Response enhancement in household surveys

Jelke Bethlehem, Fannie Cobben, Barry Schouten
Statistics Netherlands

Ann Arbor, June 7 2007

Summary

- Statistics Netherlands
- Research programme
- R-indicators; theoretical background
- R-indicators in practice
- Discussion and future research

Statistics Netherlands

Statistics Netherlands is the National Statistical Institute of the Netherlands and its mission is the dissemination of reliable, coherent and relevant statistics to support policy and research.

2004: New data collection strategy
Statistics to be based on secondary data collection unless quality insufficient, and supplementary primary data collection by efficient mixed-mode strategies given constraints on quality.

Primary data collection

- Data collection modes:
    - Face-to-face (CAPI)
    - Telephone (CATI)
    - Web
    - Paper
- All surveys based on probability samples from municipality registers
- Registered land-line phone numbers are linked from commercial databases (70% coverage)
- Web data collection only in pilot studies using letters + logins to secured website (80% coverage)
- At present no household survey employs a mixed-mode design

Secondary data collection

Statistics Netherlands Act: By law allowed to use government registers and administrative data as input to the production of statistics

Examples:
Municipality registers
Tax Board registers on wages, VAT, profits, incomes
Registers for various goverment allowances
Register on value of real estate

Municipality registers function as backbone to both probability samples and other government registers

Quality and data collection strategy

Current practice: Enforce quality of statistics through choices in the implementation of the statsitical process, i.e. best practices and survey methodology

Data collection strategy:
- Dependence on register holders; quality becomes input-based
- Mixed-mode data collection demands total survey design

In other words: Quality needs to be measured explicitly and it does not suffice to rely on implementation in statistical process

Strategic Programme Nonresponse, Difficult Groups and Mixed-mode

Research projects:
1. Nonresponse reduction
2. Nonresponse adjustment
3. Difficult groups
4. Mixed-mode data collection

Response enhancement
Differentiated data collection protocols
Responsive/adaptive designs

Response enhancement

Goal: Differentiate data collection with respect to known auxiliary information (demographics, socio-economics) and possibly dependent on paradata in order to maximize response as a whole.

Background:
• Increased data collection efforts
• Continuum-of-resistence; ease-of-contact, ease-of participation

Need:
• Relation between contact and refusal conversion strategies and auxiliary info and paradata
• Optimization criteria: R-indicators

Indicators for representative response (R-indicators)

Indicators as tools to:
- compare surveys in time
- compare different data collection strategies
- monitor and control data collection

Consequence: Focus on response behavior, i.e. independent of survey items.

Important: Auxiliary information and paradata are crucial to any indicator. An indicator must always be published together with the available external information.

R-indicators – definition and concept

Definition (strong): A response subset is representative with respect to the sample if the response propensities are the same for all units in the population and if the response of a unit is independent of the response of all other units.

Definition (weak): A response subset is representative for a categorical variable X if the average response propensity over the categories of X is constant.

R-indicators – example

Variation of response propensities in population

$$R(\widetilde{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\rho_i - \overline{\rho})^2}$$

Estimated variation of response propensities

$$\hat{R}(\widetilde{\rho}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\rho_i - \overline{\rho}_{HT})^2}$$

