# How to minimize survey errors in adaptive survey designs?

Melania Calinescu

Joint work with Barry Schouten (CBS) & Sandjai Bhulai (VU)

*NPSO workshop*
*The Hague, 2012*

# Ingredients

- Survey strategies: survey mode, contact strategy, refusal conversion, incentives, type of reporting
- Population subgroups - based on registry/frame data
- Cost and quality functions
- Input parameters – from historical data
- Optimization problem: how to allocate subgroups to strategies?

# Quality functions

ASDs focus mostly on nonresponse error:

- Response rate and subgroup response rates
- Variation in subgroup response rates (R-indicator), possibly adjusted for response rate (Schouten, Cobben, Bethlehem 2009)
- Fraction of Missing Information (Wagner 2010)

ASDs should **also** include measurement error:

- Minimize risk of undesirable measurement profiles / response styles

# Measurement error

- *Response representativity and measurement profile risks in the LFS*, Schouten, B., Leufkens, K. (2011)
- Measurement error
  - response error
  - measurement profile risk
- Response styles
  - social desirability
  - acquiescence
  - satisficing
  - straightlining
- Measurement profiles
  - Independent of survey variables
  - Respondent's state of mind

# Measurement error and Dutch LFS

- Population of interest
- LFS design features: type of reporting, number of calls
- Motivation for proxy-reporting
- *Approach 1:* quality objective function - the proportion of sample units that respond without the specified measurement profile.
- *Approach 2:* constraint - the proportion of respondents showing an undesirable measurement profile must be smaller than a specified threshold .

# Numerical results

- **Survey type**: uni-mode: face-to-face
- **Design features**: reporting type (self vs. proxy)
- **Objective**: maximize response rate $\max \bar{\rho} = \sum_{g} w_g \rho_g$

$$\rho_g = \sum_{t,m} f_g(t-1) x_g(t,m) p_g(t,m) r_g(t,m)$$

- **Constraints**:
  - reporting type – setting 1
  - number of visits – setting 2
  - measurement risk – setting 3
  - response representativity – setting 4
- **Decision variables**: binary decision $x_g(t,m)$ – approach or not

# Numerical results

| Reporting type \ Subgroups | | 15-25 | 26-55 | 56-65 |
|---|---|---|---|---|
| Self-report | Contact | 0.261 | 0.303 | 0.434 |
| | Cooperation | 0.594 | 0.651 | 0.69 |
| | Risk | 0.06 | 0.028 | 0.041 |
| Proxy-report | Contact | 0.392 | 0.367 | 0.461 |
| | Cooperation | 0.594 | 0.651 | 0.69 |
| | Risk | 0.078 | 0.035 | 0.047 |

Sample size: 10000

Subgroup proportions:
(0.196, 0.624, 0.18)

Max 10 visits pp.

# Setting 1a

Only self-report

| # visits | Group | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | Group response (%) | Response (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20000 | 15-25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | 43.4 |
| | 26-55 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 49.7 | |
| | 56-65 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 68.8 | |
| 25000 | 15-25 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15.5 | 52.6 |
| | 26-55 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 59.9 | |
| | 56-65 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 67.7 | |
| 30000 | 15-25 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 46.3 | 60.9 |
| | 26-55 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 63.3 | |
| | 56-65 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 68.3 | |

# Setting 1b

Only proxy-report

| # visits | Group | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | Group response (%) | Response (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20000 | 15-25 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23.3 | 50.5 |
| | 26-55 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 54.7 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 65.9 | |
| 25000 | 15-25 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 56.4 | 62.4 |
| | 26-55 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 62.4 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 68.9 | |
| 30000 | 15-25 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 59 | 64.2 |
| | 26-55 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 64.4 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 68.9 | |

# Setting 2a

Constraint on the number of available visits $\sum_g w_g c_g \leq C$

| # visits | Group | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | Group response (%) | Response (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20000 | 15-25 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23.3 | 50.5 |
| | 26-55 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 54.7 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 65.9 | |
| 25000 | 15-25 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 56.4 | 62.4 |
| | 26-55 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 62.4 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 68.9 | |
| 30000 | 15-25 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 59 | 64.2 |
| | 26-55 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 64.4 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 68.9 | |

# Setting 2b

Response representativity $$1-2\sqrt{\sum_g w_g(\rho_g-\bar{\rho})^2} \geq \alpha \qquad \alpha \in \{0.8, 0.85, 0.9\}$$

| # visits | Group | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | Group response (%) | Response (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20000 $\alpha = 0.8$ | 15-25 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 37.4 | 50 |
| | 26-55 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48.6 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 65.9 | |
| 20000 $\alpha = 0.9$ | 15-25 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 37.4 | 46.5 |
| | 26-55 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48.6 | |
| | 56-65 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 50 | |
| 20000 setting 2a | 15-25 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23.3 | 50.5 |
| | 26-55 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 54.7 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 65.9 | |

# Setting 3a

Approach 1: adjusted objective function $\max \sum_{g,m,t} w_g \rho_{g,m,t} (1-\theta_{g,m,t})$

| # visits | Group | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | Group response (%) | Response (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20000 | 15-25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | 48.5 |
| | 26-55 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 58.8 | |
| | 56-65 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 65.9 | |
| 20000 setting 2a | 15-25 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23.3 | 50.5 |
| | 26-55 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 54.7 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 65.9 | |

