Indicators predicting response and data quality in Dutch person and household surveys

Sjoertje Vos, Annemieke Luiten and Nino Mushkudiani

Centraal Bureau voor de Statistiek

Two years ago....

– Can we find key indicators that predict response, representativeness and data quality in person and household surveys

  - Across modes
  - And across subgroups in the population
  - Across designs
  - Across surveys
  - Across target variables.

….

- How can we to use these indicators
  - in monitoring and managing the data collection
  - in decisions concerning modes and strategy for subgroups

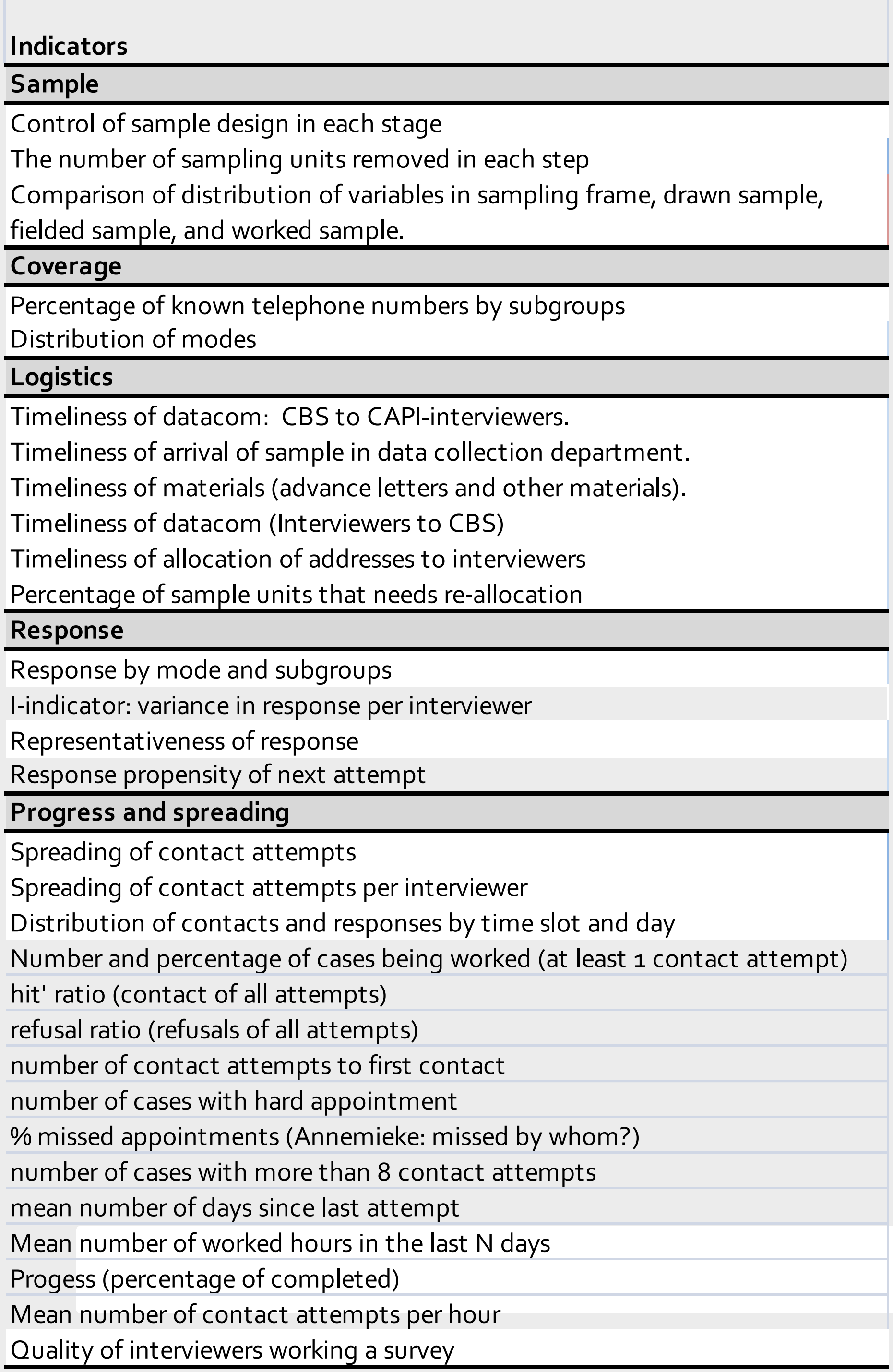
Needed:

- Frame data, auxiliary data, paradata and response data
- Real time monitoring
- Data warehouse / data store
- Automated queries
- Monitoring dashboard to visualise results

