Using Measurement Models to locate the Sources of Mode Bias

Thomas Klausch *(Utrecht University)*
Joop Hox *(Utrecht University)*
Barry Schouten *(Statistics Netherlands)*

Statistics Netherlands

Presented at the ITSEW 2012
September 2-4, Santpoort, The Netherlands

Background of this study

- Mixed-mode surveys & designs
- Equality of measurements needs to be assured
- Use multiple group confirmatory factor analysis (MCFA) to determine the type of measurement effects modes can have:
  - Change scale of a given item sensitive to mode
  - Change random error of a given item sensitive to mode
  - Introduce differential systematic bias and variance across sets of items
- Deal with: Selection error and ordinal answer scales

Data Collection Design

- National probability sample of persons (The Netherlands)
  - Gross sample 8800 persons
  - Net sample about 4048 persons
- Random assignment to one of four modes
  - Capi (Response Rate: 64%), Cati (67%/45%), Mail (49%), Web (29%)
- Analysis of 3 scales:
  - Police visibility
  - Neighbourhood traffic pressure
    - Both Explored and cross-validated on a different data set (Safety Monitor 2010)
  - Duty to obey the Police
    - Pretested in the European Social Survey (ESS round 5)

Overview on the three scales

Neighborhood Traffic Pressure (NTP), early position

- Aggressive behavior in traffic
- Traffic noise nuisance
- Speeding in traffic
- Parking problems

3 Answer categories
Explicit DK in Web/Mail

Police Visibility (PV), middle position

- The police offer protection to people in this neighborhood.
- The police have contact with people from this neighborhood.
- The police react to problems in this neighborhood.
- The police do their best in this neighborhood.

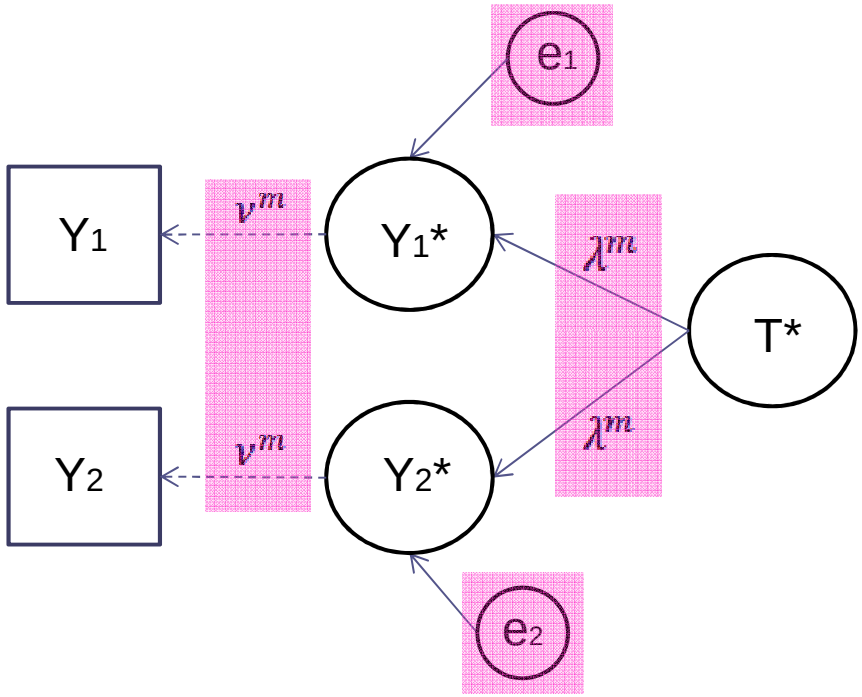
5 Answer categories
Explicit DK in Web/Mail

Duty to obey the police (DTO), late position

- Support the decisions of the police, also if I disagree.
- Do what the police say, also if I disagree.
- Do what the police say, also if I am treated unpleasantly.

