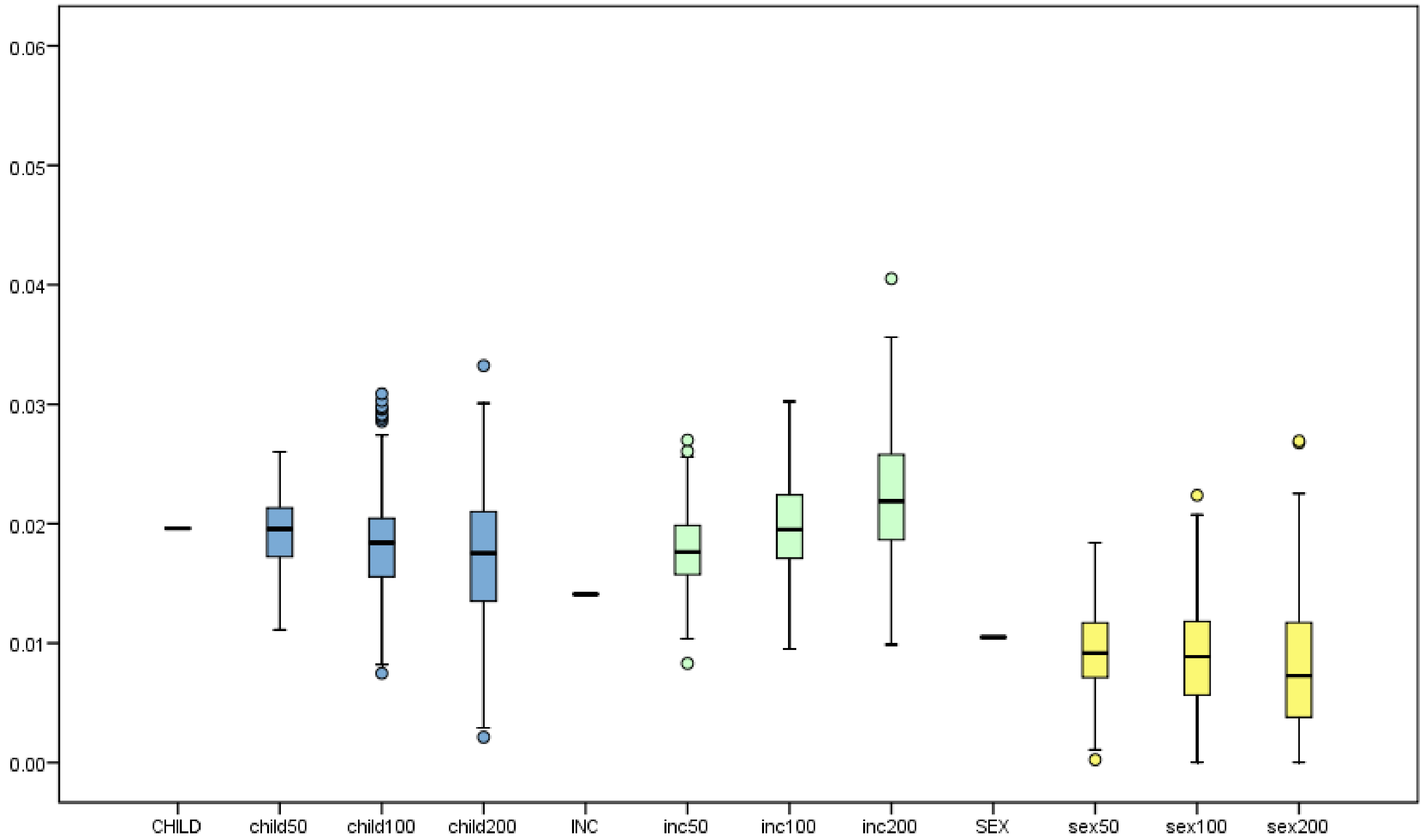
- The predictions from this model serve as the “true” response propensities for our simulations

- **400 samples were drawn**
- **Three sampling fractions:**
  - **1:50 (sample size of 15.074)**
  - **1:100 (sample size of 7.537) and**
  - **1:200 (sample size of 3.679)**

