Bayesian Factor Mixture Modeling with Response Time for Detecting Careless Respondents

Lijin Zhang[1], Esther Ulitzsch[2,3], Ben Domingue[1]

[1] Graduate School of Education, Stanford University
[2] Centre for Educational Measurement, University of Oslo
[3] Centre of Research on Equality in Education, University of Oslo

Introduction

- Scales have been extensively used to investigate latent variables in social science research.
- The effectiveness of survey data depends on the assumption that responses accurately represent the latent constructs.

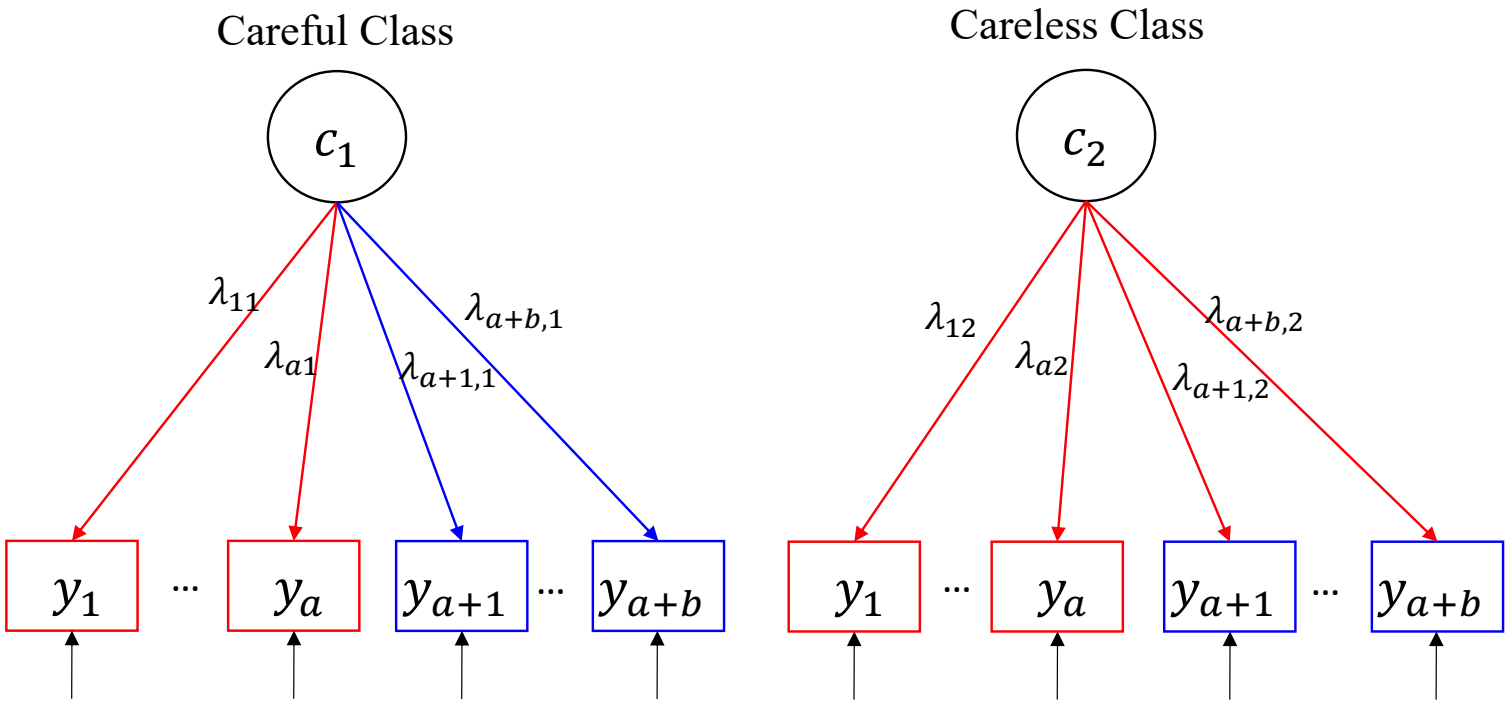
Careless Respondents

- Those whose answers are not the result of careful thought but rather lack of attention, misunderstanding, or lack of interest (Arias et al., 2020).
- Research into scales has revealed a widespread occurrence of careless behaviors, with reported rates varying from 3% to 50% (Meade & Craig, 2012).

