# A Bayesian analysis of survey design parameters for nonresponse, costs and survey outcome variable models

Eva de Jong, Nino Mushkudiani and Barry Schouten

**ASD workshop, November 6 - 8, 2017**

The Leverhulme Trust

Centraal Bureau voor de Statistiek

# Outline

- Bayesian analysis framework revisited;
- Extension to survey outcome variables;
- Research questions;
- Results based on Dutch Health Survey;
- Discussion;

# Key elements to ASD

1. Auxiliary information (frame, administrative data, paradata);
2. Design features or treatments (e.g. modes or number of calls);
3. Quality and cost functions;
4. Optimization strategy:

In a Bayesian analysis, the performance of different treatments for different population strata is modeled and updated through prior distributions and observed data.

As a result, quality and cost functions become random variables, and the optimization strategy needs to account for uncertainty. .

# Bayesian analysis for response and costs

## Strategy

Regression coefficients and variances in contact, participation and costs models are assigned a distribution (prior) that is updated using survey data (posterior). Posterior is the new prior for a next round or wave.

## Elicitation of prior distribution parameters (hyperparameters):

- Expert knowledge;
- Historic survey data

## Numerical approximation of posterior distribution through MCMC;

Schouten, Mushkudiani, Shlomo, Durrant, Lundquist, Wagner (2017)

# Bayesian ASD – research areas

- Prior elicitation from expert knowledge;
- Time change in survey design parameters;
- Extension to survey outcome variables and quality indicators;
- Optimization of ASD;

# Model - survey design parameters

Three types of survey design parameters are sufficient to compute most quality and cost functions:

- $\rho_i(s)$ : Response propensity for a unit and strategy;
- $C_i(s)$ : Costs for a unit and strategy;
- $D_i(y, s)$ : Method effect on $y$ for a unit and strategy;

For interviewer modes, response propensities and costs are detailed for different types of nonresponse.

Design parameters are modelled through generalized linear models using a selection of the available covariates.

# Model – extension to survey variables

Model based on covariates only
Continuous survey variable ($Y_k$):

$$Y_{k,i}(s) = \theta_{0,k}(s)x_i + \varepsilon_i(s)$$

Categorical: Through latent variable and link function

Model based on covariates and other survey variables ($Y_l$)
Continuous survey variable ($Y_k$):

$$Y_{k,i}(s) = \theta_{0,k}(s)x_i + \theta_{1,k}(s)Y_{l,i}(s) + \varepsilon_i(s)$$

Categorical: Again through latent variable and link function

# Model – extension to survey variables

Expected value for survey variable $Y_k$ for a strategy $s_{1,T}$ is a weighted mix of the expected values per action

$$Y_{k,i}(s_{1,T}) = \frac{1}{\rho_{k,i}(s_{1,T})}\left(\rho_{1,i}(s_1)Y_{k,i}(s_1) + \sum_{t=2}^{T}\prod_{l=1}^{t-1}\left(1-\rho_{l,i}(s_{1,l})\right)\rho_{t,i}(s_{1,t})Y_{k,i}(s_t)\right).$$

Expected values can be computed for
- other actions (potential outcomes) → MAR(Y,X) assumption
- nonrespondents → MAR(Y,X) assumption

Current model: Independence in $\theta_{0,k}(s)$ and $\theta_{1,k}(s)$

# Quality and cost – covariate-based (type 1)

Examples ($d_i$= sample inclusion weight):

- Response rate: $RR(s) = \frac{1}{N}\sum_{i=1}^{n} d_i \rho_i\,(s)$
- Total costs: $B(s) = \sum_{i=1}^{n} c_i\,(s)$
- R-indicator of response propensities for relevant X

$$R(X,s) = 1 - 2\sqrt{\frac{1}{N}\sum_{i=1}^{n} d_i(\rho_i\,(s) - RR(s))^2}$$

- CV of response propensities for relevant X

$$CV(X,s) = \frac{1-R(X,s)}{2RR(s)}$$

# Quality and cost – item-based (type 2)

- Method effect towards a benchmark strategy BM:

$$D(Y_k, X, s; BM) = \left| \frac{\sum_{i=1}^{n} d_i \rho_i(s) \left(Y_{k,i}(s) - Y_{k,i}(BM)\right)}{\sum_{i=1}^{n} d_i \rho_i(s)} \right|$$