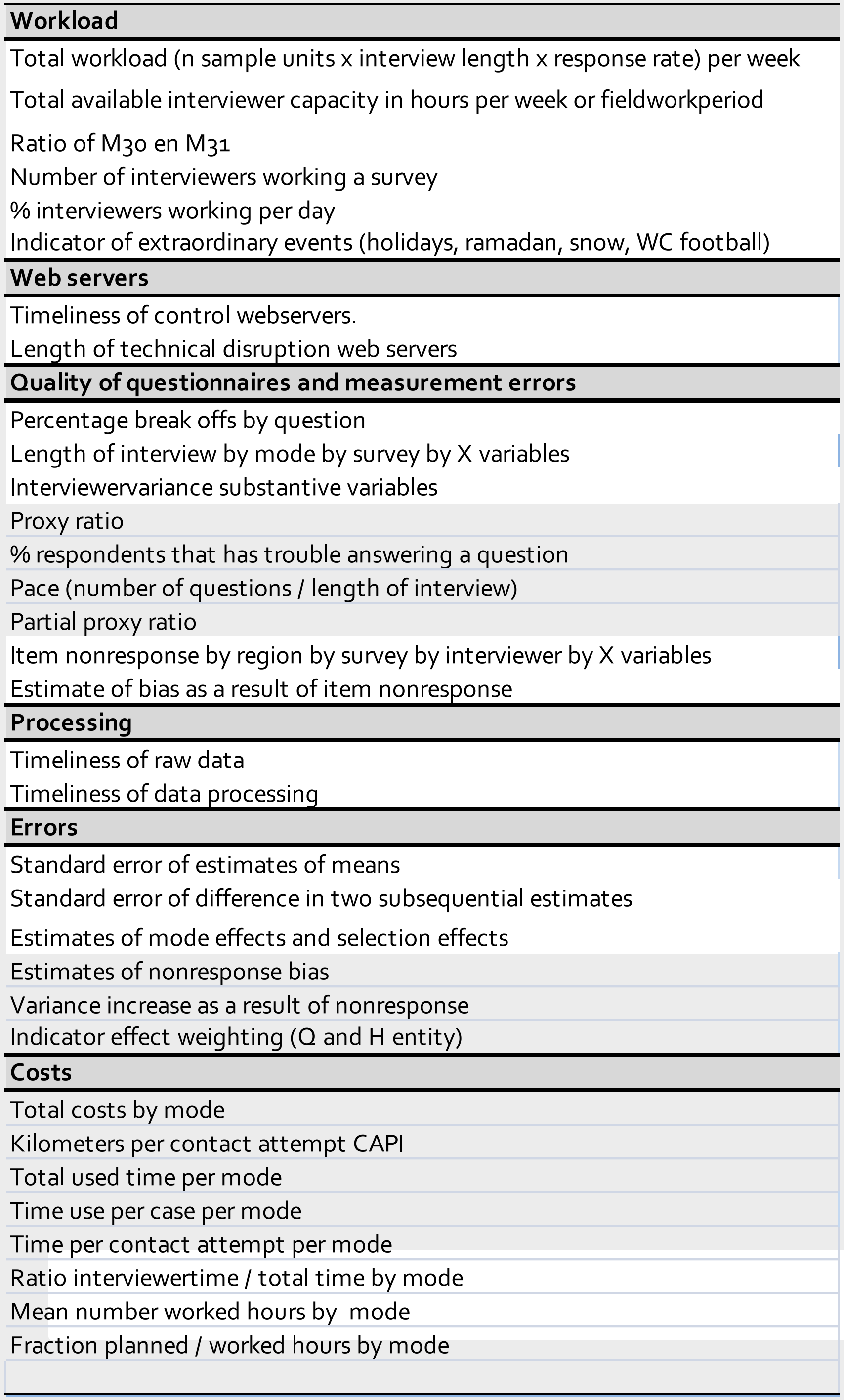
Selecting possible indicators

- Experts on data collection, methodology and subject matter rated list of indicators on
  - Relevance
  - Measurability
- Literature
- International practices

→ 63 potential indicators were selected



Indicators
Sample
Control of sample design in each stage
The number of sampling units removed in each step
Comparison of distribution of variables in sampling frame, drawn sample, fielded sample, and worked sample.
Coverage
Percentage of known telephone numbers by subgroups
Distribution of modes
Logistics
Timeliness of datacom: CBS to CAPI-interviewers.
Timeliness of arrival of sample in data collection department.
Timeliness of materials (advance letters and other materials).
Timeliness of datacom (Interviewers to CBS)
Timeliness of allocation of addresses to interviewers
Percentage of sample units that needs re-allocation
Response
Response by mode and subgroups
I-indicator: variance in response per interviewer
Representativeness of response
Response propensity of next attempt
Progress and spreading
Spreading of contact attempts
Spreading of contact attempts per interviewer
Distribution of contacts and responses by time slot and day
Number and percentage of cases being worked (at least 1 contact attempt)
hit' ratio (contact of all attempts)
refusal ratio (refusals of all attempts)
number of contact attempts to first contact
number of cases with hard appointment
% missed appointments (Annemieke: missed by whom?)
number of cases with more than 8 contact attempts
mean number of days since last attempt
Mean number of worked hours in the last N days
Progess (percentage of completed)
Mean number of contact attempts per hour
Quality of interviewers working a survey



Workload
Total workload (n sample units x interview length x response rate) per week
Total available interviewer capacity in hours per week or fieldworkperiod
Ratio of M30 en M31
Number of interviewers working a survey
% interviewers working per day
Indicator of extraordinary events (holidays, ramadan, snow, WC football)
Web servers
Timeliness of control webservers.
Length of technical disruption web servers
Quality of questionnaires and measurement errors
Percentage break offs by question
Length of interview by mode by survey by X variables
Interviewervariance substantive variables
Proxy ratio
% respondents that has trouble answering a question
Pace (number of questions / length of interview)
Partial proxy ratio
Item nonresponse by region by survey by interviewer by X variables
Estimate of bias as a result of item nonresponse
Processing
Timeliness of raw data
Timeliness of data processing
Errors
Standard error of estimates of means
Standard error of difference in two subsequential estimates
Estimates of mode effects and selection effects
Estimates of nonresponse bias
Variance increase as a result of nonresponse
Indicator effect weighting (Q and H entity)
Costs
Total costs by mode
Kilometers per contact attempt CAPI
Total used time per mode
Time use per case per mode
Time per contact attempt per mode
Ratio interviewertime / total time by mode
Mean number worked hours by mode
Fraction planned / worked hours by mode

Indicators
Sample
Control of sample design in each stage
The number of sampling units removed in each step
Comparison of distribution of variables in sampling frame, drawn sample, fielded sample, and worked sample.
Coverage
Percentage of known telephone numbers by subgroups
Distribution of modes
Logistics
Timeliness of datacom: CBS to CAPI-interviewers.
Timeliness of arrival of sample in data collection department.
Timeliness of materials (advance letters and other materials).
Timeliness of datacom (Interviewers to CBS)
Timeliness of allocation of addresses to interviewers
Percentage of sample units that needs re-allocation
Response
Response by mode and subgroups
I-indicator: variance in response per interviewer
Representativeness of response
Response propensity of next attempt
Progress and spreading
Spreading of contact attempts
Spreading of contact attempts per interviewer
Distribution of contacts and responses by time slot and day
Number and percentage of cases being worked (at least 1 contact attempt)
hit' ratio (contact of all attempts)
refusal ratio (refusals of all attempts)
number of contact attempts to first contact
number of cases with hard appointment
% missed appointments
number of cases with more than 8 contact attempts
mean number of days since last attempt
Mean number of worked hours in the last N days
Progess (percentage of completed)
Mean number of contact attempts per hour
Quality of interviewers working a survey

Workload

Total workload (n sample units x interview length x response rate) per week
Total available interviewer capacity in hours per week or fieldworkperiod
Ratio of M30 en M31
Number of interviewers working a survey
% interviewers working per day
Indicator of extraordinary events (holidays, ramadan, snow, WC football)

Web servers

Timeliness of control webservers.
Length of technical disruption web servers

Quality of questionnaires and measurement errors

Percentage break offs by question
Length of interview by mode by survey by X variables
Interviewervariance substantive variables
Proxy ratio
% respondents that has trouble answering a question
Pace (number of questions / length of interview)
Partial proxy ratio
Item nonresponse by region by survey by interviewer by X variables
Estimate of bias as a result of item nonresponse