Estimated variation of estimated response propensities

$$\hat{R}(\hat{\widetilde{\rho}}) = 1 - 2\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\frac{s_i}{\pi_i}(\hat{\rho}_i - \hat{\overline{\rho}}_{HT})^2}$$
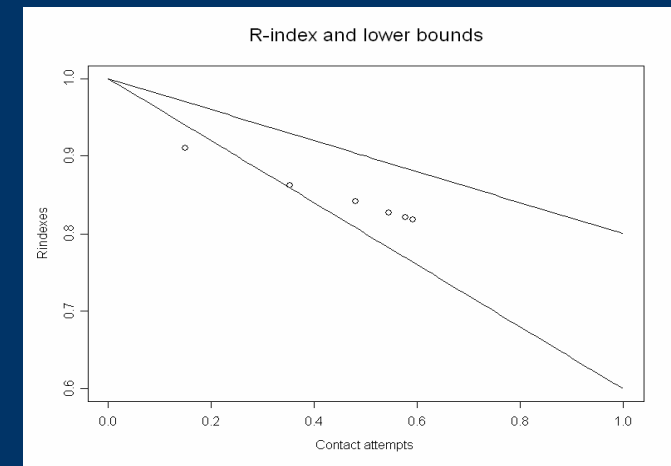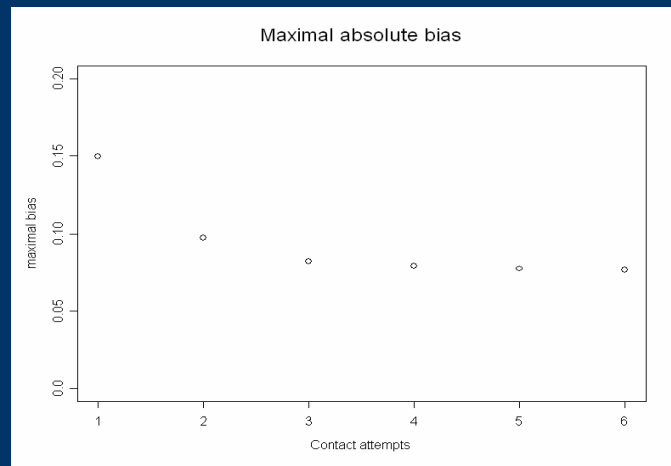
R-indicators – features

Interpretation: Dependence on X's and n

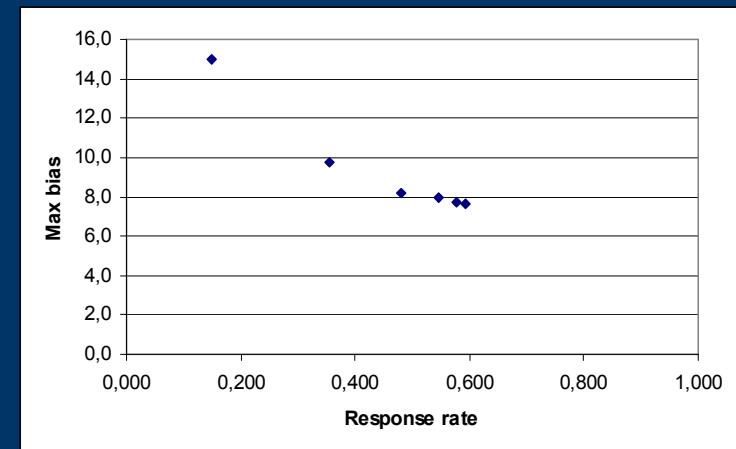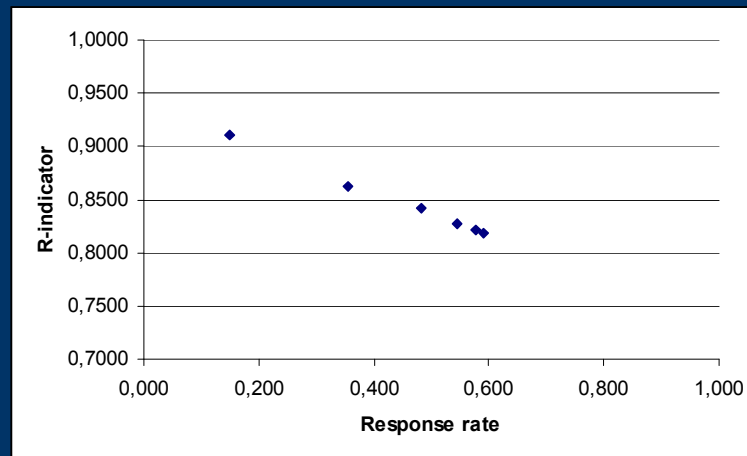Normalization of R-indicators: Relate to non-response bias under worst case scenario