- Fraction Missing Information (FMI) for relevant X:

$$\mathrm{FMI}(Y_k, X, s) = \frac{(1 - RR(s)) 1 - CD(Y_k(s), X))}{RR(s) + (1 - RR(s)(1 - CD(Y_k(s), X))}$$

- Predicted nonresponse bias (NRB) for relevant X:

$$\mathrm{NRB}(Y_k, X, s) = \frac{\mathrm{cov}(Y_k(s), \rho(s))}{RR(s)}$$

# Research questions

1. What are properties of the posteriors of the survey variable quality indicators?
2. Is their added value in including other survey variables in the models?
3. How to employ these indicators in conjunction with the covariate-based indicators?

# Case study - Dutch Health Survey

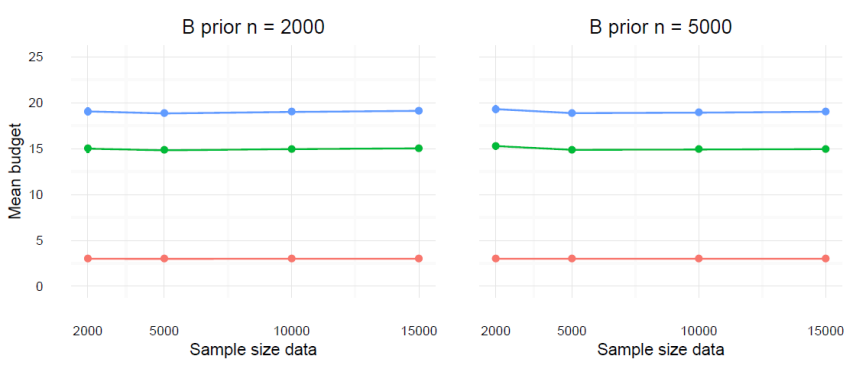
Features:

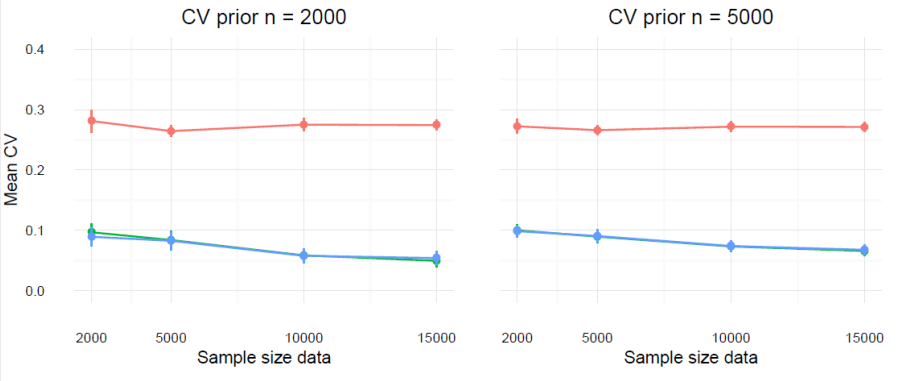
- Health survey: monthly, on-going person survey;
- Three phases Web → F2F follow up → extended F2F follow-up, where phases 2 and 3 are optional;
- Stratification based on age, gender and yes/no Web break-off;
- Survey variables: Yes/no good health, BMI, and yes/no smoking;

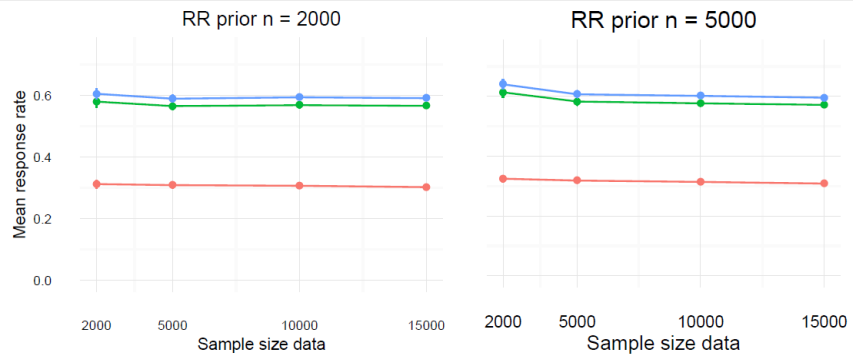
- Two priors (based on historic data of size n=2000, 5000) and four sample sizes (n=2000, 5000, 10000, 15000);