- **Boxplots show the “true” population value for each variable, the mean, the median and the spread of the distribution for each partial R-indicator.**

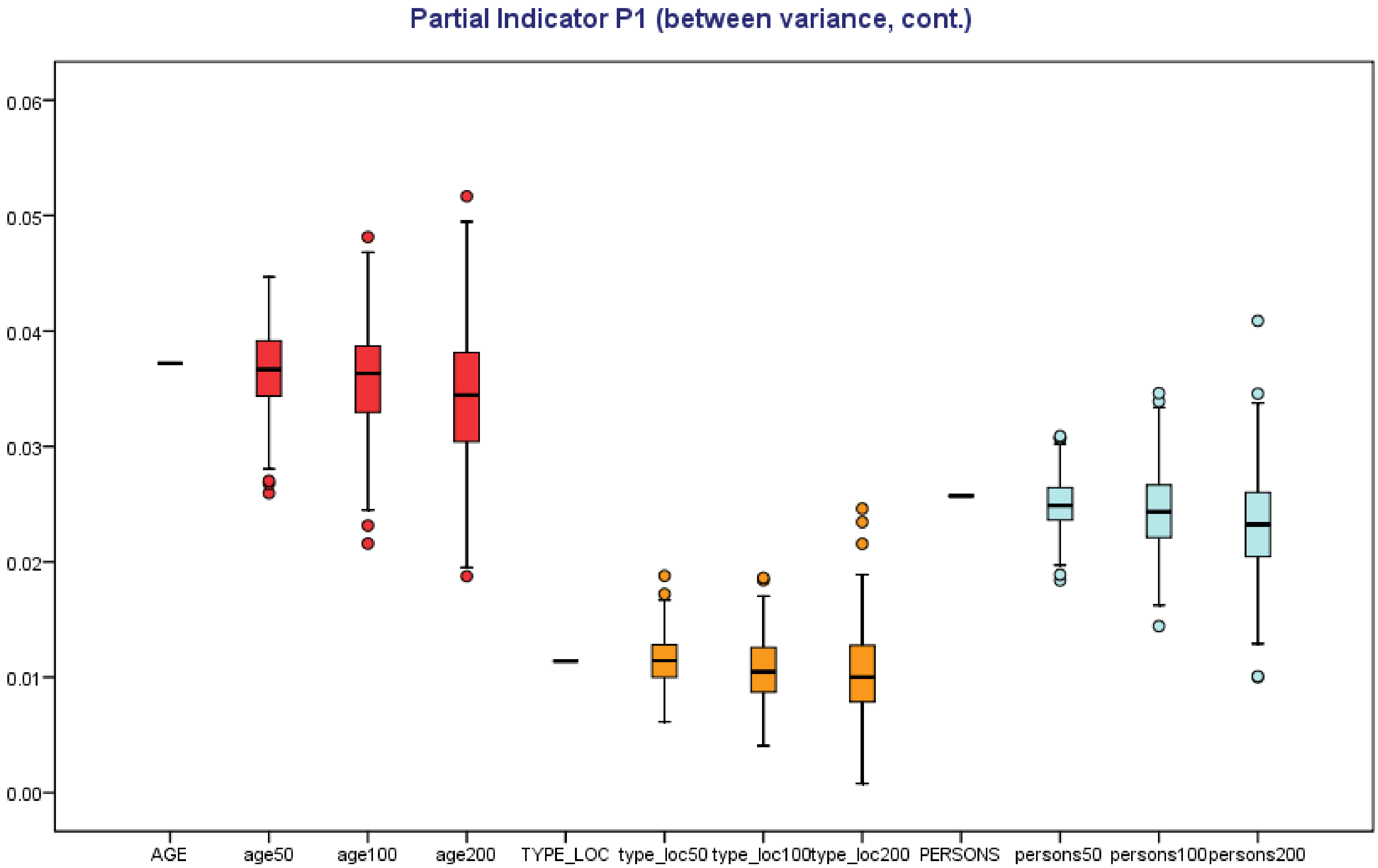
Representativity Indicators for Measuring Survey Quality

Partial Indicator P1 (between variance)



