Erasmus
School of
Economics

Identifying periods of careless responding in rating-scale surveys

Max Welz   Andreas Alfons

ICORS, July 6, 2022

Erasmus University Rotterdam

Motivation

- Rating-scale surveys are ubiquitous in empirical research

- Not all respondents may respond accurately and truthfully

  $\longrightarrow$ insufficient effort responding (IER; also called careless responding)

Careless responding

Definition: Careless Responding (Huang et al., 2012)

A response set in which the respondent answers a survey measure with low or little motivation to comply with survey instructions, correctly interpret item content, and provide accurate responses.

Identifying careless responding is important:

$\longrightarrow$ Threat to internal validity (e.g. Huang et al., 2015)

$\longrightarrow$ Already $\leq 10\%$ contamination is problematic (Arias et al., 2020)

Carelessness example: Random responding

Random responding: Tendency to randomly choose answer categories, regardless of item content

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
ICORS 2022 is awesome.	□	□	□	■	□
I like chocolate.	□	□	□	□	■
Oxygen is important.	□	□	■	□	□
I am paid biweekly by leprechauns.	□	□	□	■	□
Not getting bitten by a shark would be fun.	■	□	□	□	□

Careless responding as statistical outlier

- Careless respondents can be considered shape outliers example
- Existing outlier detection methods from robust statistics not suitable due to discrete nature of rating-scale data
- In long surveys, a large proportion of respondents may become careless towards the end due to fatigue (Bowling et al., 2021a)

$\longrightarrow$ Partial rowwise outlier

$\longrightarrow$ Goal: Identify *when* a given respondent responds carelessly (if at all)!

Proposed method (Part I)

Search for evidence of carelessness in two dimensions:

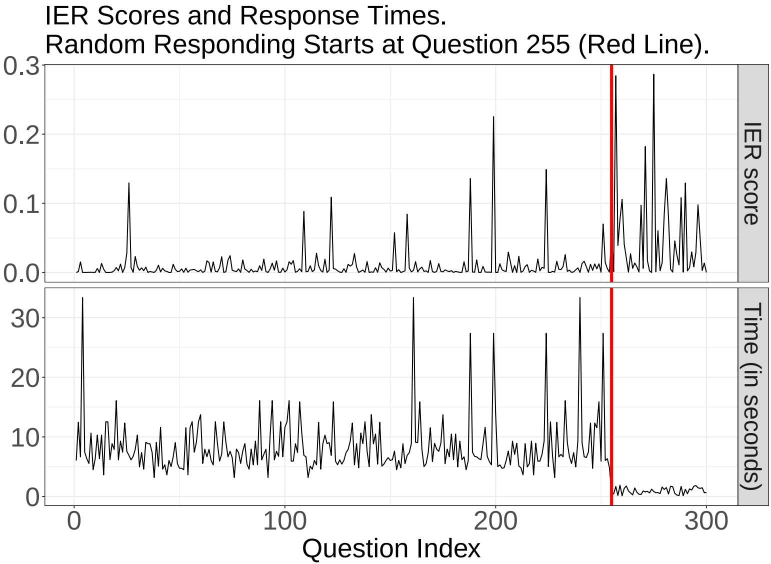
1. IER scores from autoencoder reconstruction (Kramer, 1992):

$$\mathrm{IER}_{ij} = \left(\frac{X_{ij} - \widehat{X}_{ij}}{L_j} \right)^2,$$

   for reconstruction $\widehat{X}_{ij}$ of response X_{ij} of i-th respondent to j-th item, for which there are L_j answer categories

2. Per-item response time (cf. Bowling et al., 2021b)

Illustration of change points



Proposed method (Part II)

Use Self-Normalization (Zhao et al., 2021) to detect change points in the two-dimensional series details

- Loops recursively over many nested segmentations of a series
- For each segmentation, calculate some test statistic
- If test statistic exceeds some threshold, there is a change point

$\longrightarrow$ Recursively narrow segment to isolate location of a change point

$\longrightarrow$ Asymptotic size control through the choice of the threshold

Simulation experiment

Generate 100 datasets details

- Sample 500 rating-scale observations for 15 constructs with 20 items each (random order)

