Assessing risk of nonresponse bias and dataset representativeness during survey data collection

Gabriele Durrant

Joint work with Jamie Moore, Solange Correa and Peter W.F. Smith
University of Southampton

RSS Social Statistics Section, 3 March 2016

The Research Project

- Was funded by
  - ESRC research grant on “Paradata”
  - ESRC National Centre for Research Methods, Workpackage 1 “Data Collection for Data Quality”
  - ESRC Administrative Research Centre for England (ADRCE).

Introduction

- Focus has shifted from nonresponse rate to nonresponse bias

- Key question: How to monitor, assess and minimise (risk of) nonresponse bias?
  - Post or during data collection

- Questions from survey practice: when to stop calling?

Introduction

- Fully observed information on both respondents and nonrespondents necessary
- Sample frame information from
  - register / Census
  - administrative data
  - previous wave

- Datasets (face-to-face surveys):
  - ONS Census nonresponse link study
  - Understanding Society

How to assess the risk of nonresponse bias?

- Main idea: measure similarity between sample data obtained and frame data in terms of variation in response rates

- Use of a response propensity model to obtain estimated response propensities

- Representativeness indicators: estimate variation in these response propensities (SD = Standard deviation of the response propensities)

- Low variability in response propensities imply high representativeness

Representativeness Indicators

- R indicator: $$R = 1 - 2SD$$

  SD= standard deviation of response propensities

  Ranges between 0 and 1

  Close to 1 indicates high representativeness

- CV (Coefficient of Variation): $$CV = \frac{SD}{r}$$

  r = response rate

  CV close to 0 indicates high representativeness

- Here computed at each call (visit to a household by interviewer)

Applying these Methods – Key Research Objectives

1. Visualise trends in dataset representativeness
2. Are trends in representativeness generalizable across surveys (of the same population)?
3. Can we derive stopping points for an adaptive data collection strategy – can these be generalised?

Data

Data

- ONS 2011 Census Non-Response Link Study (CNRLS)
- Links response indicator from three UK social surveys to survey call record data and census household (HH) information on sample frames
- 3 (cross-sectional) face-to-face surveys:

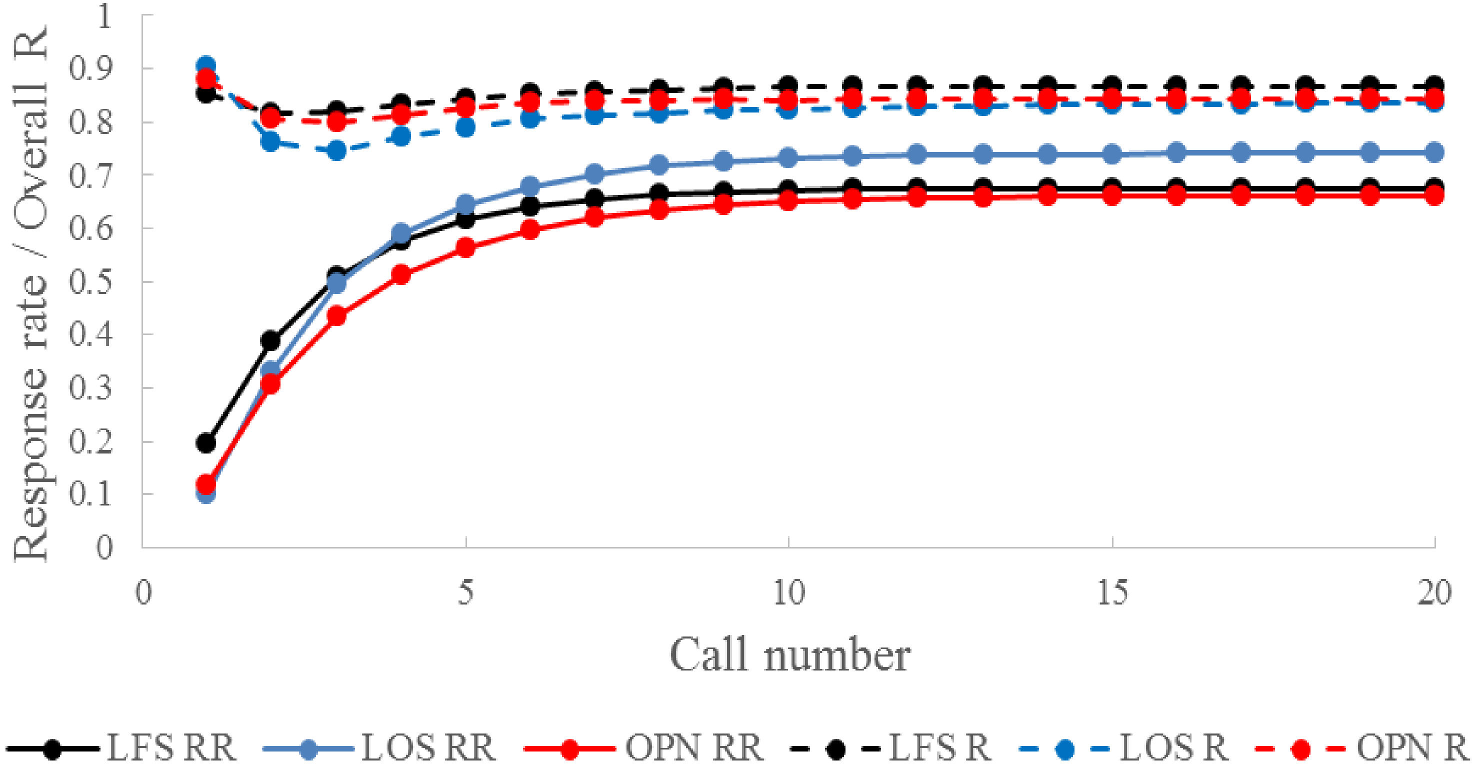
  - Labour Force Survey (LFS) (wave 1)

  - Life Opportunities Survey (LOS) (wave 1)