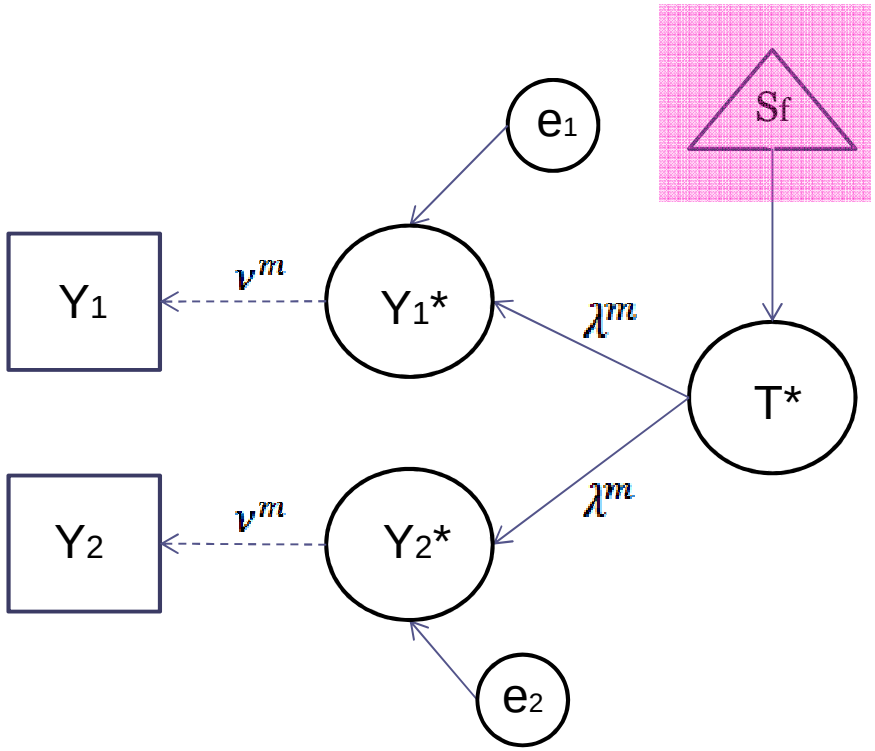
5 Answer categories
No DK in Web/Mail

Possible sources of item bias (MCFA)



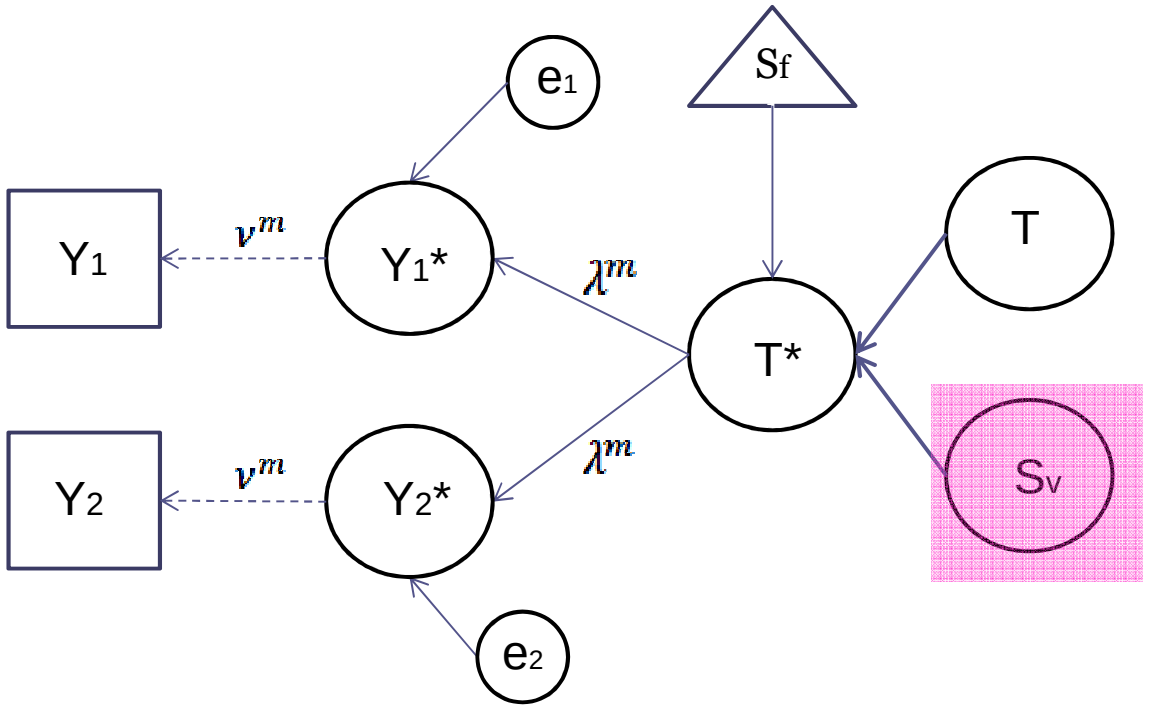
Path diagram for an ordinal CFA
(simplified illustration, not identified)