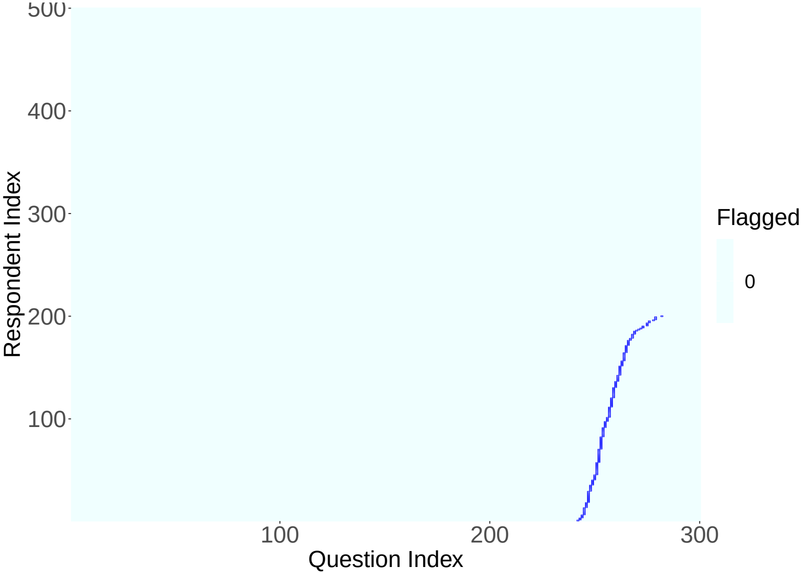
  $\longrightarrow$ Inspired by NEO PI-R (Costa and McCrae, 1992)

- Sample response times based on empirical times in Schroeders et al. (2022)

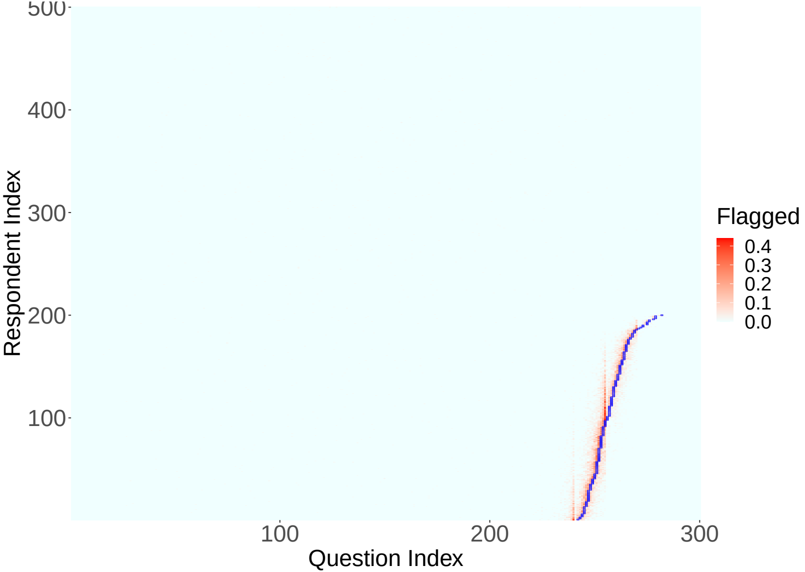
- Simulate carelessness onset from a Weibull distribution following Bowling et al. (2021a)

- Contaminate with various types of careless response styles

Simulated Carelessness onset for 40% contamination



Simulation results at level 2.5%



Simulation results details

Using the Adjusted Rand Index (ARI; Hubert and Arabie, 1985) as performance measure, we find that our method

- accurately detects the change points for all response styles (ARI ≥ 0.85)
- controls Type I error at levels 0.01, 0.025, 0.05

$\longrightarrow$ Results are fairly consistent across contamination levels (up to 100%)

Conclusions and outlook

⟶ Proposed methodology seems promising for detecting careless periods

⟶ Next steps:

- Tweaks in methodology: e.g., add third dimension based on longstring index to change point detection
- Lots of simulations
- Design a survey and collect emprical data

Conclusions and outlook

⟶ Robust methods for discrete data are underdeveloped

⟶ Big interest in the social sciences for robustness

⟶ Potential for novel research ideas

My goal: develop theory and methods for robustness in discrete data

⟶ Measure transportation (e.g. Peyré and Cuturi, 2019) seems interesting for this purpose

⟶ I'd love to chat about with you about discrete robustness!