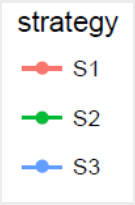
# Case study - Properties

## Properties cost and response indicators



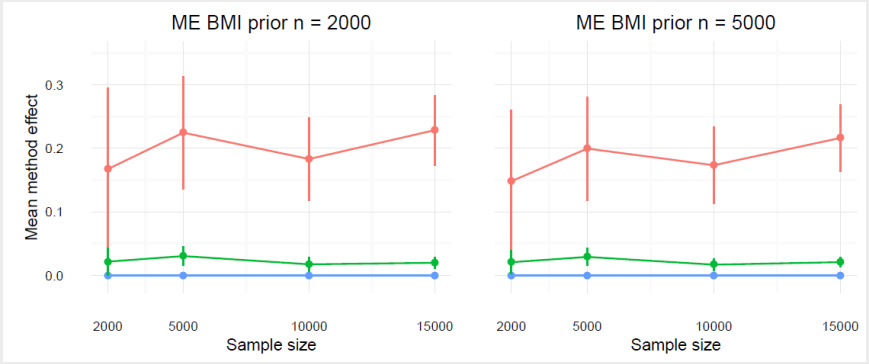



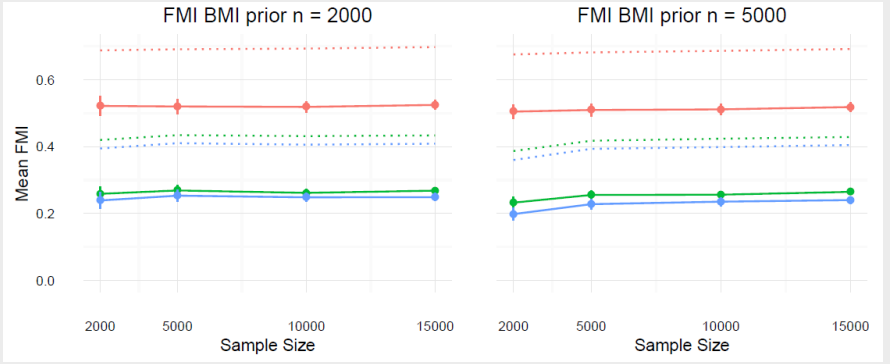



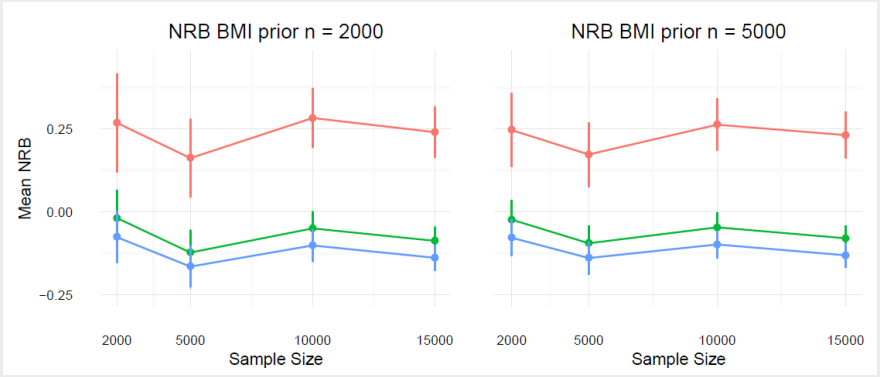



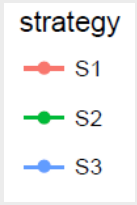