$$| B(\hat{y}) | \leq \frac{S(\widetilde{\rho})S(y)}{\overline{\rho}} \leq \frac{S(\widetilde{\rho})}{2\overline{\rho}} = \frac{1 - R(\widetilde{\rho})}{4\overline{\rho}} = B_{\max}$$

$$R(\widetilde{\rho}) \geq 1 - 4\overline{\rho}B_{\max}$$

Maximal absolute bias

R-index and lower bounds

Example – contact attempts

Survey POLS 1998, sample size n=35.893
CAPI in first month, CATI in second month
X= Age, ethnic group, region
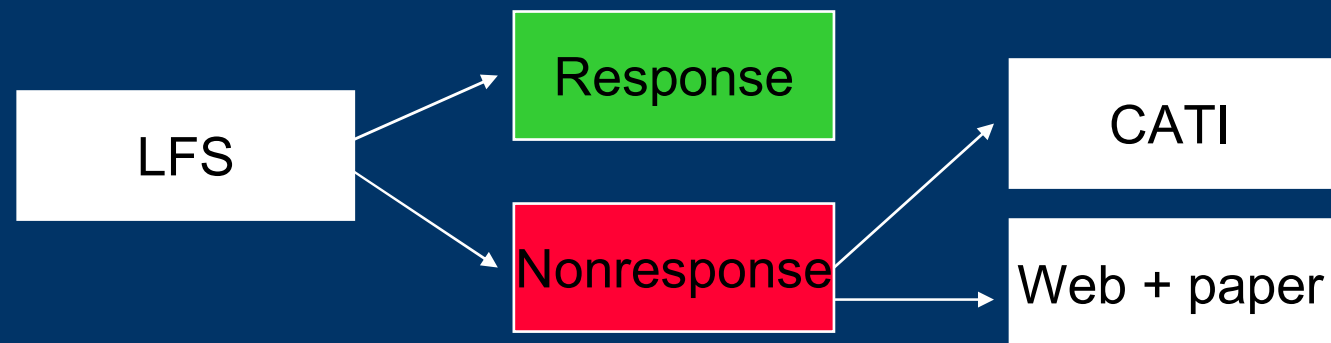
Example – call back & basic question

Survey LFS July – October 2005
Call back approach (Hurwitz 1949)
• Selection of best performing interviewers
• Additional training of interviewers
• Incentives
• Paper summaries of household characteristics

Basic question approach (Kersten & Bethlehem 1984)
Condensed questionnaires in CATI, paper, web

Example – call back & basic question

LFS n=18.076, CBA n=785

X=phone, region, ethnic group, household type, urbanity

	Response	R-indicator	Max bias
LFS	62%	79%	9%
LFS + CBA	77%	85%	5%

LFS n=18.076, BQA n=942

X=household type, urbanity, age, gender, job, allowance

	Response	R-indicator	Max bias
LFS	62%	79%	9%
LFS, phone	68%	85%	5%
LFS + CBA	76%	77%	8%
LFS + CBA, phone	83%	87%	4%

Example – mixing modes (1)

Security Monitor 2006

Reference survey

```
        ┌──────────┐
        │  Sample  │
        └──────────┘
         ╱        ╲
   ┌────────┐  ┌────────┐
   │  CATI  │  │  CAPI  │
   └────────┘  └────────┘
```