References

Arias, V. B., Garrido, L., Jenaro, C., Martínez-Molina, A., and Arias, B. (2020). A little garbage in, lots of garbage out: Assessing the impact of careless responding in personality survey data. Behavior Research Methods, 52(6):2489–2505.

Bowling, N. A., Gibson, A. M., Houpt, J. W., and Brower, C. K. (2021a). Will the questions ever end? person-level increases in careless responding during questionnaire completion. Organizational Research Methods, 24(4):718–738.

Bowling, N. A., Huang, J. L., Brower, C. K., and Bragg, C. B. (2021b). The quick and the careless: The construct validity of page time as a measure of insufficient effort responding to surveys. Organizational Research Methods. In press.

Costa, P. T. and McCrae, R. R. (1992). Revised NEO personality inventory (NEO-PI-R) and NEO five-factor (NEO-FFI) inventory: prof Psychological Assessment Resources, Odessa, FL.

Goldfeld, K. and Wujciak-Jens, J. (2020). `simstudy`: Illuminating research methods through data generation. Journal of Open Source Software, 5(54):2763.

Huang, J. L., Curran, P. G., Keeney, J., Poposki, E. M., and DeShon, R. P. (2012). Detecting and deterring insufficient effort responding to surveys. Journal of Business and Psychology, 27(1):99–114.

References (cont.)

Huang, J. L., Liu, M., and Bowling, N. A. (2015). Insufficient effort responding: examining an insidious confound in survey data. Journal of Applied Psychology, 100(3):828–845.

Hubert, L. and Arabie, P. (1985). Comparing partitions. Journal of Classification, 2(1):193–218.

Kramer, M. A. (1992). Autoassociative neural networks. Computers & Chemical Engineering, 16(4):313–328.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport. Foundations and Trends in Machine Learning, 11(5-6):355–607.

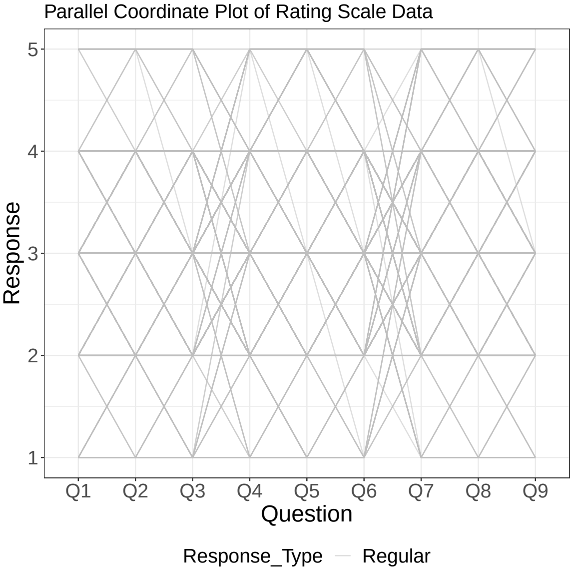
Scardapane, S., Comminiello, D., Hussain, A., and Uncini, A. (2017). Group sparse regularization for deep neural networks. Neurocomputing, 241:81–89.

Schroeders, U., Schmidt, C., and Gnambs, T. (2022). Detecting careless responding in survey data using stochastic gradient boosting. Educational and Psychological Measurement, 82(1):29–56.

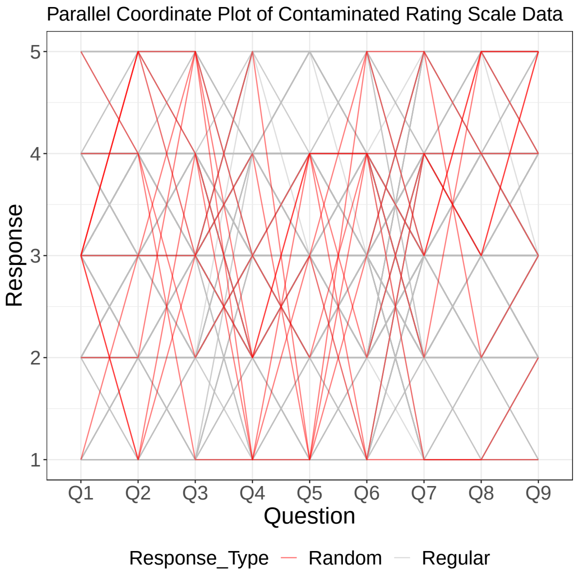
Zhao, Z., Jiang, F., and Shao, X. (2021). Segmenting time series via self-normalization. arXiv preprint arXiv:2112.05331.