Processing

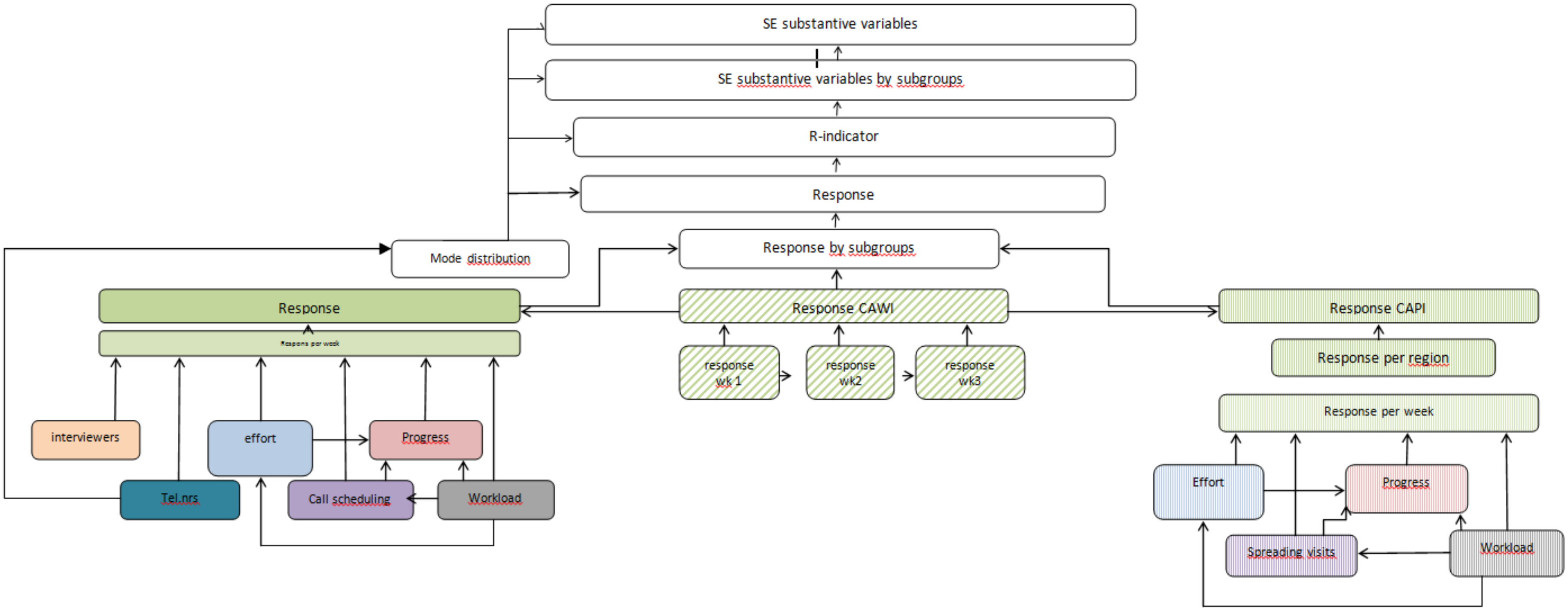
Timeliness of raw data
Timeliness of data processing

Errors

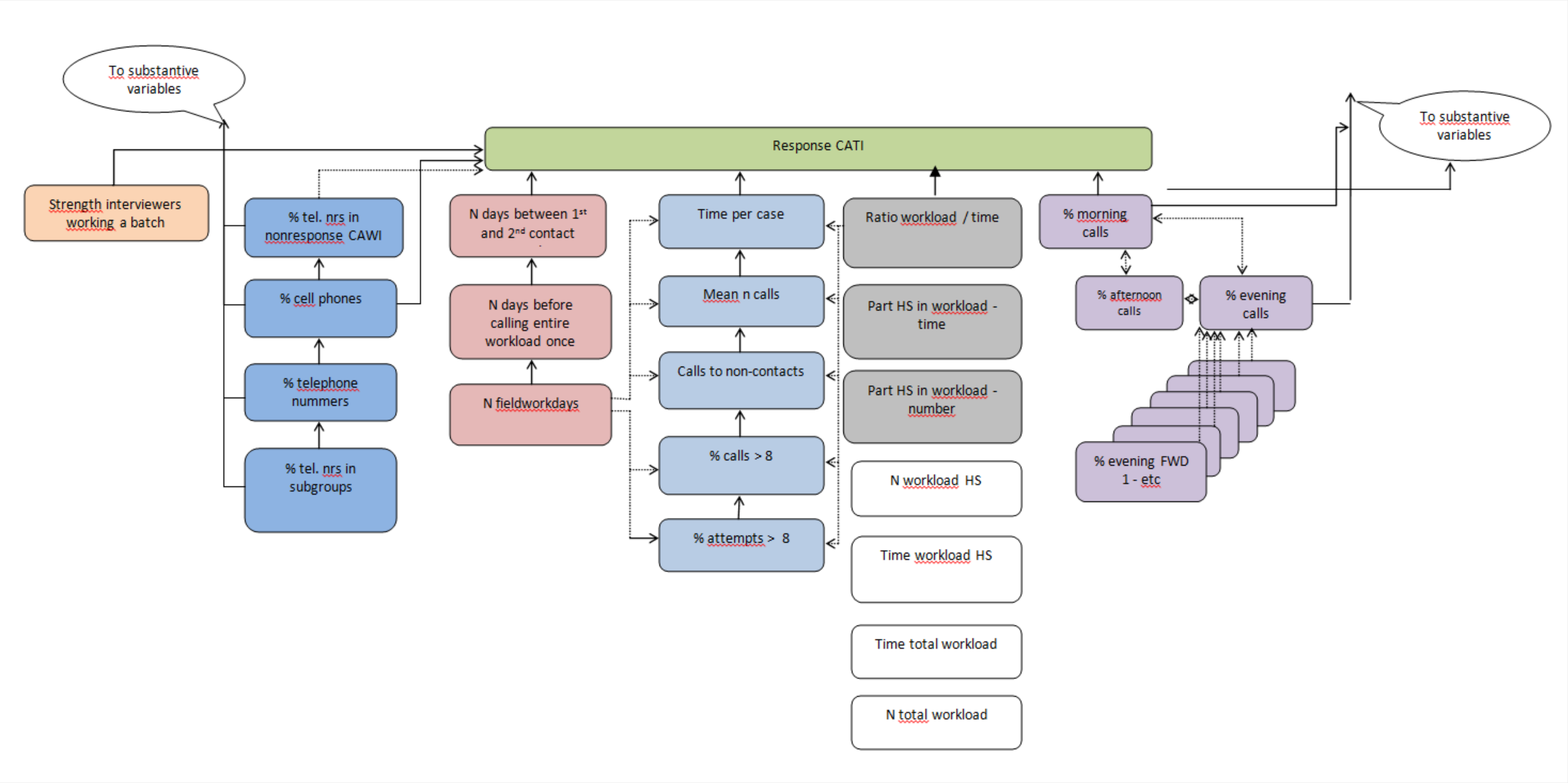
Standard error of estimates of means
Standard error of difference in two subsequential estimates
Estimates of mode effects and selection effects
Estimates of nonresponse bias
Variance increase as a result of nonresponse
Indicator effect weighting (Q and H entity)