Influences of Careless Respondents

- Careless responses are problematic because they are relatively unrelated to the relevant constructs.
- Biased findings. For instance, reduced correlation between variables (Kam & Meyer, 2015).
- Poor model fitting in CFA (Voss, 2023; Woods, 2006).

Factor Mixture Modeling

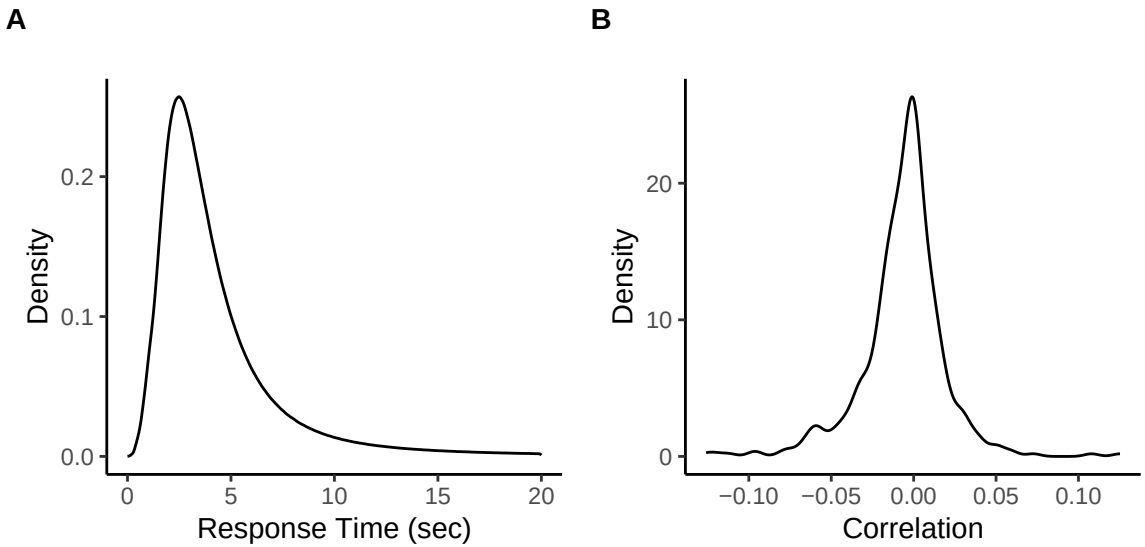


Factor Mixture Model based on Item Wording

Note: Red/Blue blocks: Positively/Negatively worded items;
Red/Blue lines: Positive/Negative loadings.

Response Time

- The duration required to respond to a question on a psychological scale tends to be short.
- Correlation between Time and Responses is around 0.



Time of 53,671,504 survey responses from IRW (Domingue et al., 2024)

Goals

- Despite the prevalent usage of CFA for scale data analysis, the integration of response time in CFA remains underexplored.
- Lack of comparative studies on models with and without response time

We seeks to fill this gap by incorporating response time metrics within the CFA framework to detect careless respondents.

The Proposed Model

For item $j \in \{1, ..., J\}$ and respondent $i \in \{1, ..., N\}$, the model is formulated as follows:

Careful Group:

$$\begin{aligned} \log\left(t_{ij}^{a}\right) &= \beta_j - \tau_i + \gamma_{ij} \\ y_{ij}^{a} &= \mu_j + \lambda_j * \omega_i + \epsilon_{ij} \end{aligned} \tag{1}$$

Careless Group:

$$\begin{aligned} \log\left(t_{ij}^{b}\right) &\sim \mathcal{N}(\mu_t^b, \sigma_t^{b2}) \\ y_{ij}^{b} &\sim \mathcal{N}(\mu_y^b, \sigma_y^{b2}) \end{aligned} \tag{2}$$

p_i: probability that i-th respondent belongs to careful group:

$$\begin{aligned} \log(t_{ij}) &= p_i \cdot \log\left(t_{ij}^{a}\right) + (1 - p_i) \\ y_{ij} &= p_i \cdot y_{ij}^{a} + (1 - p_i) \cdot y_{ij}^{b} \end{aligned} \tag{3}$$

Model Assumptions

- For the careless group, both item responses and time are independent of the item/person characteristics, maintaining homogeneity of intercept (μ_y^b and μ_t^b) and variances.
- Respondents who diligently read the item content and respond to the questions are likely to spend more time than those who exhibit less attention to details ($\beta_j > \mu_t^b$).