Appendix

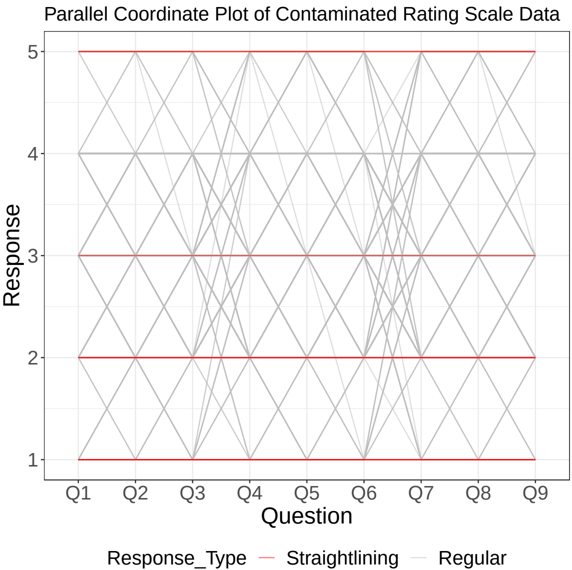
Example: Rating-scale data



Example: Random responding



Example: Straightlining



Setup

- Ordinal survey responses of n respondents to p items are collected in an $n \times p$ data matrix $\boldsymbol{X}$, with possibly $n < p$
- The observations (respondents) are assumed to be i.i.d.
- We don't know when and where (if at all) IER occurs in $\boldsymbol{X}$

Assumptions and main idea

Assumptions 1 and 2

1. The responses in $\boldsymbol{X}$ admit a low-dimensional representation, $\boldsymbol{S}$, of dimension $q \ll p$. The value of q is known.
2. The survey that generated $\boldsymbol{X}$ is reliable in the sense that if there was no IER, $\boldsymbol{X}$ would accurately measure all constructs.

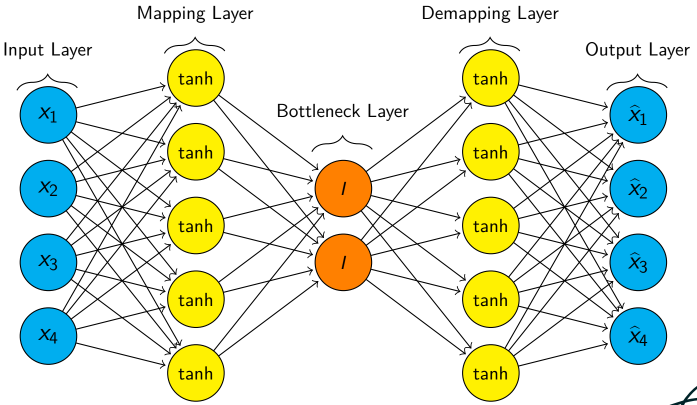
$\longrightarrow$ In the behavioral sciences, the number of constructs a survey measures is typically known

Idea: Reconstruct $\boldsymbol{X}$ from its low-dimensional representation $\boldsymbol{S}$. Poor or near-perfect reconstruction might indicate IER (more on this later).

Auto-associative neural network (autoencoder)

Autoencoder (Kramer, 1992) to estimate $\boldsymbol{S}$ from $\boldsymbol{X}$:

- Neural network with odd depth that attempts to reconstruct its input
- Central hidden layer is crucial: compresses the input to dimension q



Auto-associative neural network (autoencoder)

Network architecture:

- q nodes in bottleneck layer
- $1.5 \times p$ nodes in (de)mapping layers
- Activation functions:
  $\longrightarrow$ hyperbolic tangent in (de)mapping layers
  $\longrightarrow$ linear in bottleneck layer
- Run 100 epochs
- Use robust pseudo-Huber loss