  - Opinions Survey (OPN)
- Up to 20 calls to a household

Application and Results

R indicators

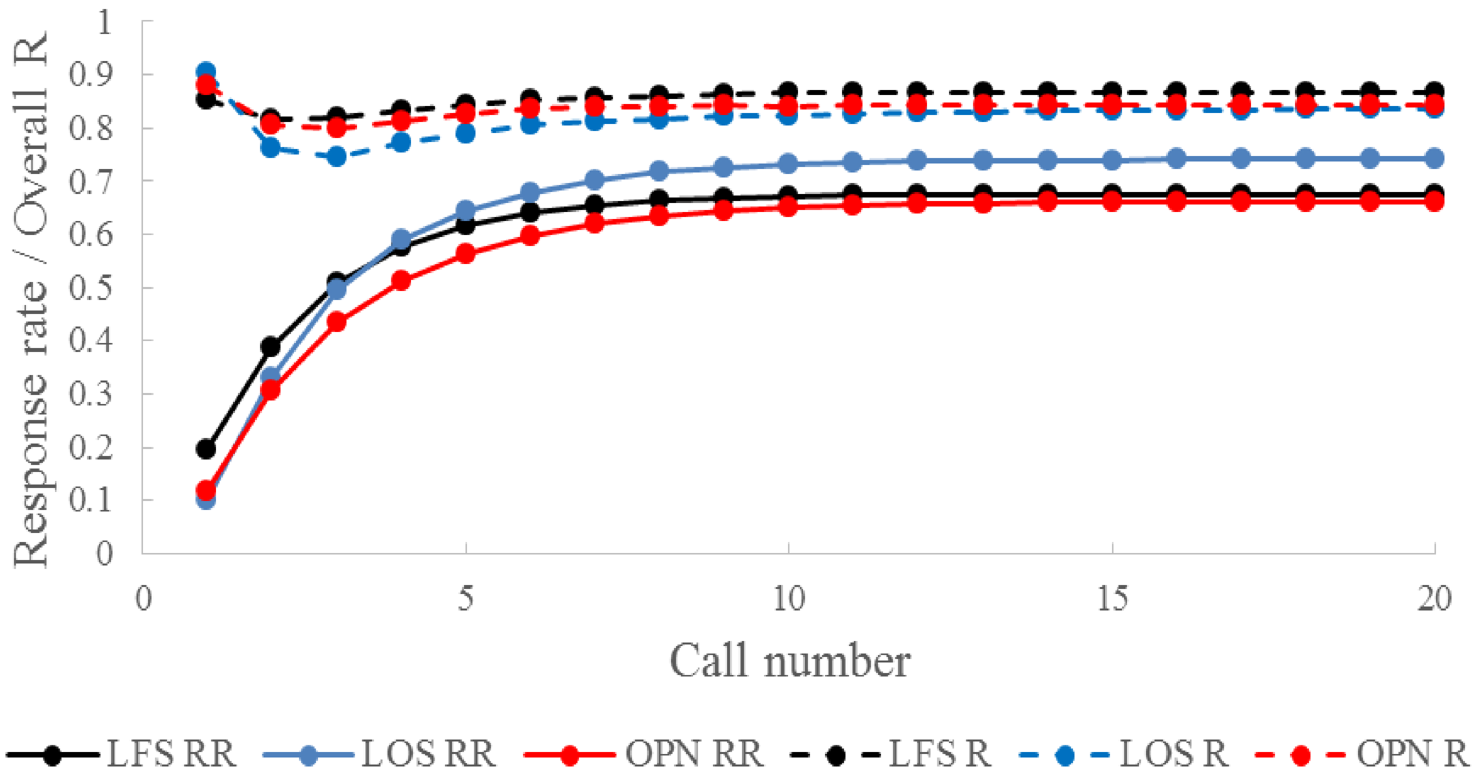


final response rate: LFS = 65.7%

LOS = 70.1%

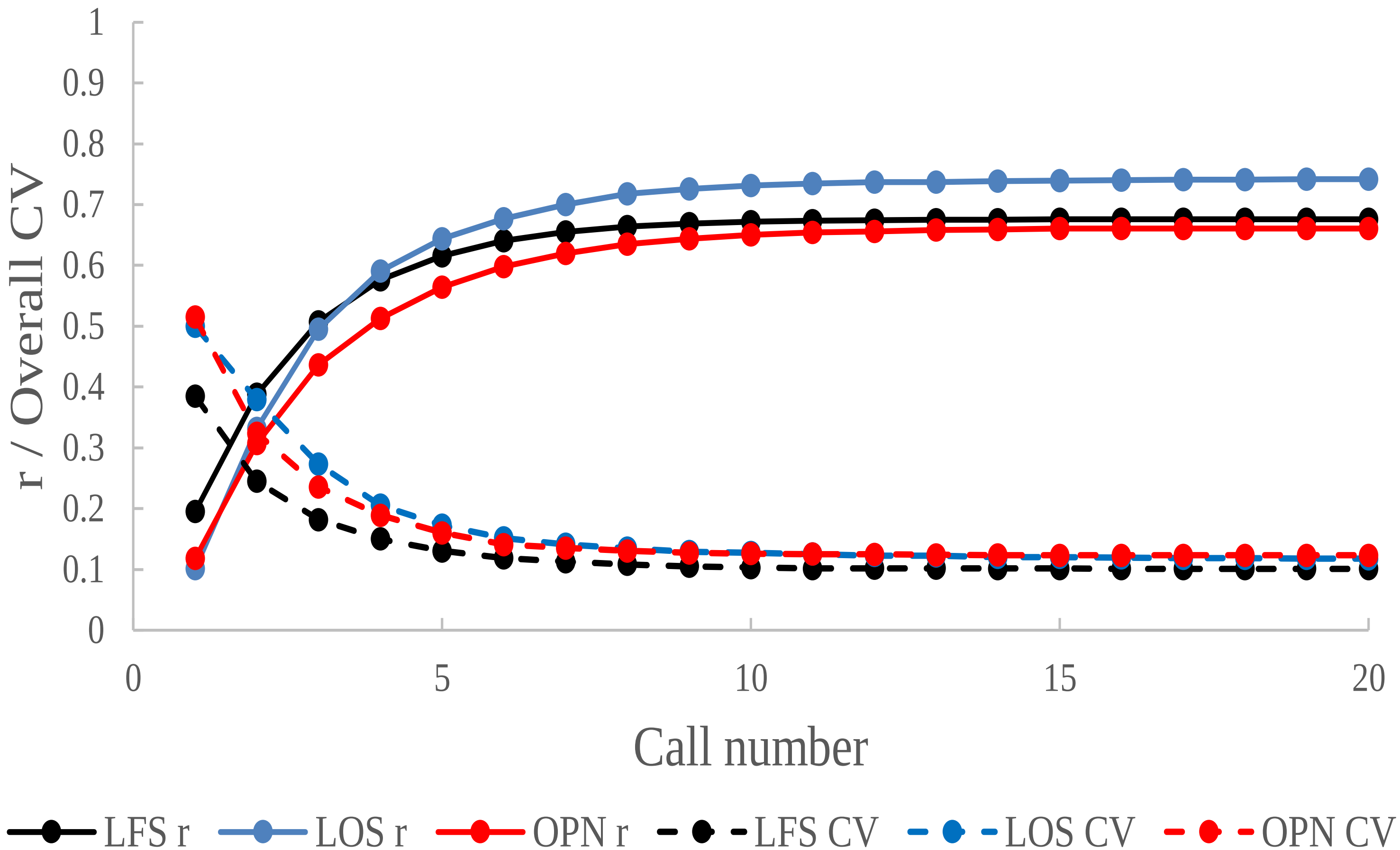
OPN= 64%.

R indicators



- In case of low response rates (as is the case early on in data collection) small response propensity variation, limited potential for response propensity divergence
- R indicators close to 1, falsely indicating high representativeness
- R-indicator can be misleading in this case

CV (Coefficient of Variation)



- CV standardises SD by r; overcomes the problem of the R indicator
- CV decreasing, close to 0 indicating high representativeness

(Unconditional) Partial Indicators

- Aim: estimate the extent to which response is representative with respect to a covariate or a particular category
- We found similarities across surveys, some variables improve across calls, some remain the same (but do not improve)