Systematic Errors in MG-CFA models



Path diagram for an ordinal CFA with a mean structure

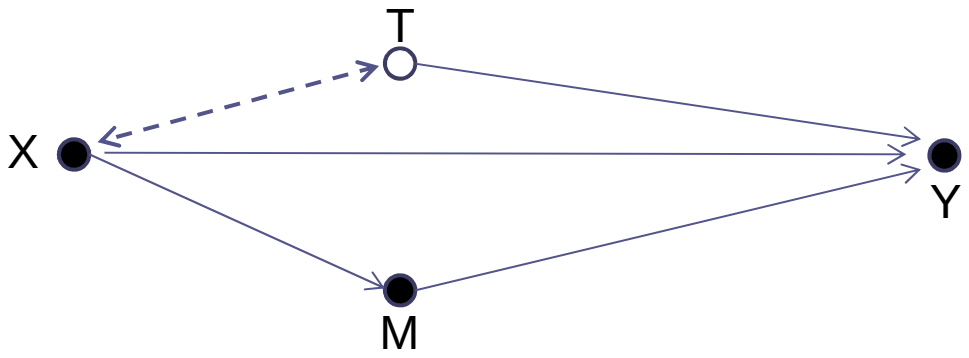
Systematic Errors in MG-CFA models



Path diagram for an ordinal CFA with systematic errors on all items
(e.g. Alwin, 2007)

Counterfactuals in mode experiments

- Sample compositions obtained by different survey modes are never homogenous
- Threat to causal inference, if measurement differs across selection variables X
- Inverse Propensity Score Weighting applied (adjusted for 8 socio-demographics)



Expectation

- Self-administered modes (web, mail) have very similar psychological properties in the answering process
  - Visual stimulus and answering
  - Anonymous situation, absence of interviewer
  - Earlier studies: no measurement effects in CFA models
- Interviewer administered modes (F2F, Telephone) also similar
  - Audible information exchange, cognitive processing without visual support
  - Social situation
- Expectation: Major differences between interviewer and self-administered modes

Summary of Results

- On all items of the PV and NTP scales there was a threshold bias on at least one of the thresholds, but not on the DTO scale
  - Difference was only present between interviewer and self-administered modes
  - Surprise: item-specific bias found regardless of item content in these scales
- On all scales there was additionally a systematic bias (factor mean difference)
- On all items of all scales, there was a difference in random errors
  - Interviewer modes produces more random error
- But no systematic variance difference (except Web DTO scale)

Illustration of threshold bias for PV scale:

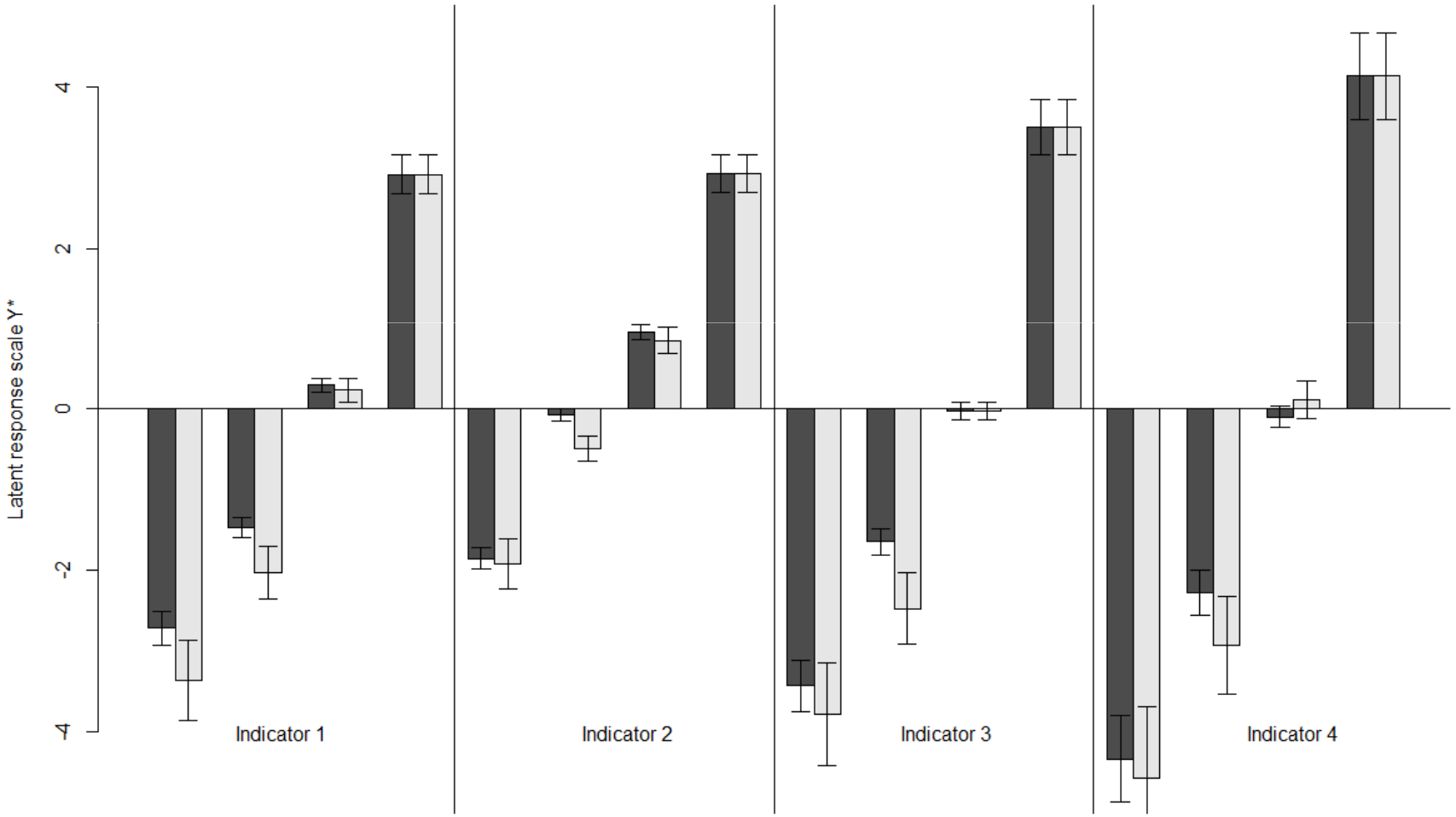


Illustration of reliability difference

	NTP Scale		PV Scale		DTO Scale		
	F2F/Tel	Web/Paper	F2F/Tel	Web/Paper	F2F/Tel	Paper	Web
Indicator 1	0.590	0.678	0.545	0.645	0.317	0.396	0.490
	(.029)	(.035)	(.019)	(.018)	(.016)	(.019)	(.028)
Indicator 2	0.444	0.472	0.432	0.532	0.767	0.870	0.908
	(.029)	(.028)	(.019)	(.022)	(.023)	(.021)	(.017)
Indicator 3	0.577	0.764	0.660	0.663	0.638	0.724	0.794
	(.030)	(.034)	(.018)	(.021)	(.020)	(.019)	(.020)
Indicator 4	0.101	0.118	0.771	0.835	-	-	-
	(.015)	(.017)	(.018)	(.019)			

Conclusion

- Modes cause systematic differences in measurement across sets of attitudinal items between self- and interviewer adm. modes
  - Item-specific variations in strength of threshold bias
  - Systematic bias across all items
- Direction of systematic bias suggests social desirability
  - However: other answering behaviours might cause this bias
- The same observed answer in interviewer and self-administered modes does not reflect the same underlying opinion
- Self-administered modes: more efficient (lower random error)