Costs

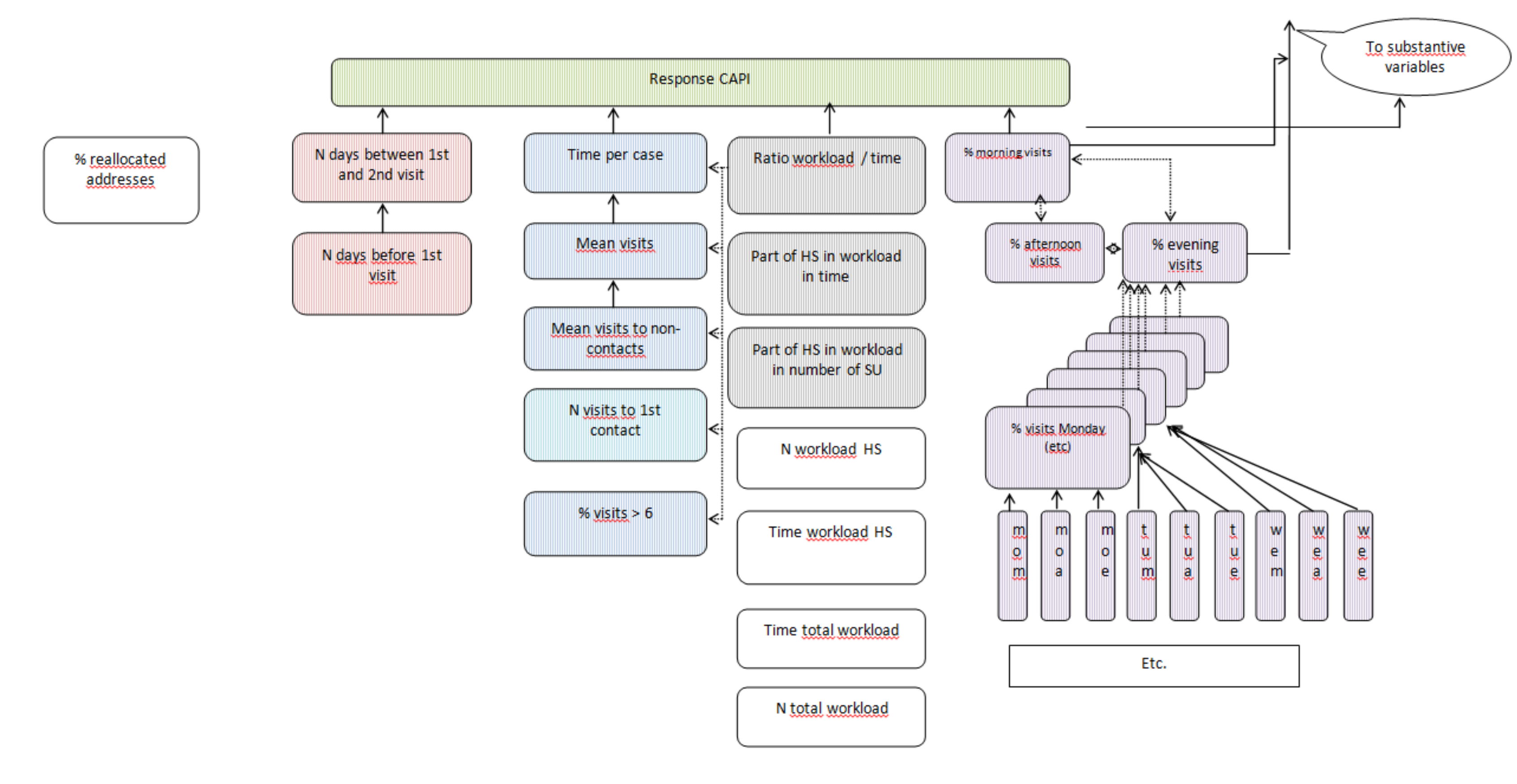
Total costs by mode
Kilometers per contact attempt CAPI
Total used time per mode
Time use per case per mode
Time per contact attempt per mode
Ratio interviewertime / total time by mode
Mean number worked hours by mode
Fraction planned / worked hours by mode



Characteristics of sample



Characteristics of sample



Response CAPI

% reallocated addresses

N days between 1st and 2nd visit

N days before 1st visit

Time per case

Mean visits

Mean visits to non-contacts

N visits to 1st contact

% visits > 6

Ratio workload / time

Part of HS in workload in time

Part of HS in workload in number of SU

N workload HS

Time workload HS

Time total workload

N total workload

% morning visits

% afternoon visits

% evening visits

To substantive variables

% visits Monday (etc)

mom

moa

moe

tum

tua

tue

wem

wea

wee

Etc.

Characteristics of sample

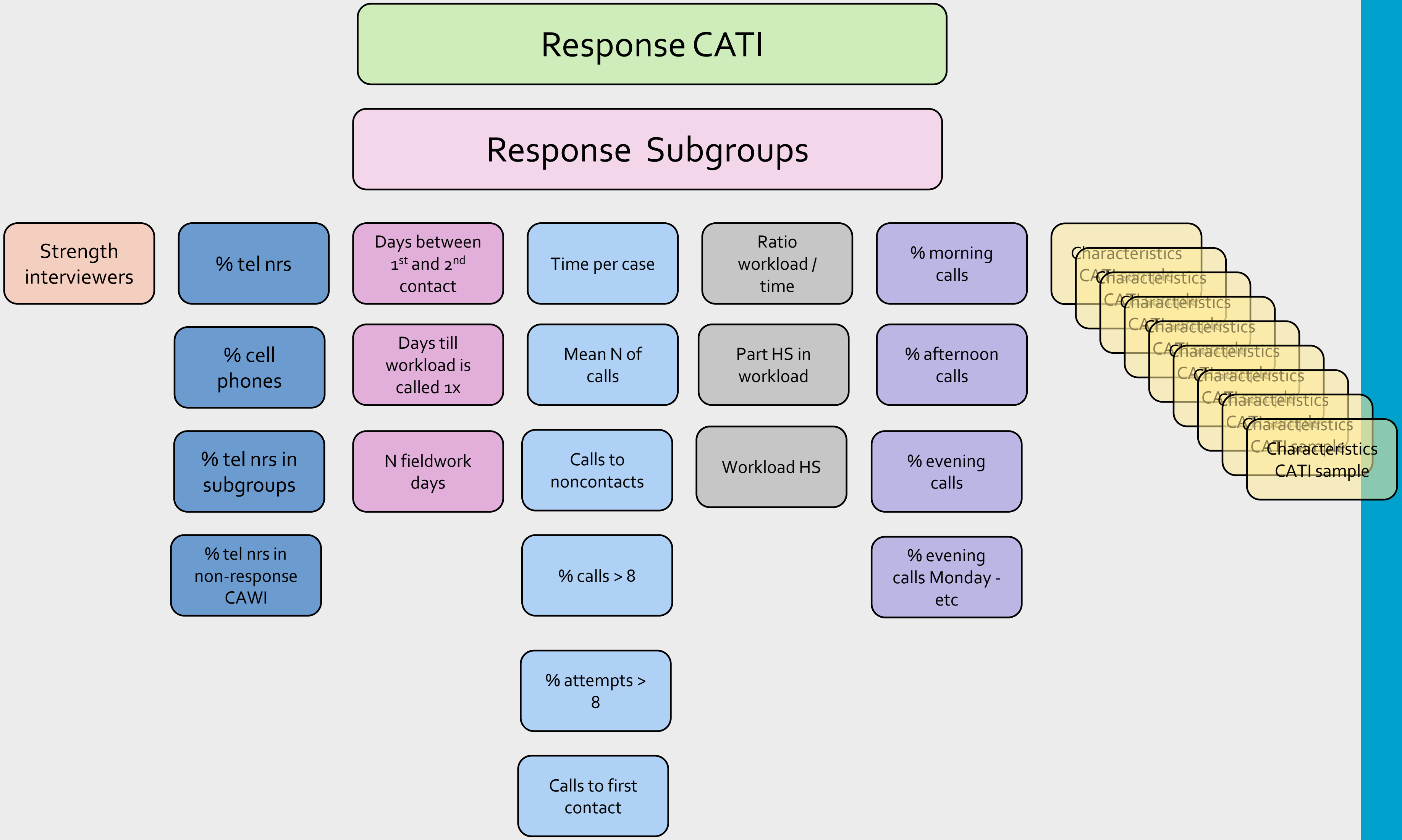
Method

- So far we calculated the 'blue' indicators for two surveys in two designs:
  - LFS: cati – capi; 33 months
  - LFS: web – cati – capi; 35 months
  - Health Survey: web – cati – capi; 47 months
  - Health Survey: web – capi; 13 months
- Even the limited set consisted of 580 (sub)indicators
- Linking the files and calculating the indicators took an enormous amount of time
  - New data store will ease this task in future