# Case study - Properties

## Properties method effect, FMI and NRB for BMI



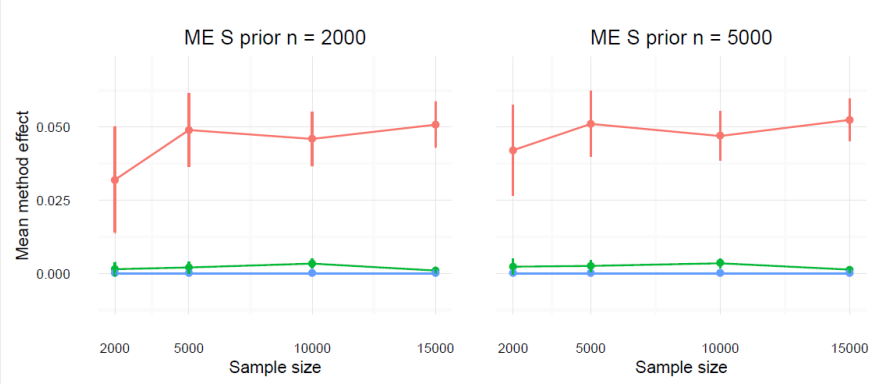



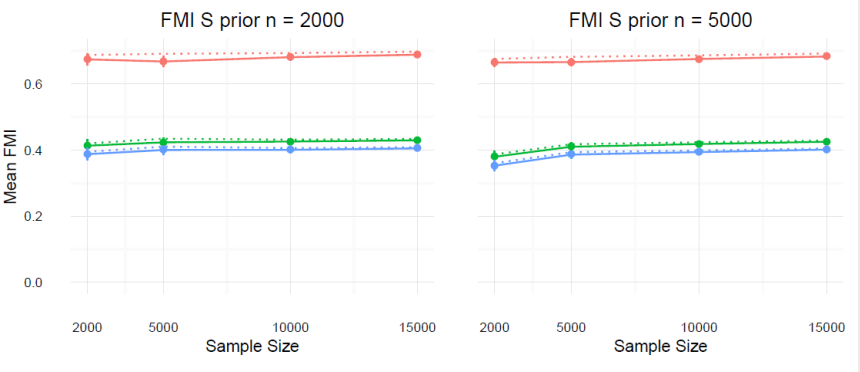



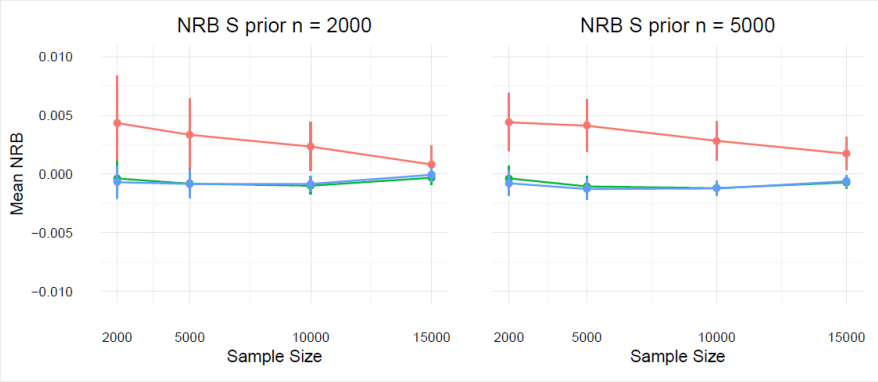



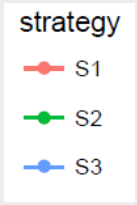