Pilot survey

```
        ┌──────────┐
        │  Sample  │
        └──────────┘
             │
        ┌──────────┐
        │   Web    │
        └──────────┘
         ╱        ╲
  ┌──────────┐  ┌──────────────┐
  │ Response │  │ Nonresponse  │
  └──────────┘  └──────────────┘
                 ╱            ╲
           ┌────────┐      ┌────────┐
           │  CATI  │      │  CAPI  │
           └────────┘      └────────┘
```
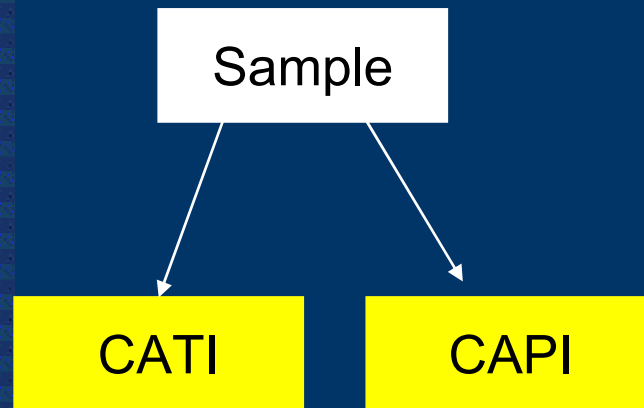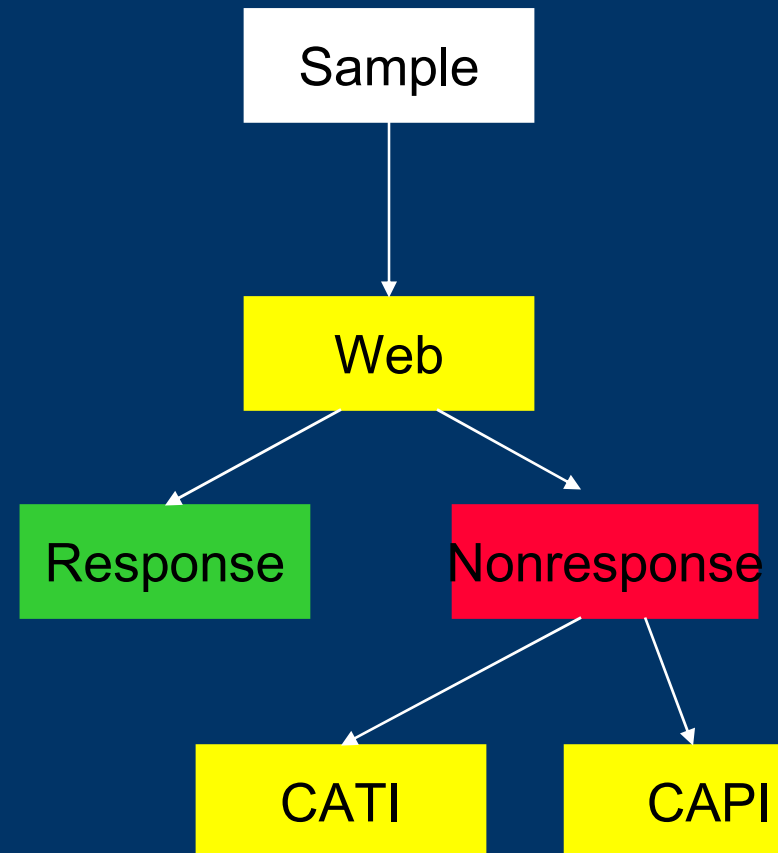
Example – mixing modes (1)

X=urbanity, household type, ethnic group, age

	n	Response	R-indicator	Max bias
Reference	30.139	69%	81%	7%
Pilot, web	3.615	23%	85%	13%
Pilot, total	3.615	63%	81%	7%

Example – mixing modes (2)