Simulation Study: Goals

- Assess the effectiveness of the proposed model;
- Illustrate the advantages of incorporating response time.

Simulation Design

- Proportions of careless respondents (π): 0.05, 0.1, 0.15, 0.2.
- Sample sizes: 300, 500, and 1000.
- Differences in log-RT between the careful and careless groups were controlled.

Responses

- Careful: A single-factor model with ten items ($\lambda_j = 0.8$)
- Careless: $y_{ij}^b \sim N(1, 2)$

Response Time

- Careful: A log-normal model with intercept following $U(0.95, 1.2)$ or $U(1.2, 1.45)$, and residual variance $\psi_{\gamma,j}$ at 1 or 1.25.
- Careless: $\log\left(t_{ij}^b\right) \sim N(0.7, 0.5)$.

Model Estimation

$$
\begin{aligned}
(p_i, 1 - p_i) &\sim Dir(1, 1) \\
\mu_j, \mu_y^b, \beta_j, \mu_t^b &\sim N(0, 100) \\
\psi_{\epsilon,j}, \sigma_y^{b2}, \phi_\tau, \psi_{\gamma,j}, \sigma_t^{b2} &\sim Inv - Gamma(0.01, 0.01) \\
\delta_j, \lambda_j &\sim TN(0, 100, 0,)
\end{aligned}
\tag{4}
$$

- JAGS (Plummer, 2004).
- Burn-in iterations: 10,000 - 100,000.
- Model convergence is assessed using the estimated potential scale reduction (EPSR) index (Gelman, 1996): EPSR < 1.1.
- Upon convergence, we generate additional 20,000 MCMC samples for model estimation.

Evaluation Criteria

- Model Convergence
- Estimation Accuracy
  - Relative Bias; Root Mean Square Error (RMSE)
- Classification Accuracy

		Estimated		
		Careful	Careless	
True	Careful	TP	FN	$\text{FNR} = \frac{FN}{TP+FN}$
	Careless	FP	TN	$\text{FPR} = \frac{FP}{FP+TN}$
		$\text{Precision} = \frac{TP}{TP+FP}$		$\text{Accuracy} = \frac{TP+TN}{N}$

Convergence and Classification Accuracy

The model convergence rates are 100% for all conditions.

Table 3: Classification Accuracy of the Proposed Model.

	N	$\psi_{\gamma,j}$	$\pi = 0.05$	$\pi = 0.1$	$\pi = 0.15$	$\pi = 0.2$
Accuracy	500	1	0.92	0.97	0.98	0.98
Sensitivity			0.92	0.97	0.98	0.98
Precision			1.00	1.00	1.00	1.00
FPR			0.01	0.01	0.01	0.01
FNR			0.08	0.03	0.02	0.02
Accuracy		1.25	0.93	0.98	0.98	0.98
Sensitivity			0.92	0.98	0.98	0.98
Precision			1.00	1.00	1.00	1.00
FPR			0.01	0.01	0.01	0.01
FNR			0.08	0.02	0.02	0.02

Note: N denotes the sample size; $\psi_{\gamma,j}$ indicates the residual variance of response time in the careful group; FPR = False Positive Rate; FNR = False Negative Rate.

Estimation Accuracy

Table 4: Estimation Results of the Proposed Model.