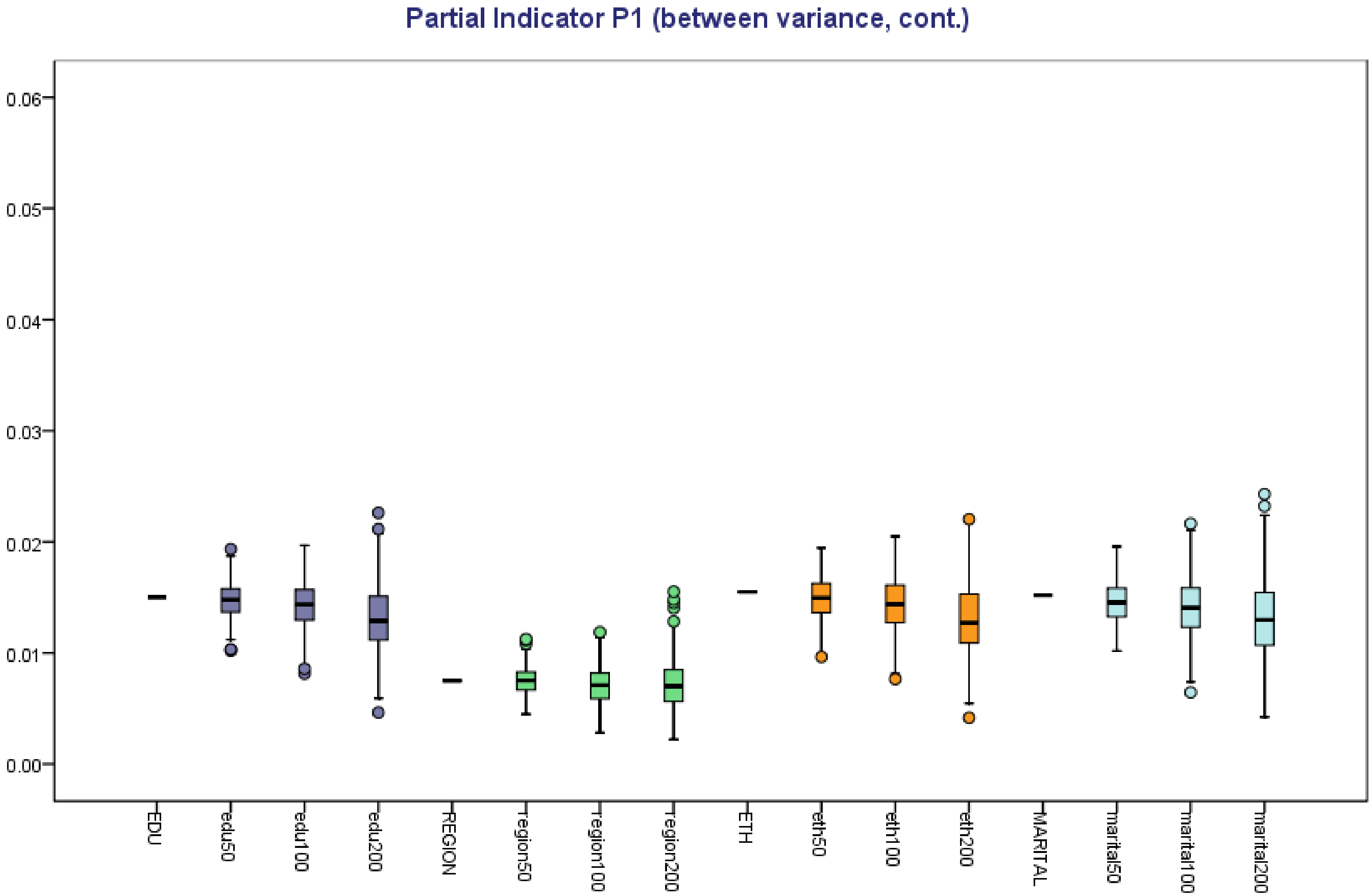
RISQ

Representativity Indicators for Measuring Survey Quality



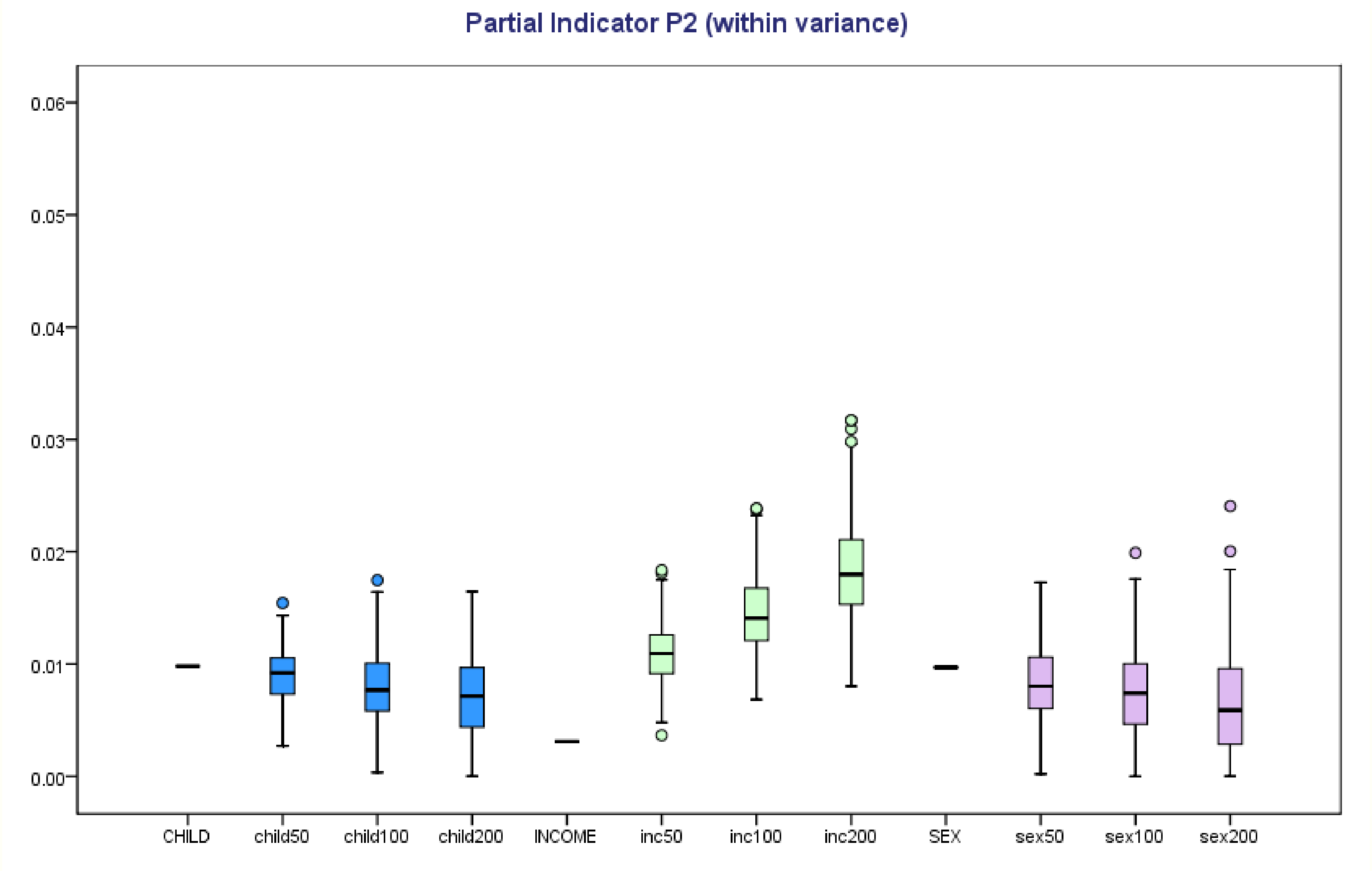
RISQ

Representativity Indicators for Measuring Survey Quality

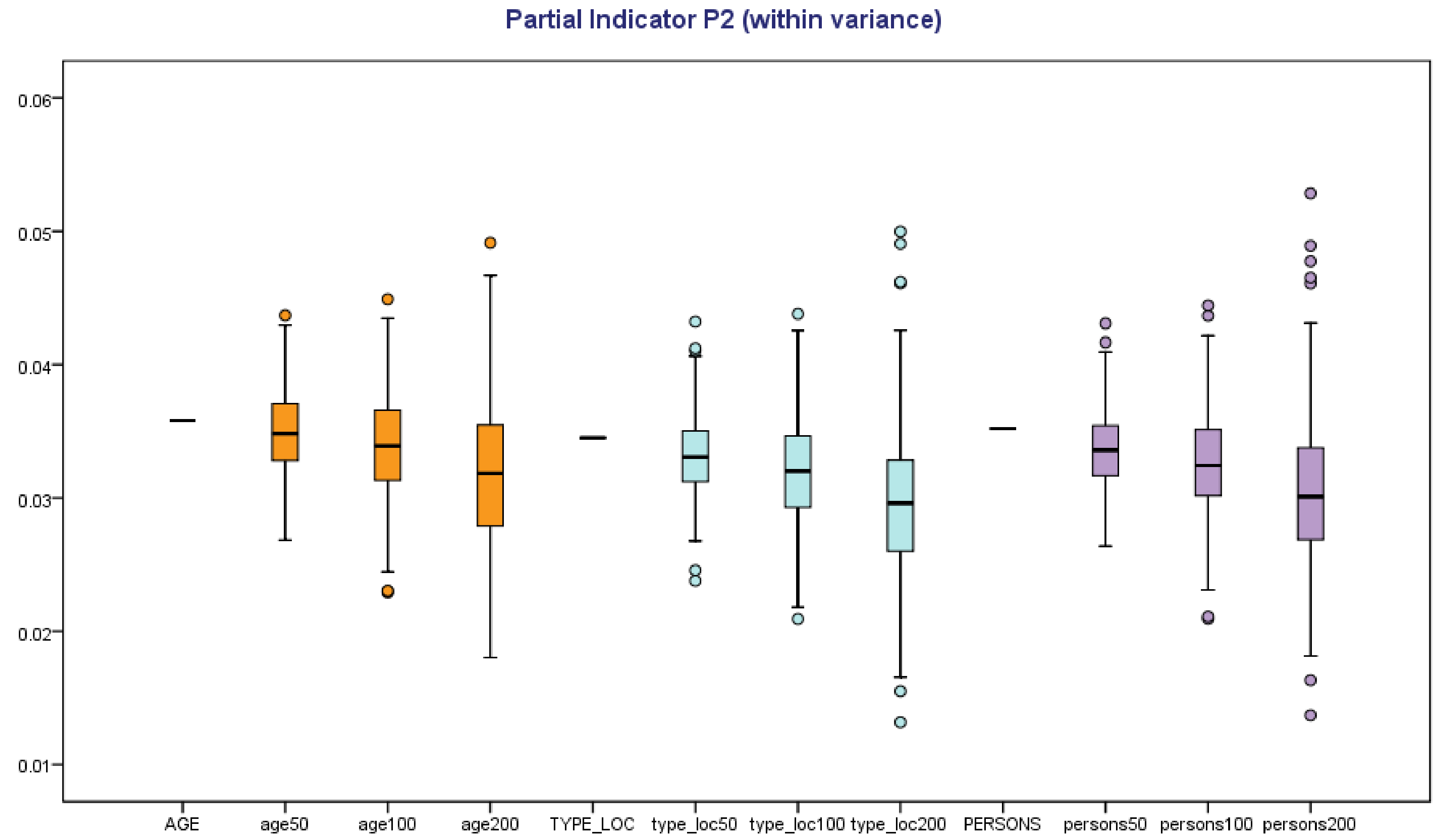


RISQ

Representativity Indicators for Measuring Survey Quality