Informal Economy 2006

Pilot survey 1

Pilot survey 2

Sample

Sample

CAPI

Web
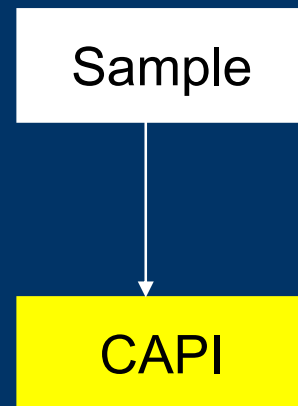
Response
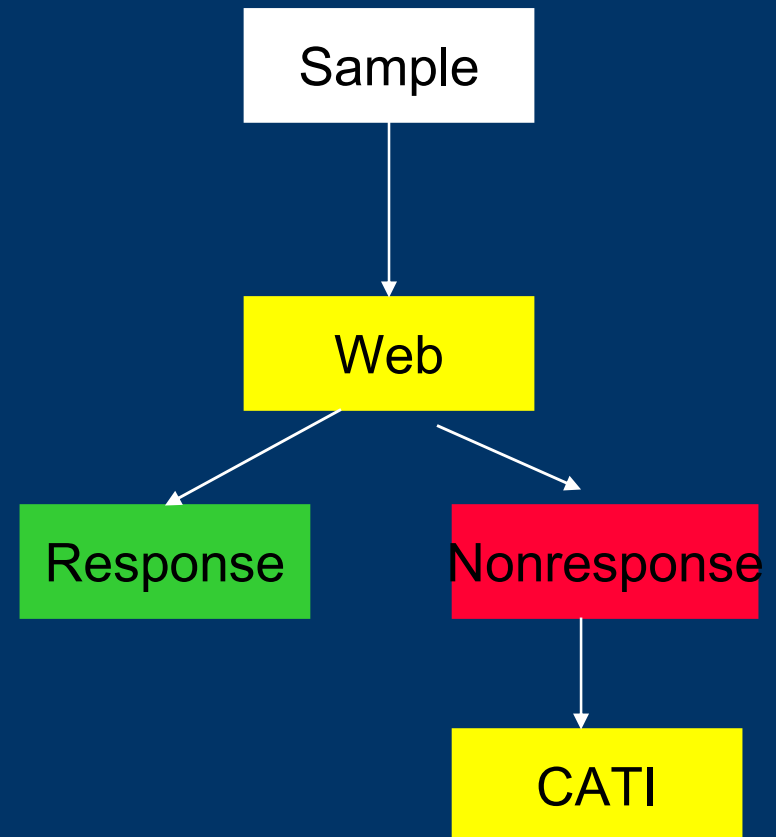
Nonresponse

CATI

Example – mixing modes (2)

X= urbanity, household type, ethnic group, age

	n	Response	R-indicator	Max bias
CAPI	2.000	57%	78%	10%
Web	2.001	34%	84%	12%
Web + CATI	2.001	49%	79%	11%

Example – incentives

Survey LFS 2005
Incentives:
1) no stamps, 2) 5 stamps , and 3) 10 stamps
*X= urbanity, average house value, ethnic group,
    size of household*

	n	Response	R-indicator	Max bias
No	11.774	66,6%	84,2%	5,9%
5	5.906	72,2%	81,6%	6,4%
10	5.982	73,8%	81,8%	6,2%