- Conceptual model led analysis
    - Modelling CATI response
    - Modelling CAPI response
    - Modelling CAWI response
    - CATI - CAPI - CAWI response underlies substantive variables
- First: identification of univariate relations
    - Within and across surveys and designs
- Multivariate analysis
    - Choose best model (lowest AIC) on all combinations of covariates
    - If more than ± 15 covariates: define core model with most significant univariate covariates as default and add all possible combinations of the rest

Results

Predicting CATI response – conceptual model – Health Survey



Response CATI

Response Subgroups

Strength interviewers

% tel nrs

% cell phones

% tel nrs in subgroups

% tel nrs in non-response CAWI

Days between 1st and 2nd contact

Days till workload is called 1x

N fieldwork days

Time per case

Mean N of calls

Calls to noncontacts

% calls > 8

% attempts > 8

Calls to first contact

Ratio workload / time

Part HS in workload

Workload HS

% morning calls

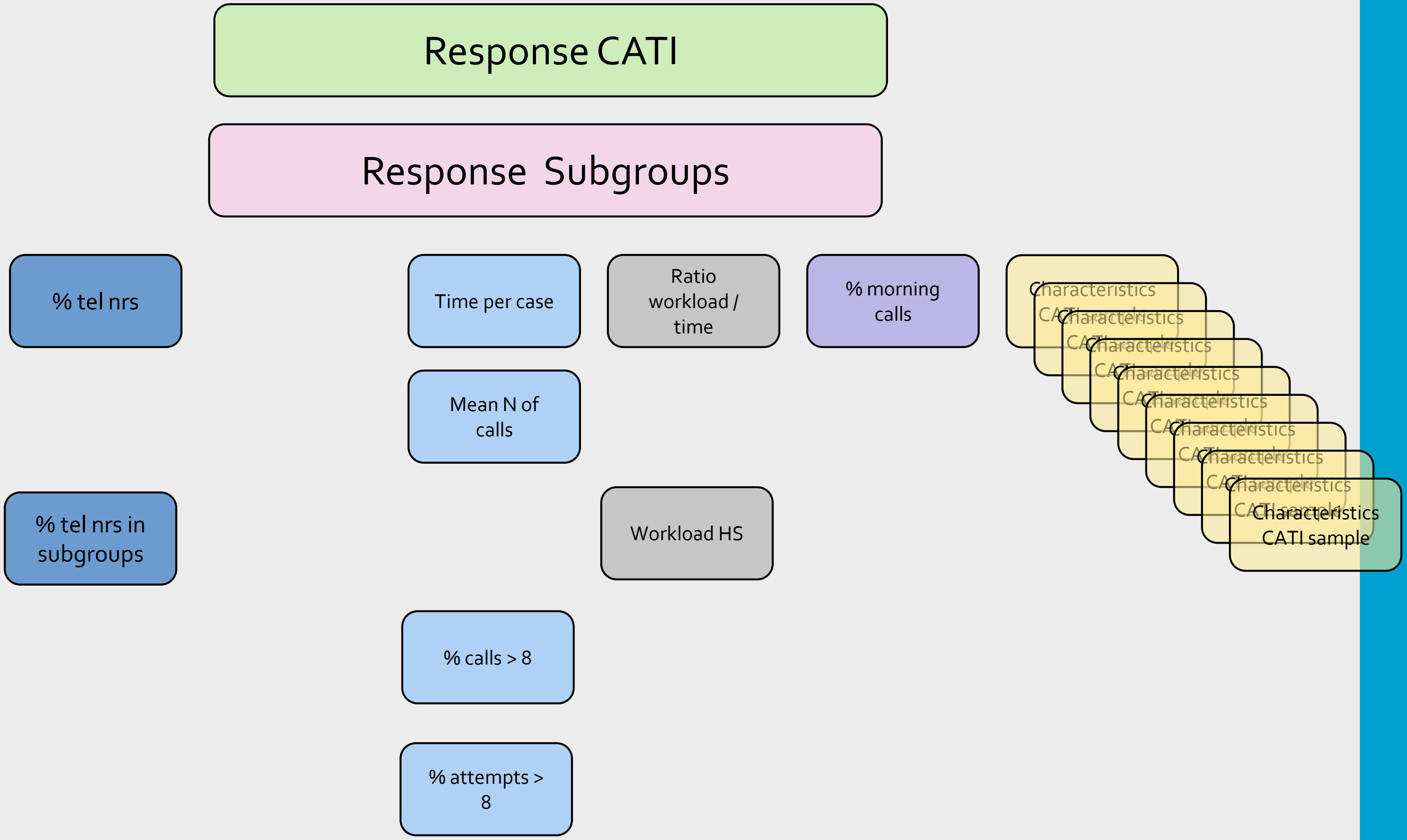
% afternoon calls

% evening calls

% evening calls Monday - etc