Phase Capacity or Stopping Points

Stopping or Phase Capacity Points

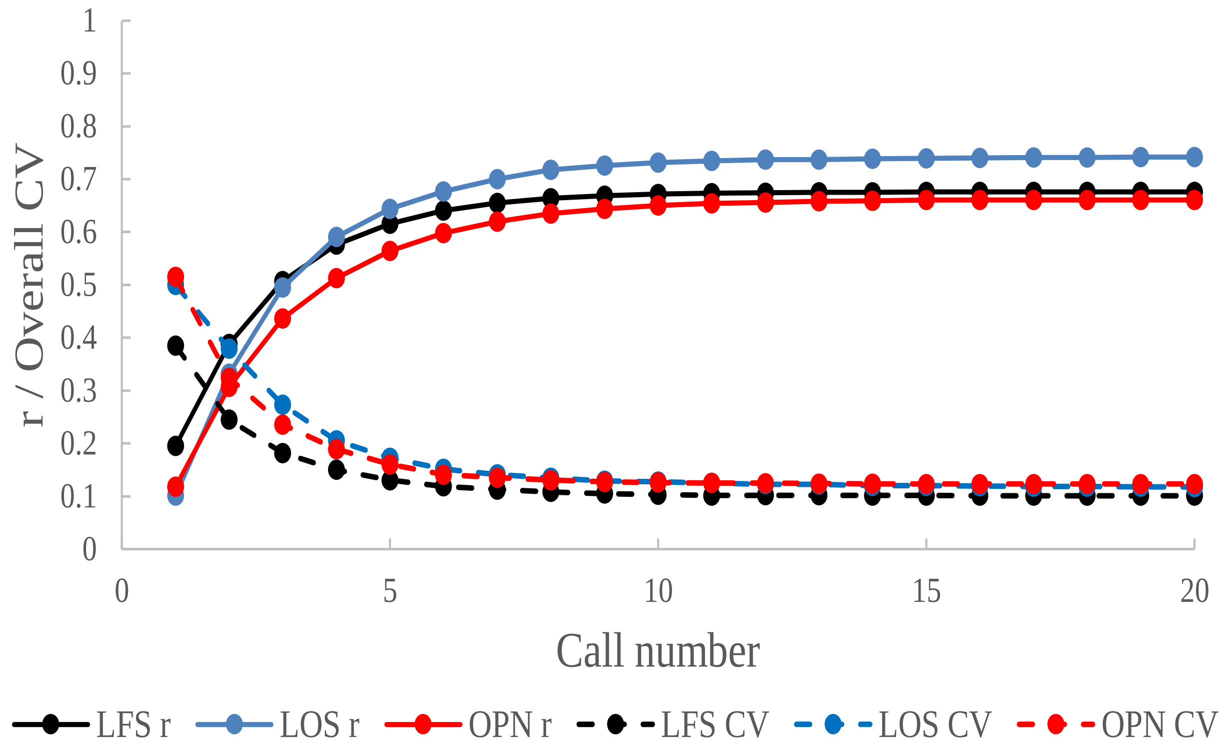
- When to change a survey data collection method?

  (Phase capacity point)

- When to stop calling?

  (Stopping point)

Stopping or Phase Capacity Points



- Adaptive Strategy: stop when indicator within 0.02 of minimum value (points later when threshold decreased)
- Responsive strategy: stop when indicator within 0.02 of previous value

Stopping or Phase Capacity (PC) Points

- Overall:

Survey	PC point (adaptive)	% calls saved	PC point (responsive)	% calls saved
LFS	6	8%	5	12%
LOS	8	15%	7	18%
OPN	6	13%	6	13%

- Also possible by variable

Further Evidence from Understanding Society

Understanding Society Data

- Longitudinal study
- Assess (risk of) nonresponse bias at each call for wave 2 for a range of survey variables as measured at wave 1

Further Data Quality Indicators

- Proposed approach
  - Dissimilarity indices (e.g. Delta index)
  - Basic idea: compare two distributions (those for respondents and those if everyone had responded)
- Comparison to
  - Coefficient of Variation (CV)