Example – differentiated contact strategies

Consumer Sentiments Survey 2003 and 2004

- Response decomposed in contact and participation
- Censored geometric model for contact propensities + unreachable class and part of day
- Logistic regression for participation propensities
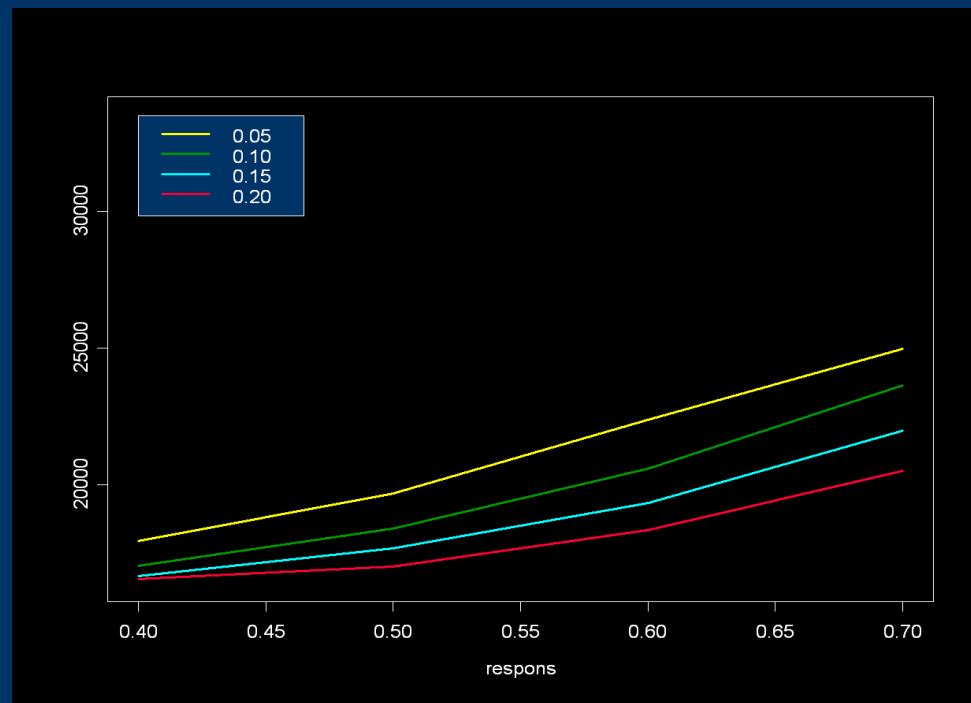
Consumer Sentiments Survey Nov-Dec 2006

- Current contact strategy
- Two differentiated contact strategies based on estimated propensities with 2003/2004 models; naive and advanced

Example – differentiated contact strategies

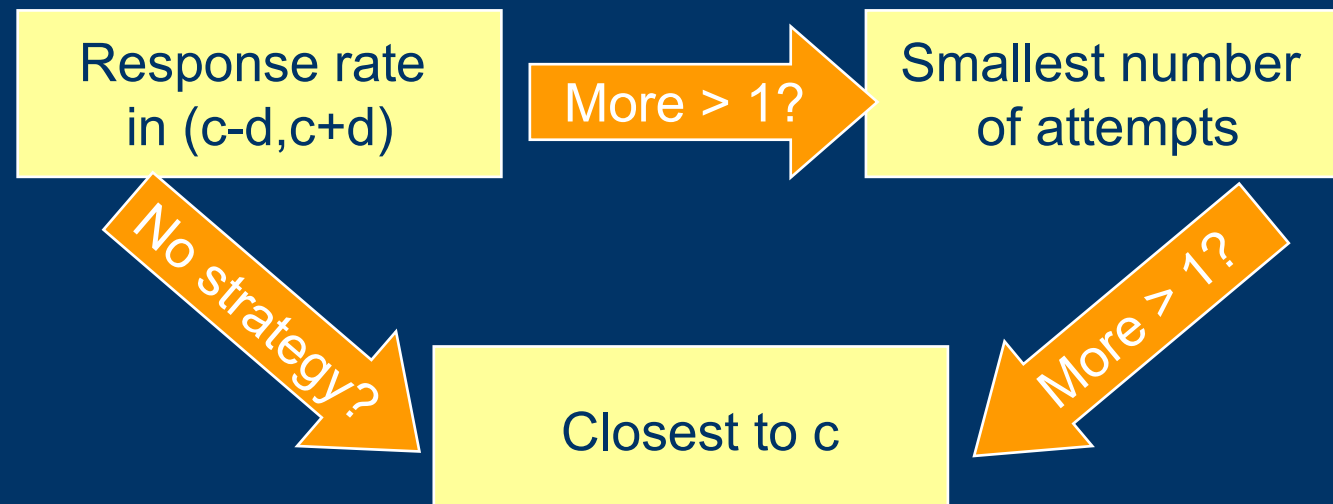Naive strategy: By hand, using experience, small number of call strata and estimated propensities
Advanced strategy: Add contact attempts until specified estimated response rate is attained.