Autoencoder: Regularization

We incorporate information on page-membership of items:

$\longrightarrow$ IER behavior can be expected to be similar within each page, but may differ between pages

$\longrightarrow$ Use group-lasso regularization (cf. Scardapane et al., 2017)

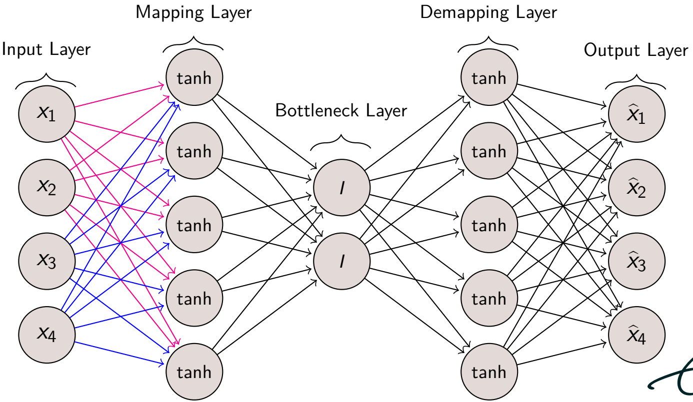
- Each group G_j holds items on the same survey page
- Regularization is applied between input layer and mapping layer

$\longrightarrow$ For instance, items from later pages (where IER can be expected to be higher) may be routed through different nodes in the mapping layer than items from earlier pages

Autoencoder: Regularization

For m pages and a parameter vector $\boldsymbol{\theta}$, the group-lasso penalty is

$$\Omega(\boldsymbol{\theta}) = \sum_{j=1}^{m} \sqrt{|G_j| \sum_{k \in G_j} \theta_k^2}$$



Autoencoder: Reconstruction error

Let $\widehat{X}_{ij}$ be the reconstruction of response $X_{ij}, i = 1, \ldots, n; j = 1, \ldots, p$

$\longrightarrow$ The associated IER score is the reconstruction error

$$\text{IER}_{ij} = \left(\frac{X_{ij} - \widehat{X}_{ij}}{L_j} \right)^2$$

with L_j denoting the number of answer categories of item j

Response times

Response times can be a good indicator of IER (Bowling et al., 2021b)...

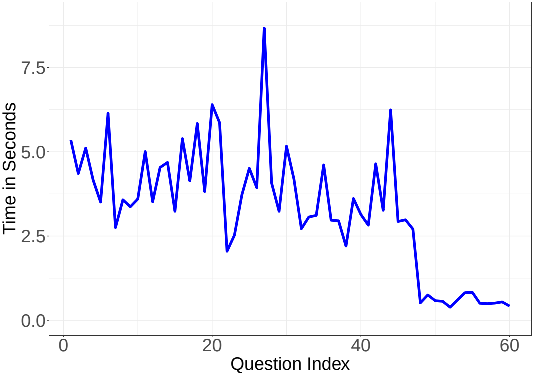


Figure: Response times of a respondent who was instructed to respond accurately and truthfully to each item in a study of Schroeders et al. (2022)

Response times

. . . but sometimes they are not!

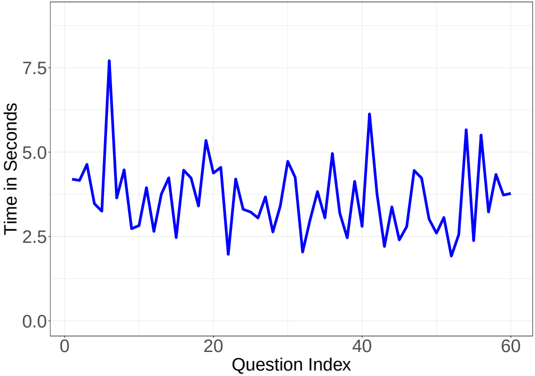


Figure: Response times of a respondent who was instructed to perform IER throughout a study of Schroeders et al. (2022)

Change point detection

Assumption 3

In periods in which a respondent engages in IER, there is a change point in either the IER scores or the response times, or both.