π	0.05				0.1				0.15				0.2			
β_j	U(0.95-1.2)		U(1.2-1.45)		U(0.95-1.2)		U(1.2-1.45)		U(0.95-1.2)		U(1.2-1.45)		U(0.95-1.2)		U(1.2-1.45)	
	RB(%)	RMSE	RB	RMSE	RB	RMSE	RB	RMSE	RB	RMSE	RB	RMSE	RB	RMSE	RB	RMSE
β_j	3.59	0.07	3.89	0.08	2.54	0.06	2.55	0.07	1.29	0.07	1.51	0.07	1.18	0.06	1.37	0.06
λ_j	5.78	0.06	5.17	0.06	2.21	0.05	1.98	0.05	1.92	0.05	1.83	0.05	2.05	0.05	1.99	0.05
μ_{j1}	**-10.51**	0.08	-9.06	0.07	-5.20	0.06	-4.55	0.06	-3.57	0.05	-3.20	0.05	-1.23	0.05	-0.91	0.05
ϕ_τ	**10.72**	0.07	5.78	0.05	3.64	0.05	1.15	0.05	3.97	0.05	2.11	0.05	1.62	0.05	0.04	0.04
$\psi_{\epsilon,j}$	-4.54	0.03	-3.95	0.03	-2.26	0.03	-1.78	0.03	-1.71	0.03	-1.43	0.03	-1.19	0.03	-0.94	0.03
$\psi_{\gamma,j}$	1.84	0.08	1.72	0.08	2.14	0.08	1.95	0.08	1.72	0.08	1.59	0.08	1.51	0.08	1.40	0.08
μ_y^b	-9.39	0.14	-8.87	0.14	-1.18	0.06	-0.82	0.06	-1.22	0.06	-1.12	0.06	-0.77	0.05	-0.78	0.05
σ_y^{b2}	**-47.43**	0.96	**-44.42**	0.91	**-18.21**	0.40	**-16.14**	0.36	-9.65	0.23	-8.72	0.21	-6.05	0.16	-5.46	0.15
μ_t^b	8.96	0.10	**14.73**	0.14	0.97	0.04	1.84	0.04	0.38	0.03	0.83	0.03	0.57	0.03	0.67	0.03
σ_t^{b2}	**61.49**	0.32	**57.79**	0.30	**14.02**	0.08	**12.11**	0.08	6.98	0.05	6.52	0.05	4.04	0.03	3.74	0.03

Benefits of Modeling Response Time

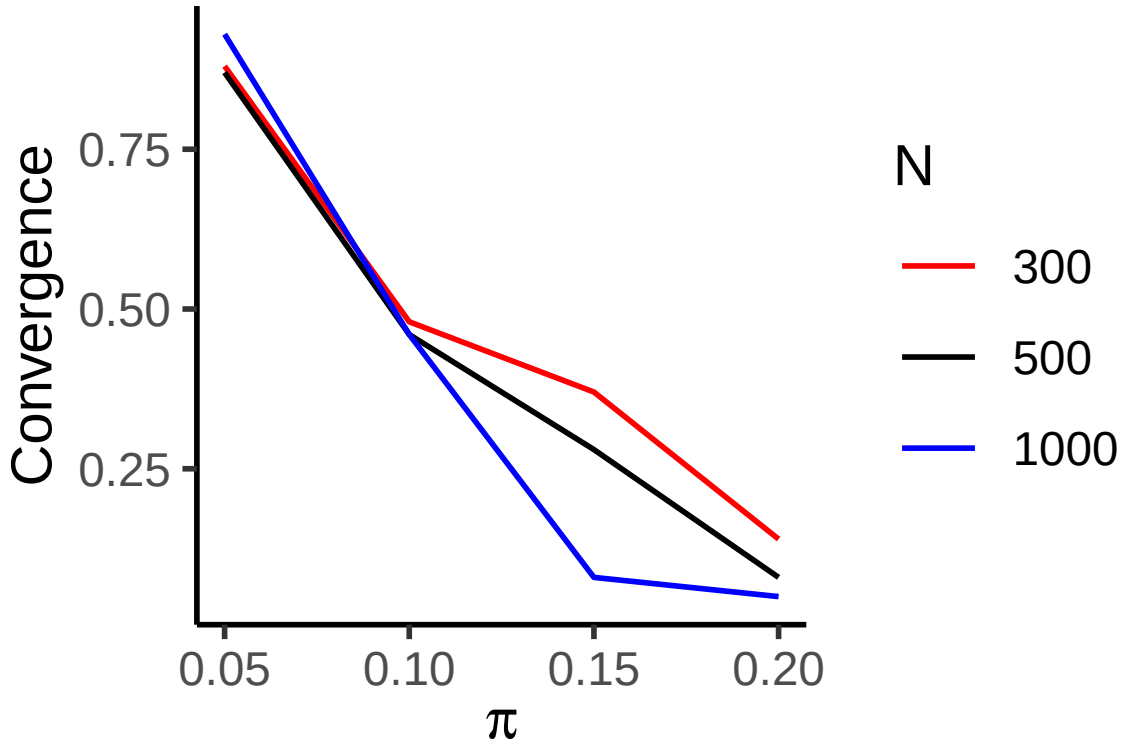
We compared the proposed model with the model without response time:

$$
\begin{aligned}
y_{ij}^{a} &= \mu_j + \lambda_j * \omega_i + \epsilon_{ij} \\
y_{ij}^{b} &\sim N(\mu_y^b, \sigma_y^{b2}) \\
y_{ij} &= p_i \cdot y_{ij}^{a} + (1 - p_i) \cdot y_{ij}^{b}
\end{aligned}
\tag{5}
$$