Characteristics CATI sample

Predicting CATI – univariate relations



Response CATI

Response Subgroups

% tel nrs

Time per case

Ratio workload / time

% morning calls

Characteristics CATI sample

Mean N of calls

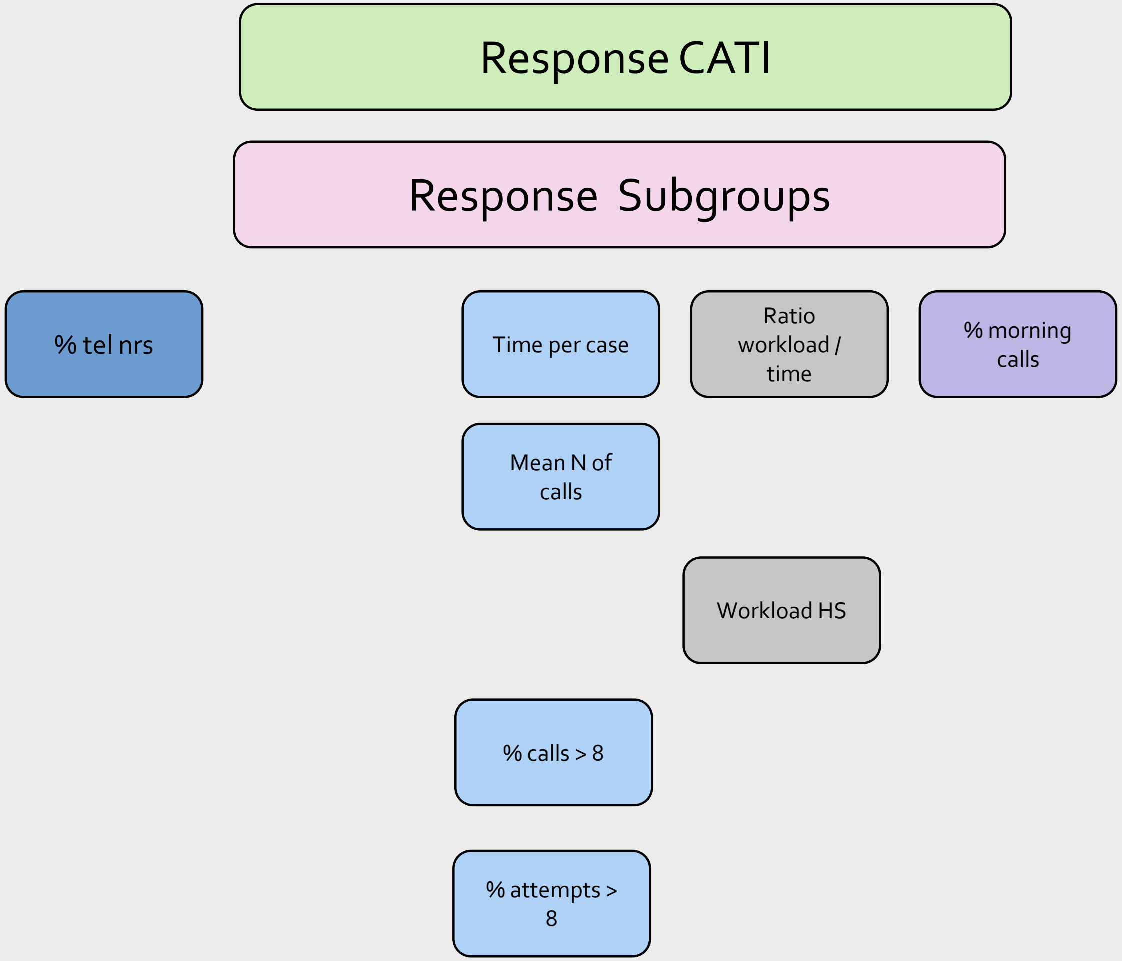
% tel nrs in subgroups

Workload HS

% calls > 8

% attempts > 8

Predicting CATI – best model



Response CATI

Response Subgroups

% tel nrs

Time per case

Ratio workload / time

% morning calls

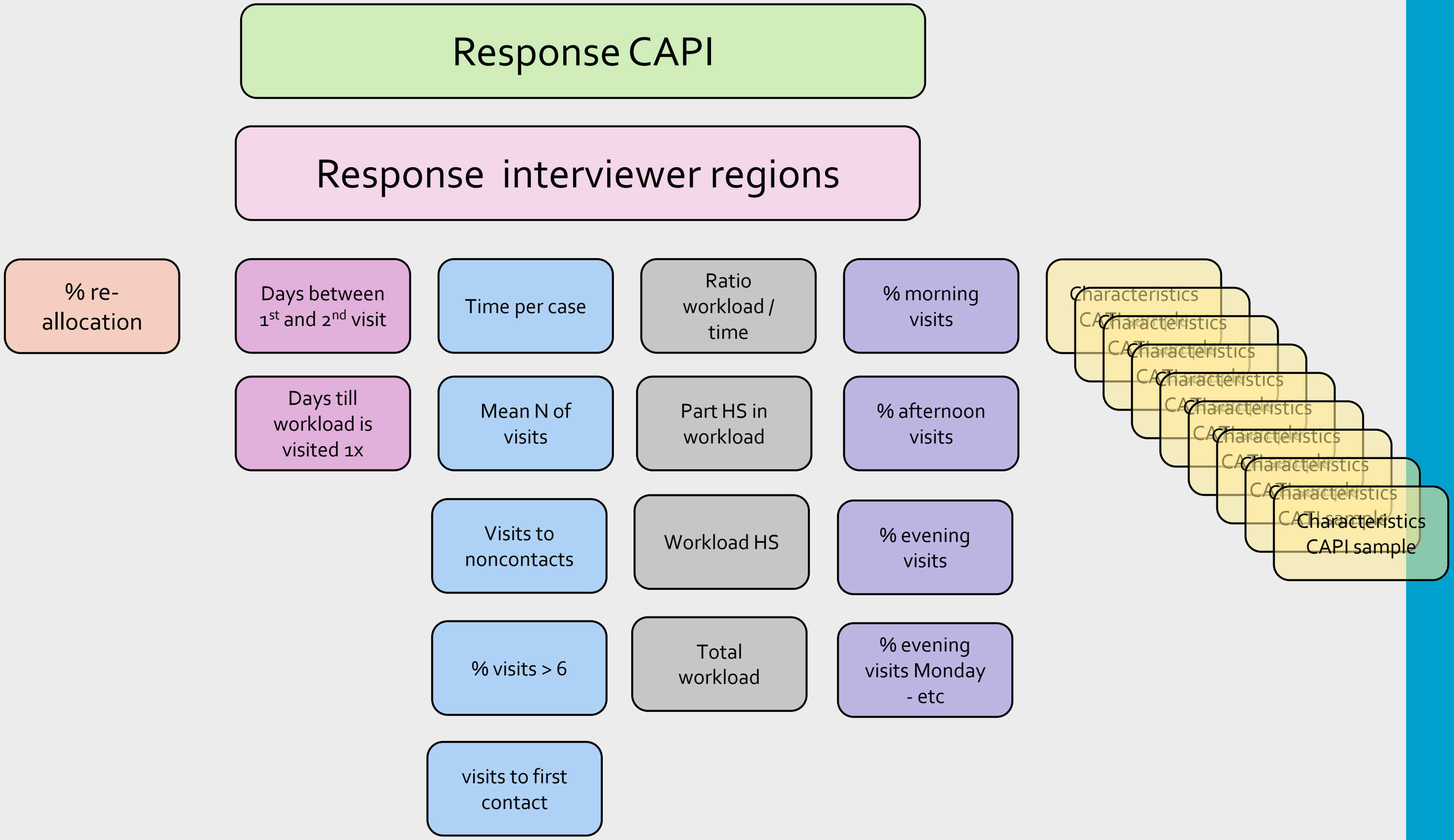
Mean N of calls

Workload HS

% calls > 8

% attempts > 8

Predicting CAPI response, contact and cooperation – conceptual model



Response CAPI

Response interviewer regions

% re-allocation

Days between 1st and 2nd visit

Days till workload is visited 1x

Time per case

Mean N of visits

Visits to noncontacts

% visits > 6

visits to first contact

Ratio workload / time

Part HS in workload

Workload HS

Total workload

% morning visits

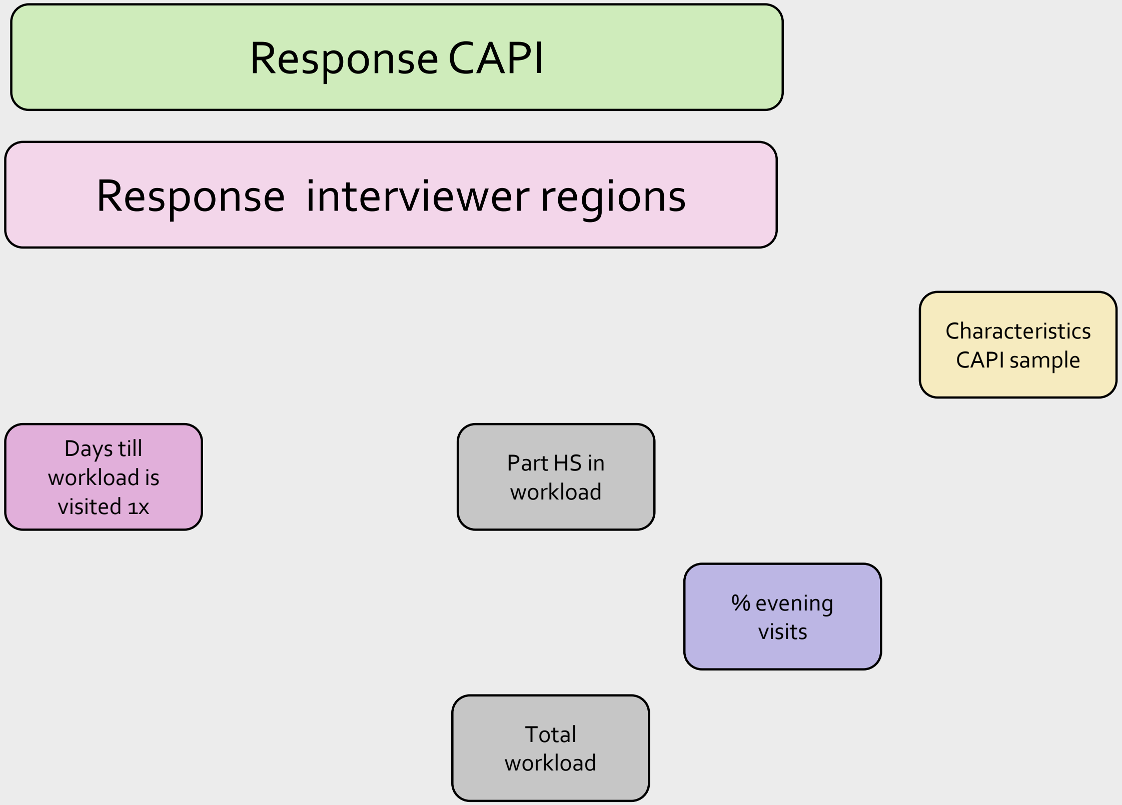
% afternoon visits

% evening visits

% evening visits Monday - etc

Characteristics CAPI sample