$\longrightarrow$ The IER scores and response times support or complement each other in identifying IER periods

$\longrightarrow$ Combine both quantities in a two-dimensional item series of length p and apply a multivariate method for change point detection

Self-normalization test for change points

Estimation of multiple change points based on self-normalization (SNCP; Zhao et al., 2021)

- Let $\{\boldsymbol{Y}_t\}_{t=1}^{p}$ be a piecewise stationary series of dimension d
- Series has $m_o \geq 0$ (unknown) change points that partition it into $m_o + 1$ segments
- Each segment is piecewise stationary and obeys CDFs $F^{(i)}, i = 1, \ldots, m_o + 1$
- CDFs are characterized by functionals $\boldsymbol{\theta}_i = \boldsymbol{\theta}\left(F^{(i)}\right) \in \mathbb{R}^d$

$\longrightarrow$ Breaks in $\{\boldsymbol{Y}_t\}_{t=1}^{p}$ are characterized by breaks in $\boldsymbol{\theta}$

$\longrightarrow$ Goal: Estimate number of change points, m_o, and their locations

Self-normalization test for change points

For $1 \leq a < b \leq p$, put $\widehat{\boldsymbol{\theta}}_{a,b} = \boldsymbol{\theta}\left(\widehat{F}_{a,b}\right)$, where $\widehat{F}_{a,b}$ is the empirical CDF of $\{\boldsymbol{Y}_t\}_{t=a}^{b}$. For a vector $\boldsymbol{v}$, denote its outer product by $\boldsymbol{v}^{\otimes 2} = \boldsymbol{v}\boldsymbol{v}^\top$.

For $k \in \mathbb{N}$ such that $1 \leq t_1 < k < t_2 \leq p$:

$$
\begin{aligned}
T_p(t_1, k, t_2) &= \boldsymbol{D}_p(t_1, k, t_2)^\top \boldsymbol{V}_p(t_1, k, t_2)^{-1} \boldsymbol{D}_p(t_1, k, t_2) \\
\boldsymbol{D}_p(t_1, k, t_2) &= \frac{(k - t_1 + 1)(t_2 - k)}{(t_2 - t_1 + 1)^{3/2}} \left(\widehat{\boldsymbol{\theta}}_{t_1,k} - \widehat{\boldsymbol{\theta}}_{k+1,t_2}\right) \\
\boldsymbol{V}_p(t_1, k, t_2) &= \boldsymbol{L}_p(t_1, k, t_2) + \boldsymbol{R}_p(t_1, k, t_2) \\
\boldsymbol{L}_p(t_1, k, t_2) &= \sum_{i=t_1}^{k} \frac{(i - t_1 + 1)^2 (k - i)^2}{(t_2 - t_1 + 1)^2 (k - t_1 + 1)^2} \left(\widehat{\boldsymbol{\theta}}_{t_1,i} - \widehat{\boldsymbol{\theta}}_{i+1,k}\right)^{\otimes 2} \\
\boldsymbol{R}_p(t_1, k, t_2) &= \sum_{i=k+1}^{t_2} \frac{(t_2 - i + 1)^2 (i - 1 - k)^2}{(t_2 - t_1 + 1)^2 (t_2 - k)^2} \left(\widehat{\boldsymbol{\theta}}_{i,t_2} - \widehat{\boldsymbol{\theta}}_{k+1,i-1}\right)^{\otimes 2}
\end{aligned}
$$

Self-normalization test for change points back

We calculate the test statistic $T_p(t_1, k, t_2)$ based on a collection of nested windows covering k. Specifically, for $\epsilon \in (0, 0.5)$, define window size $h = \lfloor \epsilon p \rfloor$. For each $k = h, h+1, \ldots, p-h$, we define its corresponding nested window set $H_{1:p}(k)$ defined by

$$H_{1:p}(k) = \left\{ (t_1, t_2) \middle| \begin{array}{l} t_1 = k - j_1 h + 1,\ j_1 = 1, \ldots, \lfloor k/h \rfloor; \\ t_2 = k + j_2 h,\ j_2 = 1, \ldots, \lfloor (p-k)/h \rfloor \end{array} \right\}$$