Theoretically, modeling response time helps the interpretation and justification of excluding those careless individuals.

Model without Response Time

A significant influence observed upon removing response time is the difficulty in achieving model convergence.



Model without Response Time

Table: Classification Accuracy of the Model without Response Time

π	0.05		0.1	
β	U(0.95-1.2)	U(1.2-1.45)	U(0.95-1.2)	U(1.2-1.45)
Accuracy	0.84	0.84	0.93	0.93
Sensitivity	0.84	0.84	0.92	0.92
Precision	1.00	1.00	1.00	1.00
FPR	0.02	0.02	0.02	0.02
FNR	0.16	0.16	0.08	0.08

Model without Response Time

Table: Estimation Results of the Model without Response Time.

π	0.05				0.1			
$\boldsymbol{\beta}$	U(0.95-1.2)		U(1.2-1.45)		U(0.95-1.2)		U(1.2-1.45)	
	RB(%)	RMSE	RB	RMSE	RB	RMSE	RB	RMSE
λ_j	**10.20**	0.09	**10.20**	0.09	5.71	0.07	5.71	0.07
μ_{j1}	**-18.49**	0.11	**-18.49**	0.11	**-10.37**	0.08	**-10.37**	0.08
$\psi_{\epsilon,j}$	-7.97	0.04	-7.97	0.04	-6.24	0.04	-6.24	0.04
μ_y^b	**-13.48**	0.18	**-13.48**	0.18	-5.88	0.09	-5.88	0.09
σ_y^{b2}	**-57.13**	1.15	**-57.13**	1.15	**-33.52**	0.68	**-33.52**	0.68

Empirical Study

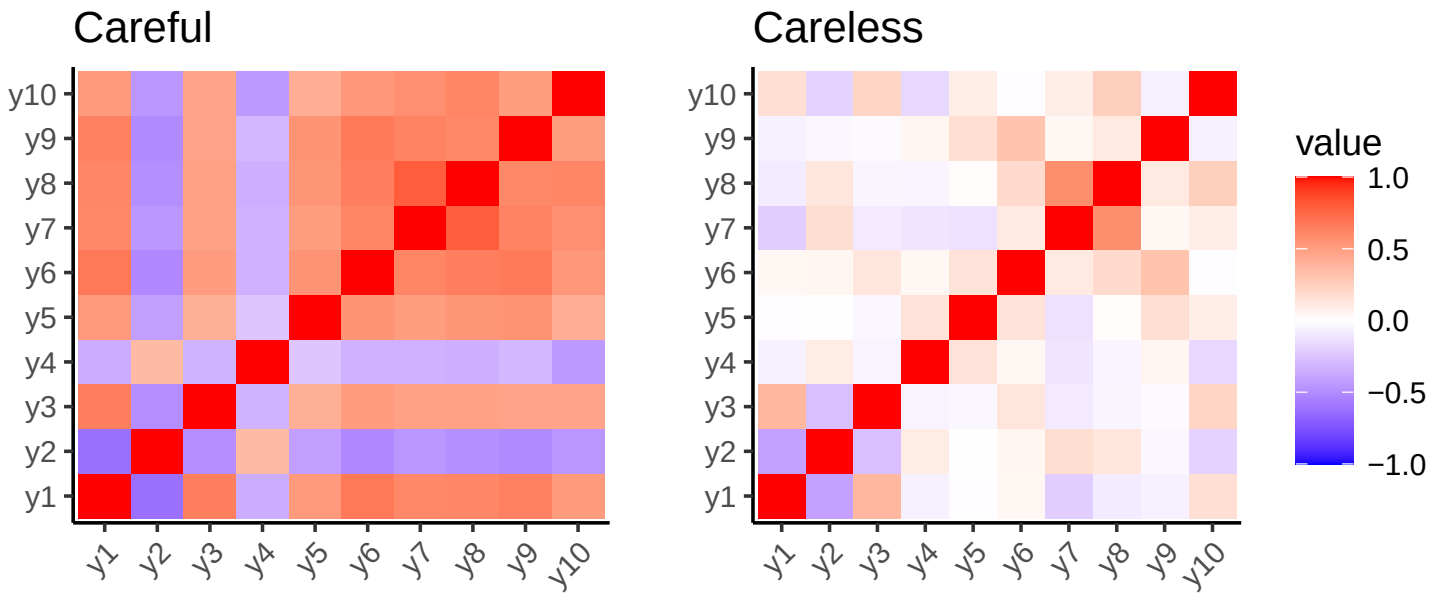
- Emotional Stability Scale
- $N = 1000$
- Number of Items: 10 (8 negatively worded items)
- Five-point Likert Scale