Conclusion

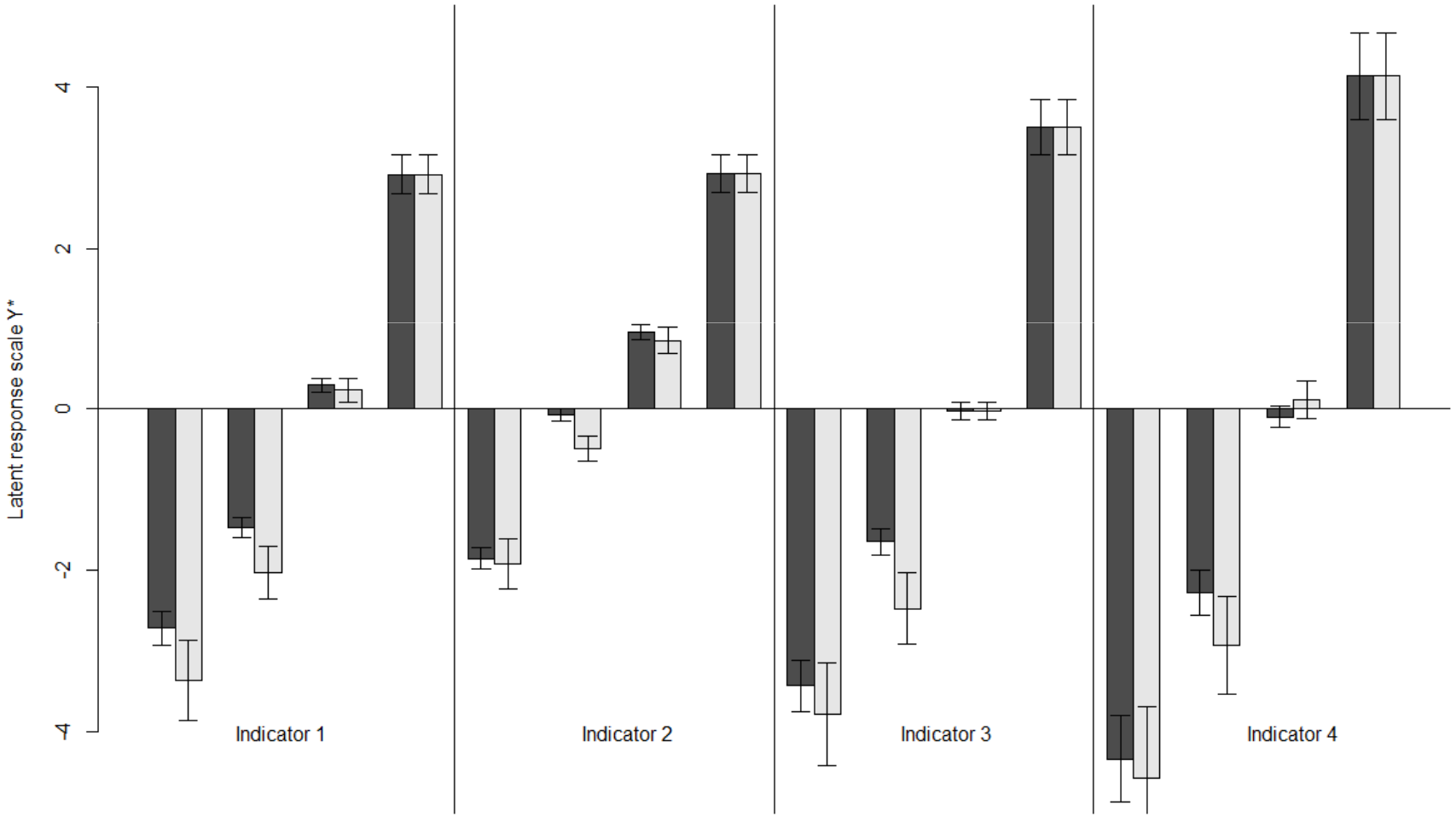
- The worse mixed-mode options:
  - Any combination of interviewer and self-adm. modes
- The good options:
  - Web-Mail or F2F-Telephone only
- Our results might be scale-dependent
  - Reproduction on more scales / items
  - Assess equivalence for your items during MM design
- Conclusions apply to surveys that focus on attitudinal constructs
  - Factual variables might behave differently
  - Talk J. Van der Laan: Employment statistics no strong ME