Dissimilarity Index: Categorical

- Delta index

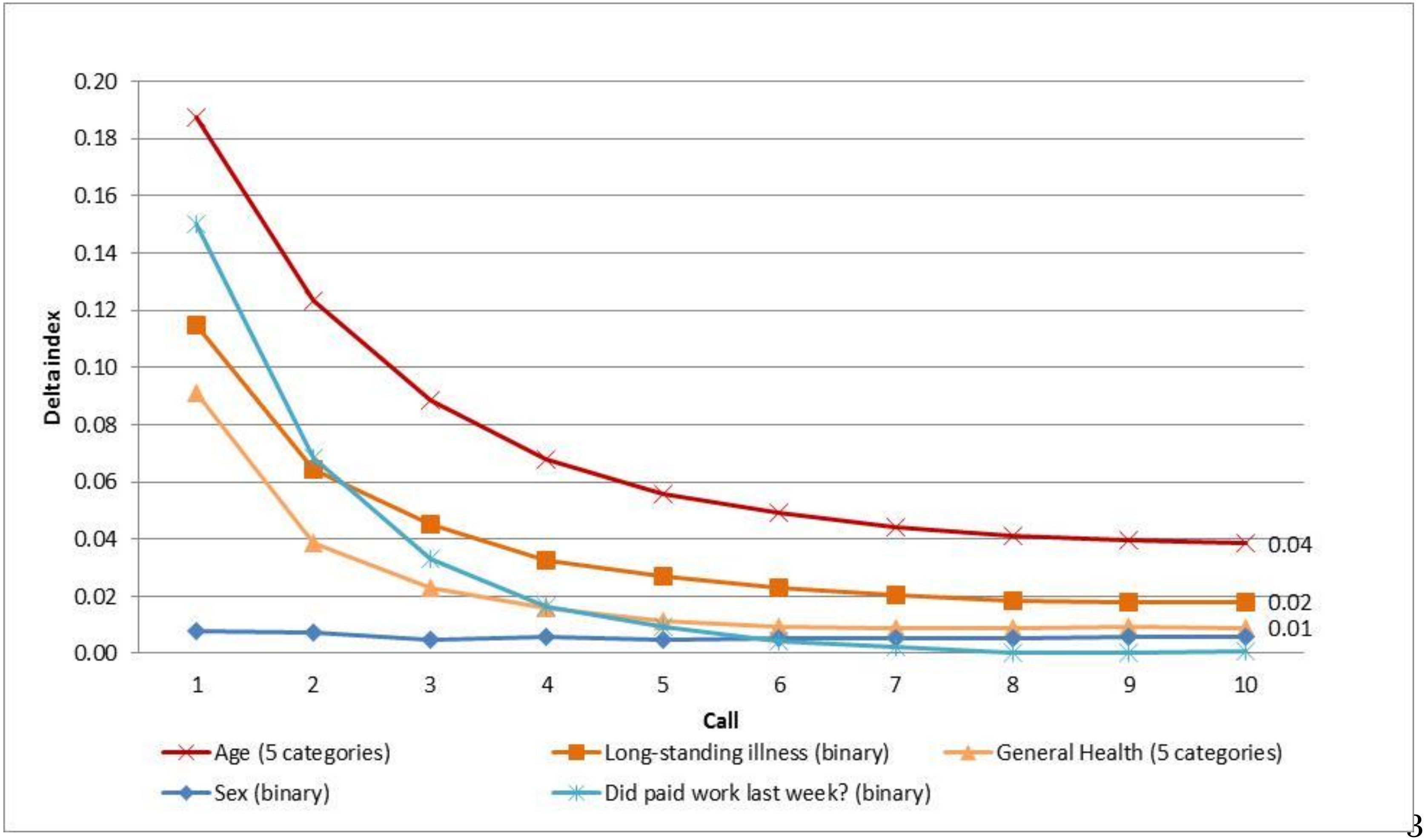
$$\Delta_z = \sum_{k=1}^{K} \left|\hat{\pi}_{z,k} - \pi_{z,k}\right|/2$$