# Case study - Properties

## Properties method effect, FMI and NRB for smoking



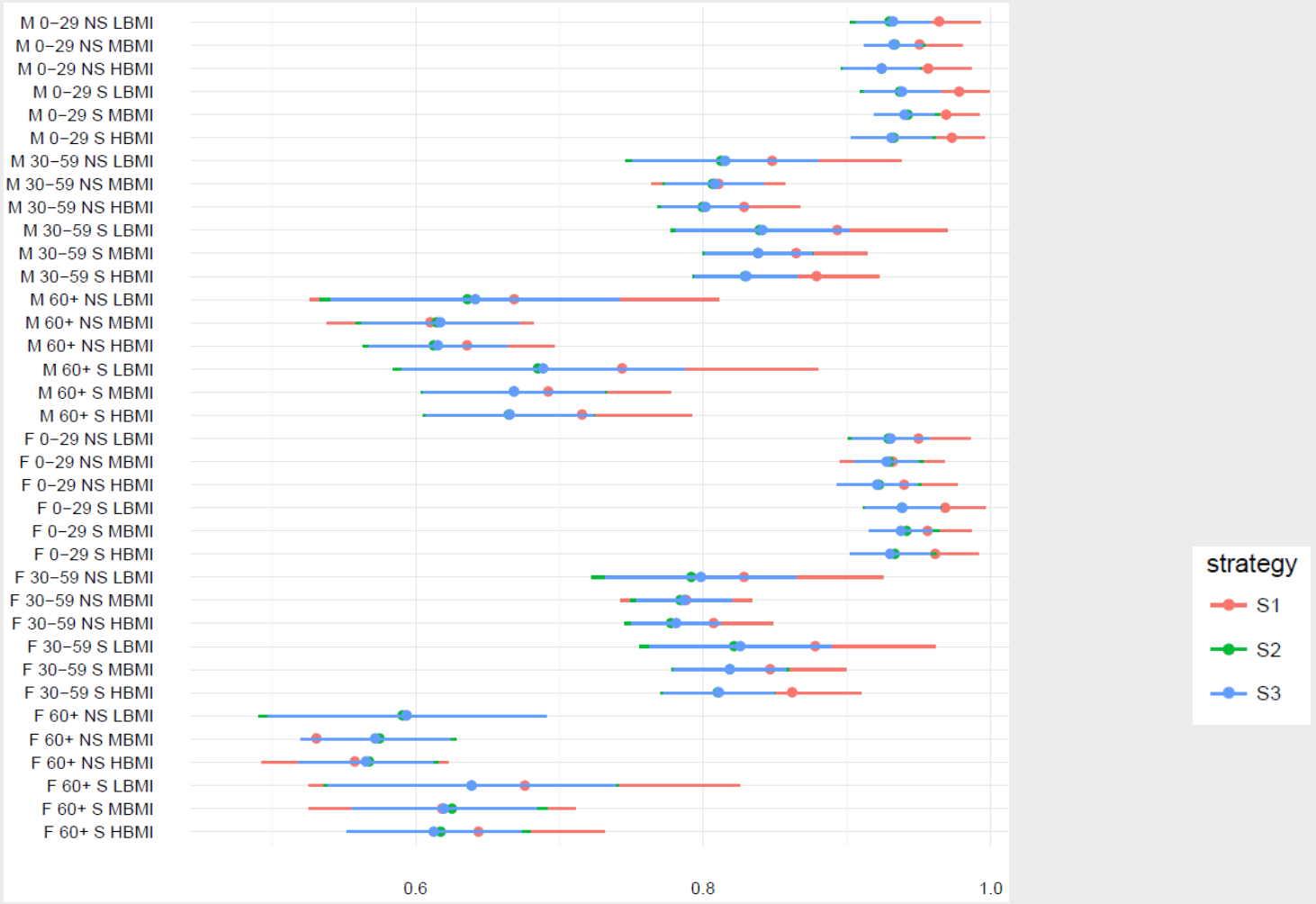



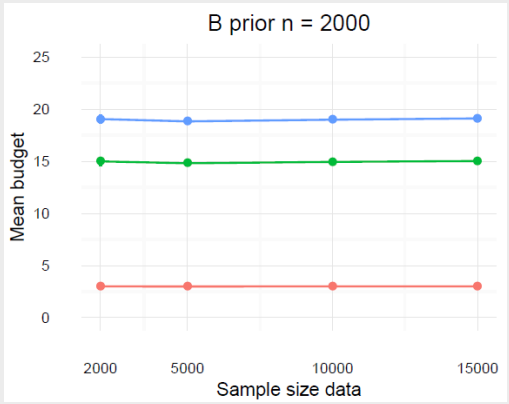
# Case study – Models

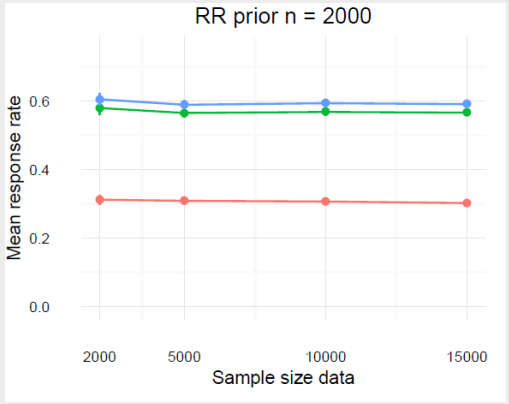
Models for yes/no good health including BMI and smoking