Example – differentiated contact strategies

Naive strategy: By hand, using experience, small number of call strata and estimated propensities

Advanced strategy: Add contact attempts until specified estimated response rate is attained.

Response rate in (c-d,c+d) — More > 1? → Smallest number of attempts

Response rate in (c-d,c+d) — No strategy? → Closest to c

Smallest number of attempts — More > 1? → Closest to c

m = *maximal number of attempts per household*
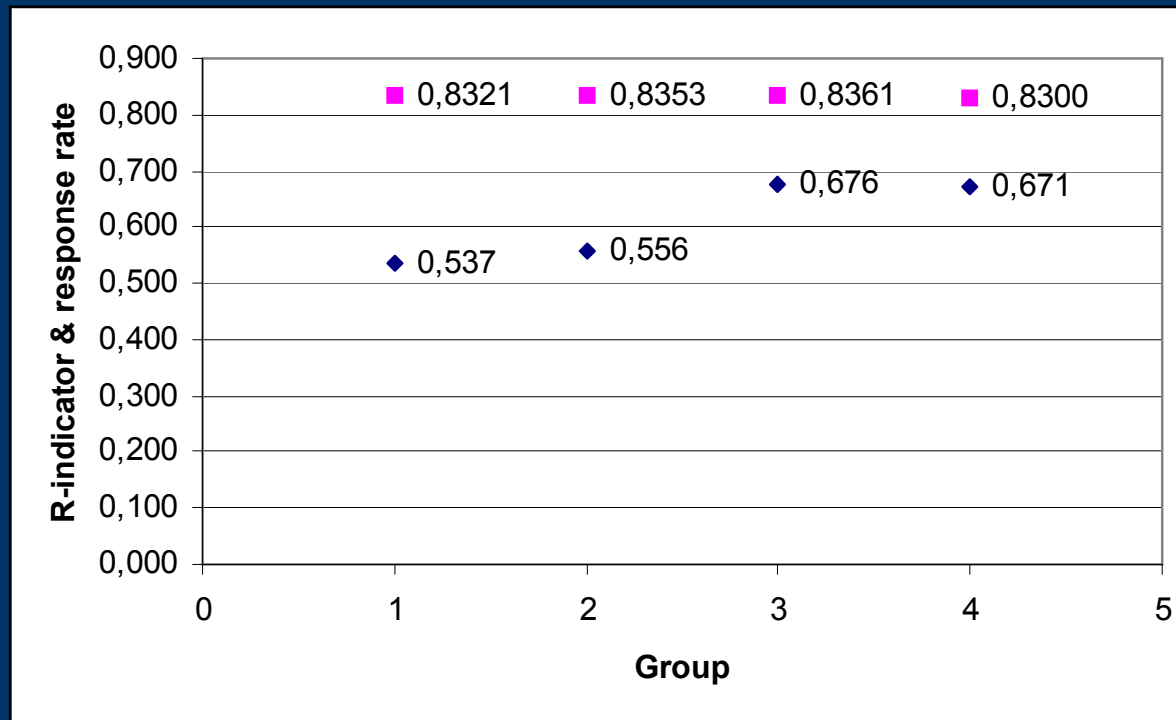
c = *target response rate*