Representativity Indicators for Measuring Survey Quality

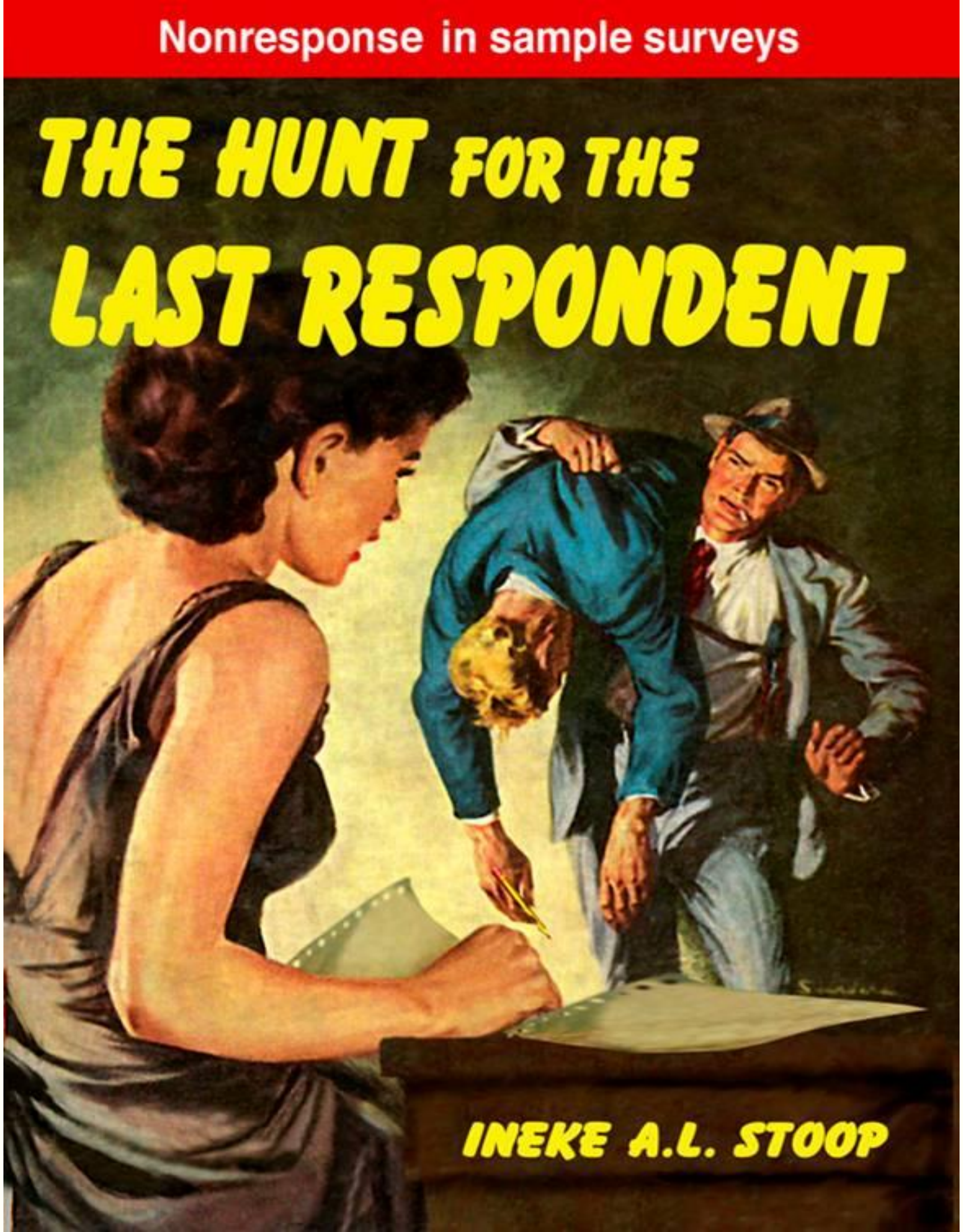


Conclusions

- This is a first exploration of partial indicators
- Partial indicators are useful to test survey methods, field monitoring and for weighting classes
- Identify variables that contribute to representativity
- Must be tested in real data sets in order to assess their impact on identifying variables and categories of variables that contribute to the lack of representativity

Conclusions

- Together with R-Indicators and response rates, survey managers can target data collection resources to specific sub-groups contributing to the lack of representativity, identify variables that might be used in survey estimation procedures to reduce non-response bias.



Nonresponse in sample surveys

THE HUNT FOR THE LAST RESPONDENT

INEKE A.L. STOOP