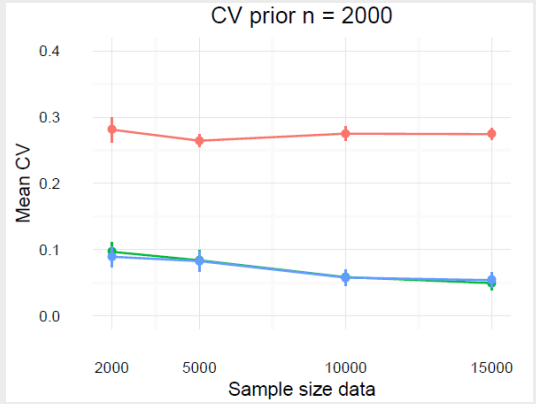


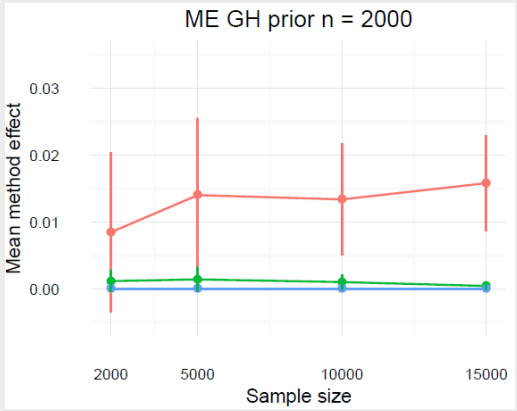
# Case study – Combining all indicators

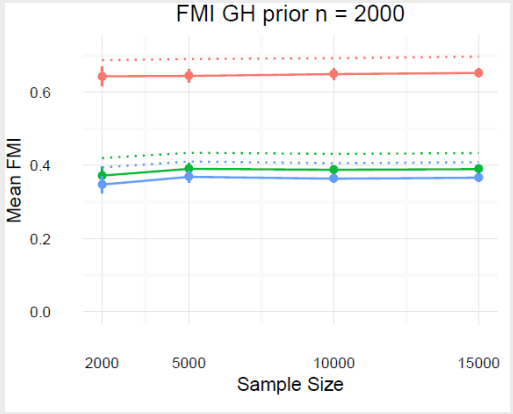
## How to employ the two types of quality indicators?

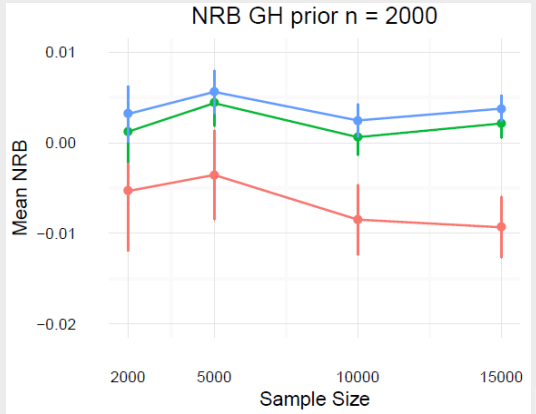


# Conclusions

- Within existing framework it is fairly straightforward to embed models for survey variables;
- Doing so, a range of new quality indicators is available which have slightly more uncertainty since they are based on respondent data;
- It is insightfull to model survey variables jointly;
- Quality indicators based on survey variables require assumptions and/or a benchmark;

# Discussion

- Should we assume dependence between regression parameters in survey variable models over different actions?
- How to deal with the implicit assumptions in the survey variable models?
- What about prior elicitation for survey variable models?
- How to combine the various indicators in ASD?