d = *allowed margin to target response rate*

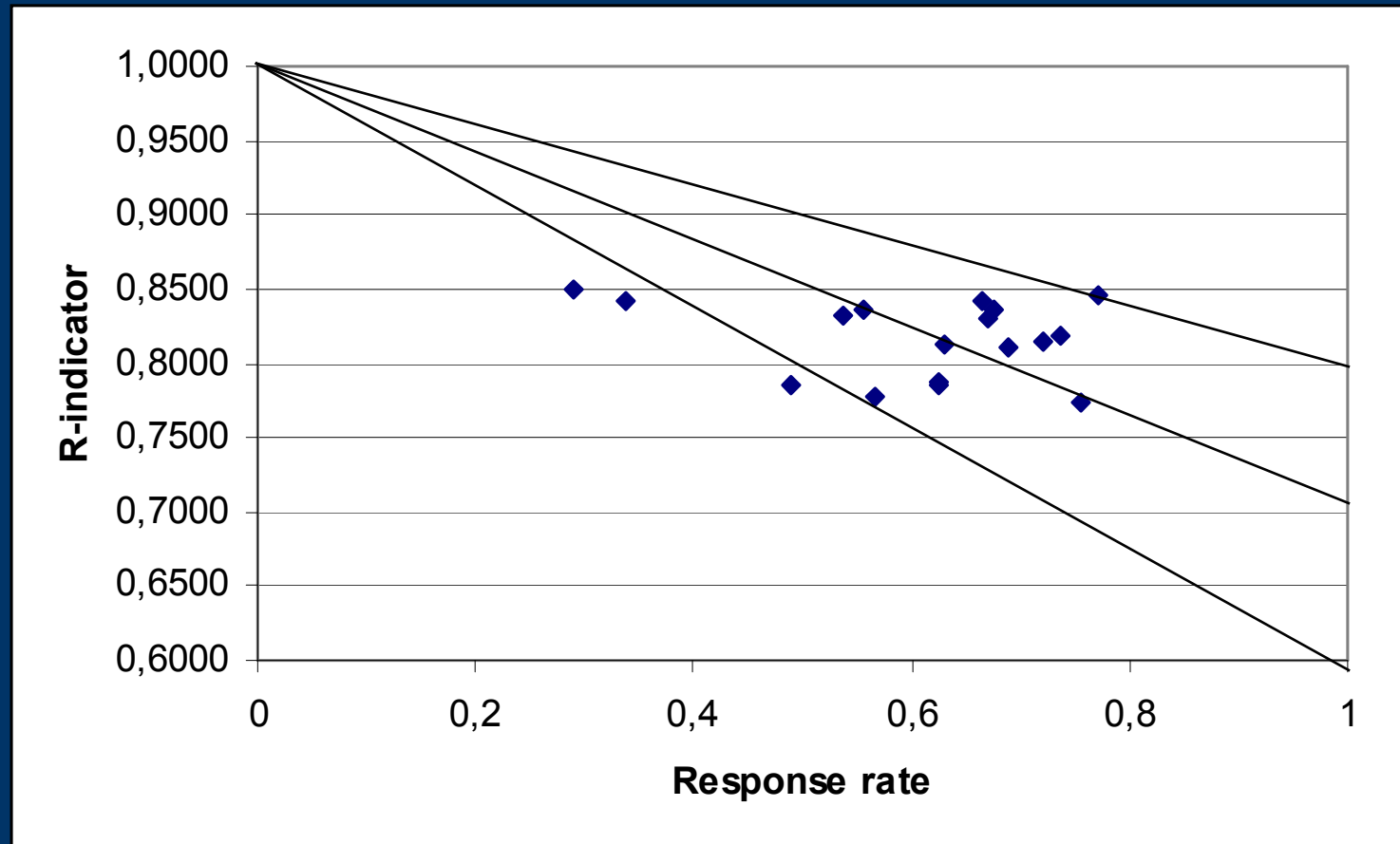# Example – differentiated contact strategies

n=1500 for all groups
X= urbanity, ethnic group, job, age, household type

Example – maximal bias

Discussion & future research

- Can we ignore survey items?
- Are there alternative R-indicators?
- Can R-indicators be tools in monitoring or even controlling survey data collection?
- Can R-indicators help in comparing different surveys (possibly over time)?
- How to interpret the values of R-indicators?

Discussion & future research

Short term:
- Extend theory to situation where only population totals are available
- Construction of R-indicator confidence intervals

Longer term:
- Joint research in 7th EU Framework Programme
- Responsive designs