# Setting 3b

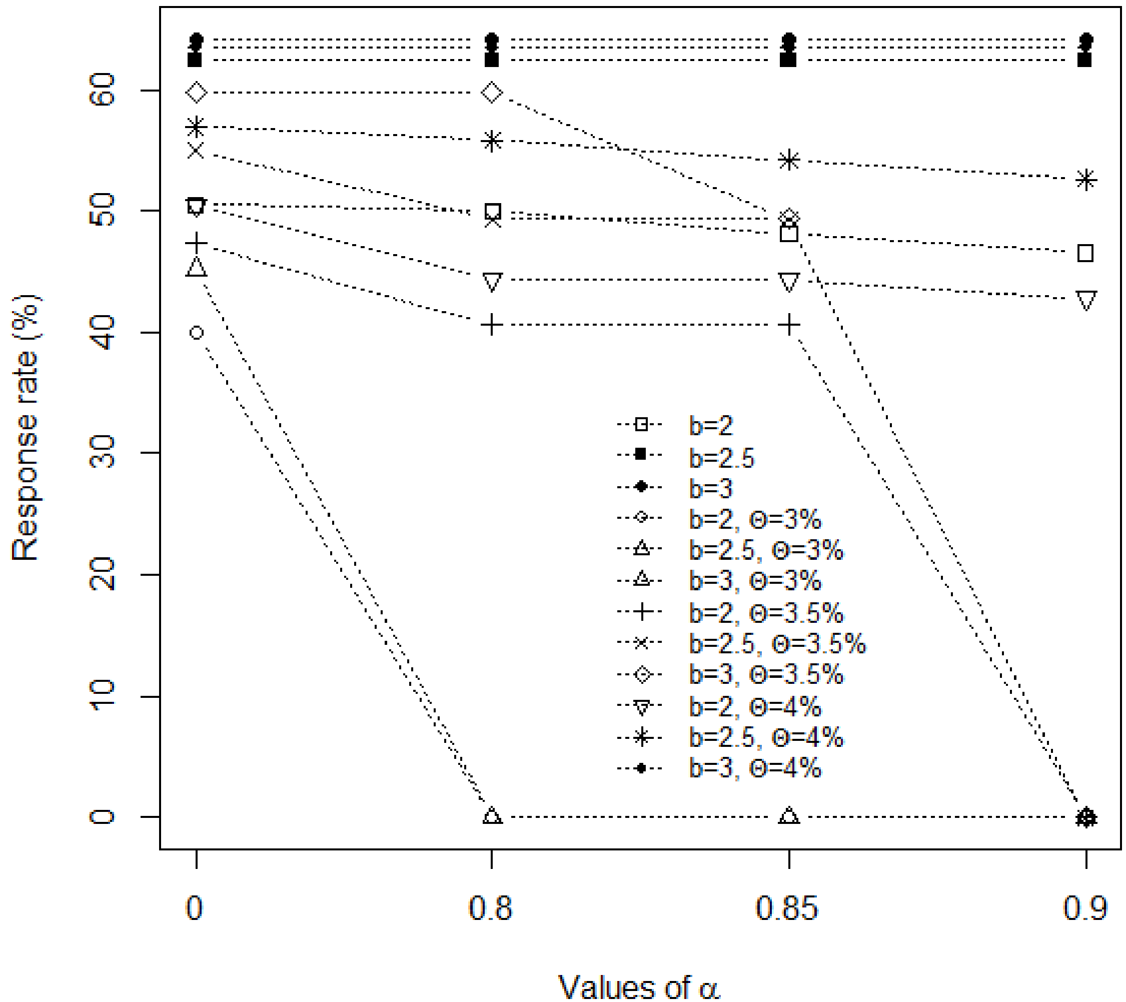
Approach 2 : constraint $\frac{\sum_{g,m,t} w_g \rho_{g,m,t} \theta_{g,m,t}}{\bar{\rho}} \le \Theta$ $\Theta \in \{3\%, 3.5\%, 4\%\}$

| # visits | Group | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | Group response (%) | Response (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20000 $\Theta = 3\%$ | 15-25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | 40 |
| | 26-55 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 64.1 | |
| | 56-65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | |
| 20000 $\Theta = 4\%$ | 15-25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | 50.4 |
| | 26-55 | 2 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 60.9 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 | 68.9 | |
| 20000 setting 2a, 4.17% | 15-25 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23.3 | 50.5 |
| | 26-55 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 54.7 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 65.9 | |

# Setting 4

| # visits | Group | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 | #9 | #10 | Group response (%) | Response (%) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20000 $\alpha = 0.8$ $\Theta = 3\%$ | 15-25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | 0 |
| | 26-55 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | |
| | 56-65 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.0 | |
| 20000 $\alpha = 0.8$ $\Theta = 4\%$ | 15-25 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 32.7 | 44.3 |
| | 26-55 | 1 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 46.9 | |
| | 56-65 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 | |
| 20000 $\alpha = 0.9$ $\Theta = 4\%$ | 15-25 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 37.4 | 42.8 |
| | 26-55 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 43.1 | |
| | 56-65 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 48 | |
| 20000 setting 2a | 15-25 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 23.3 | 50.5 |
| | 26-55 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 54.7 | |
| | 56-65 | 2 | 2 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 65.9 | |

# Results summary



# Future research

- Reliable input parameters – sufficient historical data
- Contact models – relax independence assumption
- Measurement error in other settings
- Move towards total survey error
- Include paradata and create dynamic designs
- Complex designs but still computationally feasible

Thank you!

*Adaptive survey designs that minimize nonresponse and measurement risk,* M. Calinescu, B. Schouten, S. Bhulai

Available soon at www.cbs.nl

Contact: *melania.calinescu@vu.nl*