Predicting CAPI response, contact and cooperation – univariate relations



Response CAPI

Response interviewer regions

Characteristics CAPI sample

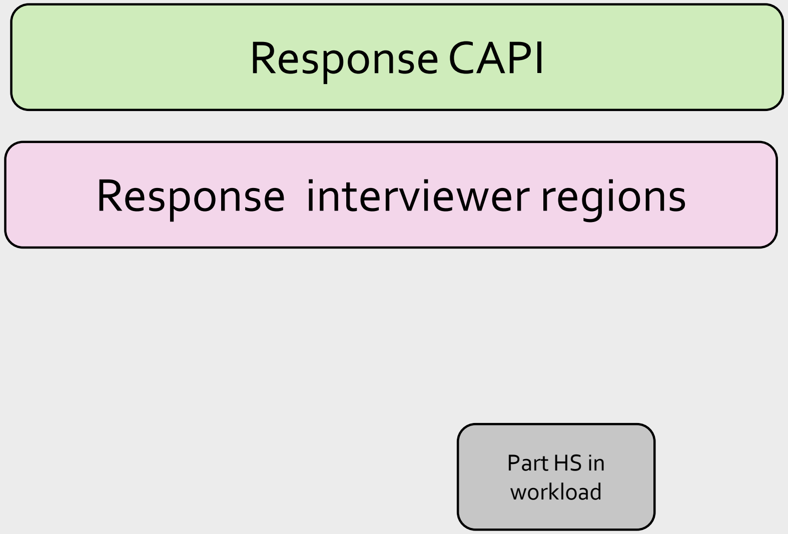
Days till workload is visited 1x

Part HS in workload

% evening visits

Total workload

Predicting CAPI response – best model

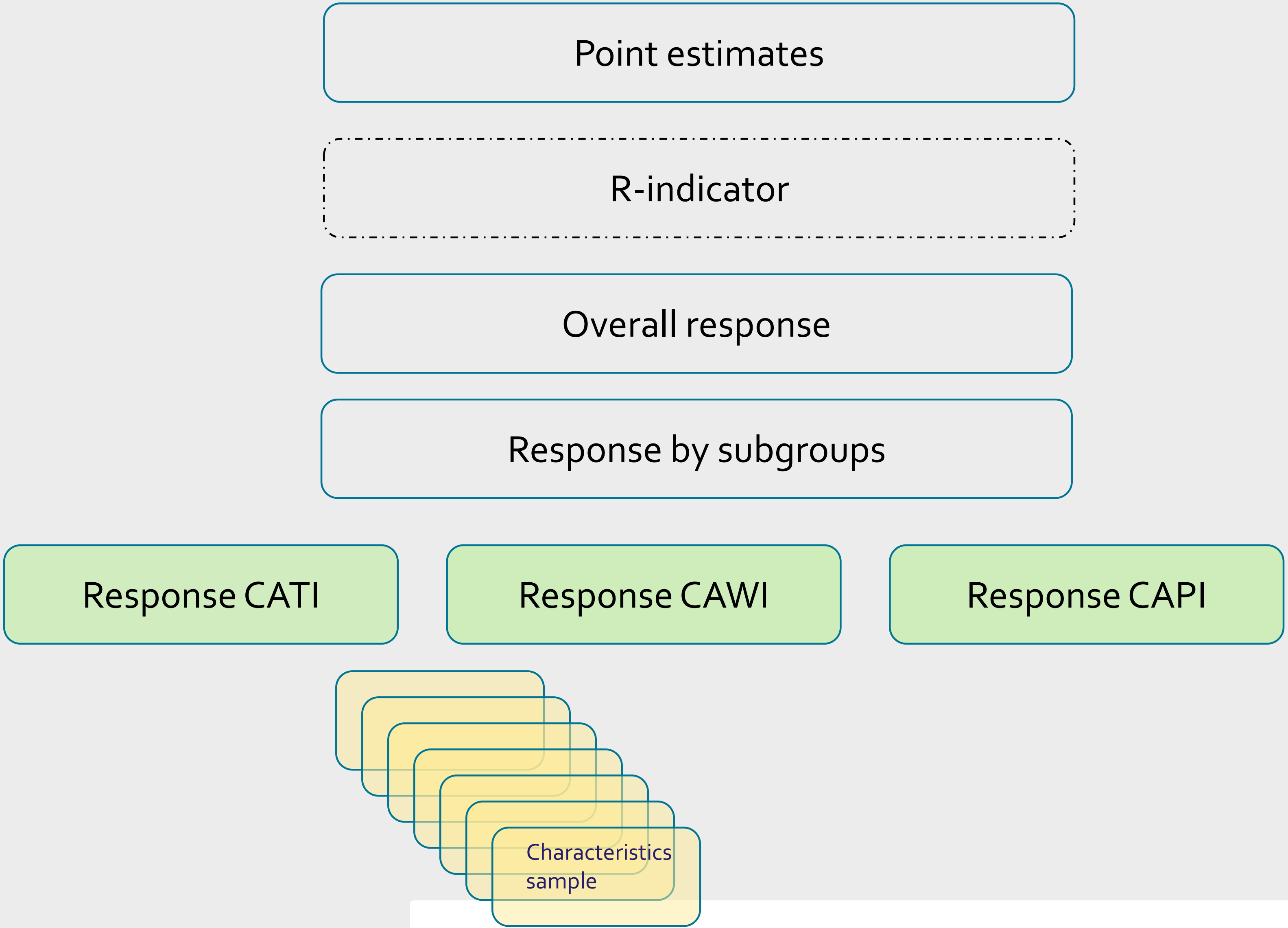


Response CAPI

Response interviewer regions

Part HS in workload

Predicting point estimates - conceptual model



Point estimates

R-indicator

Overall response

Response by subgroups

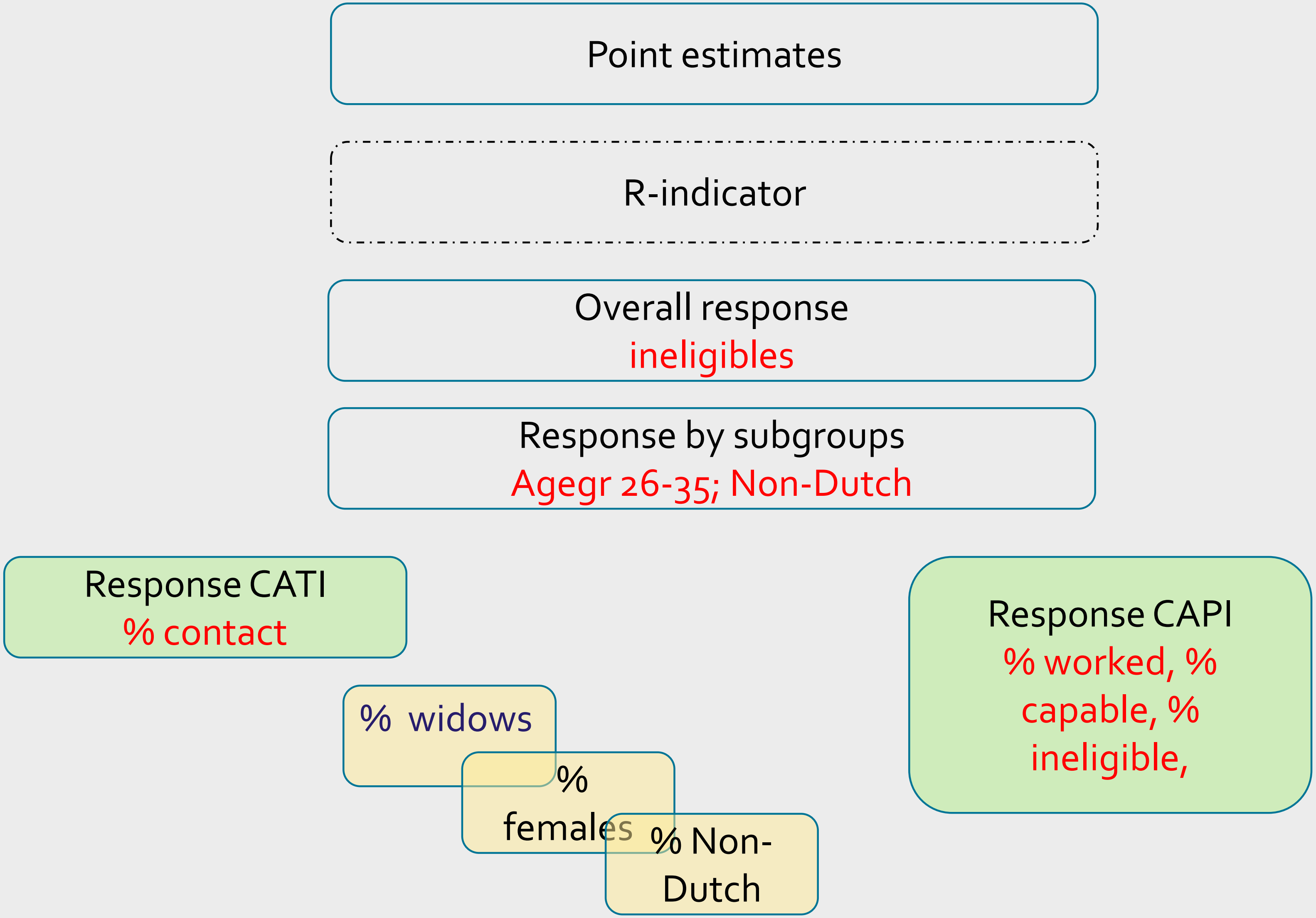
Response CATI

Response CAWI

Response CAPI

Characteristics sample

Predicting point estimates - univariate relations & final model



Point estimates

R-indicator

Overall response
ineligibles

Response by subgroups
Agegr 26-35; Non-Dutch

Response CATI
% contact

% widows

% females

% Non-Dutch

Response CAPI
% worked, % capable, % ineligible,

SE by subgroups: education

SE point estimates by subgroup		
lowest education	middle education	highest education
v39	v228	v50
v47	v311	v425
v48	v326	
v54	v153	
v56	v408	
v67		
v70		
v317		
v325		
v326		
v327		
v334		
v407		
v408		
v416		
v419		
v421		
v425		
v426		
R^2_{adj} = .741	R^2_{adj} = .637	R^2_{adj} = .172

(preliminary) conclusions (1)

- Design (change) has large impact on indicators
  → re-evaluate your indicators after change
- Large differences between subgroups in relevant indicators
  →May mean we still need a lot of indicators
- What happens in the field has impact on weighted point estimates and variance estimates
- Workload is one of the most consistently relevant indicators
- What happens in the interviewer regions has large influence on end result; don’t know yet what determines region response rates

(preliminary) conclusions (2)

- Sample fluctuations (= what happens in CAWI) influence CATI and CAPI response

  → Weighted response rates are needed to compare monthly results
- Still small N with many indicators
  - Univariatly high correlations don’t end up in model. Power issue or really not important?

  → Keep building
- From indicator to dashboard
- To be continued….

Thank you!

- Questions?
- Suggestions on how to proceed?