Backup

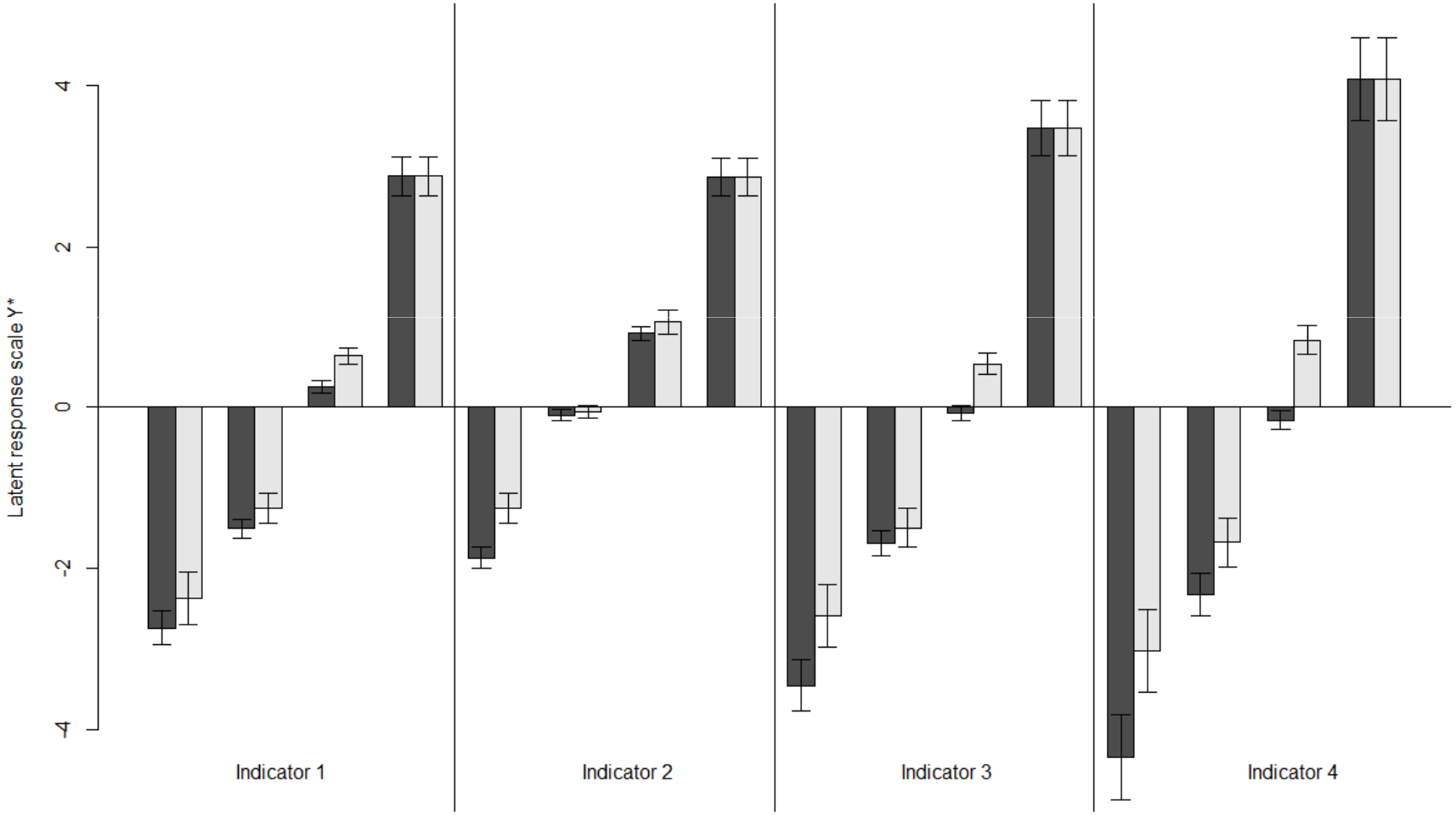
What did adjustment weighting change?

- All bias was a bit reduced
- Model Fit increased (about -0.02 change in RMSEA)
- Important: Systematic variance difference was present before adjustment (i..e. difference in factor variance)
  - Again between interviewer and self-administered modes
  - Effect of adjustment? Perhaps. Could also be an increase in noise.
- All selection effects adjusted? Maybe – effects found conform to theoretical expectations!

Illustration of threshold bias for PV scale:



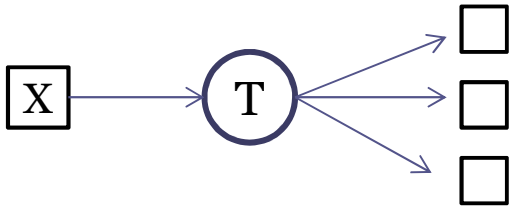
The effect of a systematic bias:



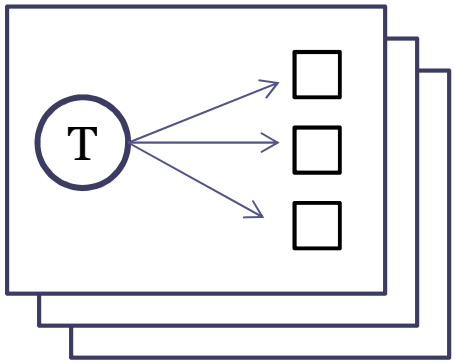
Conditioning in CFA models

Options to condition on X include (e.g. Morgan & Winship, 2007)

1. Ancova type adjusments



2. Stratifying all estimation on X
   - Sparseness problems
   - Tedious



3. Propensity score methods, e.g. weighting
   - Own simulation: inverse propensity score weighting works best

IPW estimation

- Probit model with socio-demographics and interactions with all mode indicators
  - Gender
  - Age
  - Income
  - Nationality
  - Civil Status
  - Household Size
  - Urbanity
  - Living in one of the 3 big Dutch cities

- Available from national registries on sample level