The single-factor model did not adequately fit the data (BCFI = 0.921, BTLI = 0.882, BNFI = 0.918).

Results

The proposed model identified 261 respondents as careless.

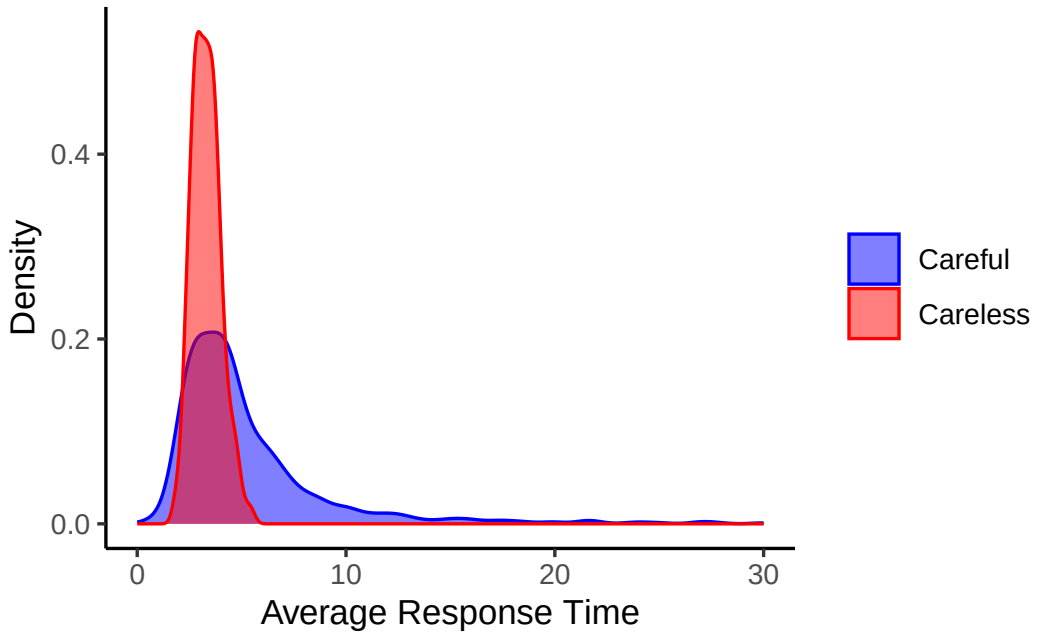
Upon removal of these careless respondents, the fit of the CFA model to the data improved significantly (BCFI = 0.955, BTLI = 0.930, BNFI = 0.952).



Results

Respondents classified as careless took less time per item (3.31 seconds vs 6.46 seconds).

Time taken to respond by the careless group showed minimal fluctuation across ten items, with an average standard deviation (SD) of 1.51, compared to 7.96 in the careful group.



Summary

- We integrate response time into factor analysis to detect careless respondents by factor mixture modeling.
- It is independent of item wording, applicable to scales that do not include reverse-worded items.
- Simulation and empirical studies demonstrated its effectiveness and the benefits of modeling response time.
- Extension: Careless Respondents Detection for Visual Analogue Scale Data (Zhang, Domingue, Vogelsmeier, Ulitzsch, 2024).

Preprint

https://osf.io/preprints/psyarxiv/qc9jb

Thank you!