$\hat{\pi}_{z,k}$ observed proportion in category k of survey variable z

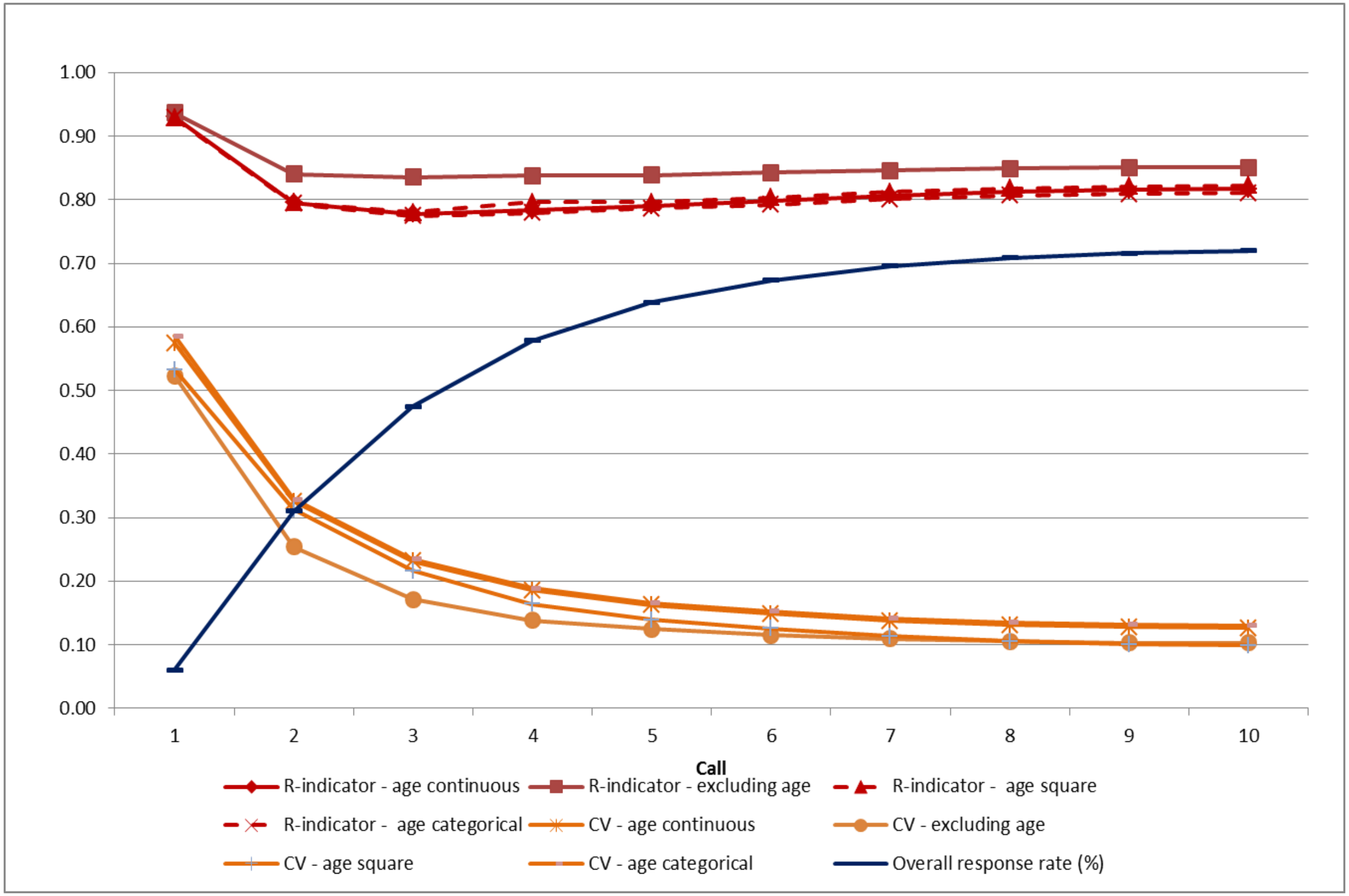
$\pi_{z,k}$ corresponding expected proportion

- ranges from 0 to 1
- the higher the delta index the more dissimilar is the estimated distribution to the true distribution
- values below 0.03 may indicate similarity (negligible nonresponse bias)
- no model required

Delta Index
Binary and Categorical Variables



Response Rate, R-indicator and CV



Summary

- Representativeness increases similarly in the surveys over call records
  - Sources of non-representativeness are under-representation of economically active HHs, HHs located in London / SE, and single adult HHs
- CV preferred over the R-indicator
- Data collection stopping points differ (slightly) between surveys
- Dissimilarity index:
  - Can monitor categorical variables with several categories
  - Allows monitoring of several variables in the same graph
  - Does not require the fit of a model at every call
- Results for CV very similar to Dissimilarity Indices – reassuring

Implications for Survey Practice

- Number of calls could be reduced (no more than 8 calls)
- Implications for cost savings without potentially much loss of data quality

Thank you.

g.durrant@southampton.ac.uk

Acknowledgements

This work contains statistical data from ONS which is Crown Copyright. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.