Self-normalization test for change points

For $k \in \{1, \ldots, p\}$ and $1 \leq s < e \leq p$, denote

$$W_{s,e} = \{(t_1, t_2) | s \leq t_1 < t_2 \leq e\}$$

and

$$H_{s:e}(k) = H_{1:p}(k) \cap W_{s,e}$$

and the subseries maximal test statistic by

$$T_{s,e}(k) = \max_{(t_1,t_2) \in H_{s:e}(k)} T_p(t_1, k, t_2)$$

$\longrightarrow$ Algorithm 1 (next slide) uses these test statistics to estimate the number of change points, $\widehat{m}$, and their locations in $\{1, \ldots, p\}$

Self-normalization test for change points

Algorithm 1: SNCP for multiple change point detection

Input: Time series $\{\boldsymbol{Y}_t\}_{t=1}^{p}$, threshold K_p, window size $h = \lfloor p\epsilon \rfloor$.
Output: Estimated change-points set $\widehat{\boldsymbol{k}} = (\widehat{k}_1, \ldots, \widehat{k}_{\widehat{m}})$
Procedure: SNCP(s, e, K_p, h), for $1 \leq s < e \leq p$
Initialization: SNCP$(1, p, K_p, h)$

if $e - s + 1 < 2h$ **then**
    Stop
else
    $\widehat{k}^* = \arg\max_{k=s,\ldots,e} T_{s,e}(k)$;
    if $T_{s,e}(\widehat{k}^*) \leq K_p$ **then**
        Stop
    else
        $\widehat{\boldsymbol{k}} = \widehat{\boldsymbol{k}} \cup \widehat{k}^*$;
        SNCP$(s, \widehat{k}^*, K_p, h)$;
        SNCP$(\widehat{k}^* + 1, e, K_p, h)$;
    end
end

Self-normalization test for change points

Algorithm 1 ...

- runs in $O(p/\epsilon^2)$ time
- identifies almost surely the correct number of change points and their correct locations, as $p \to \infty$

Under the null hypothesis of no change points, the SNCP test statistic $\max_{k=1,\dots,p} T_{1,p}(k)$ converges in distribution to a limit distribution $G_{\epsilon,d}$

$\longrightarrow$ Offers (asymptotic) size control

$\longrightarrow$ For fixed ϵ, d, and $\alpha \in (0, 0.5)$, choose threshold K_p in Algorithm 1 as the $(1-\alpha)$-th quantile of $G_{\epsilon,d}$

Generating correlated ordinal variables

Sampling scheme of Goldfeld and Wujciak-Jens (2020):

- Goal: Sample independent $X_j \in \{1, \ldots, K\}$, for $j = 1, \ldots, p$, that jointly follow a positive semidefinite covariance matrix $\boldsymbol{\Sigma}$

$\longrightarrow$ Ordinal variables with finite support (K answer categories)

- Specify probabilities of choosing k-th answer category as

$$P_{j,k} = \mathbb{P}[X_j = k],$$

where $\sum_{k=1}^{K} P_{j,k} = 1$, for $j = 1, \ldots, p$

$\longrightarrow$ Implies distribution $X_j \sim F_j$, where, for $x \in \mathbb{R}$,

$$F_j(x) = \sum_{k=1}^{K} \mathbb{1}\{k \leq x\} P_{j,k}$$

Generating correlated ordinal variables

- For F_{logistic} the standard logistic CDF, put, for $k = 1, \ldots, K$,

$$T_{j,k} = F_{\text{logistic}}^{-1}(F_j(k)),$$

  with conventions $F_{\text{logistic}}^{-1}(0) = -\infty$ and $F_{\text{logistic}}^{-1}(1) = +\infty$

- Sample $(Y_1, \ldots, Y_p)^\top \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, then calculate probabilities $U_j = \Phi(Y_j)$ and quantiles $Z_j = F_{\text{logistic}}^{-1}(U_j)$, for $j = 1, \ldots, p$

- Finally, obtain the desired ordinal variables X_j as

$$X_j = 1 + \sum_{k=1}^{K} \mathbb{1}\{Z_j > T_{j,k}\}$$

$\Longrightarrow \quad X_j \in \{1, \ldots, K\}$ and $\text{Var}[X_1, \ldots, X_p] \approx \boldsymbol{\Sigma}$

Setup

Generation of 100 uncontaminated datasets:

- Generate $n = 500$ rating-scale observations
- 15 constructs with 20 items each ($\implies q = 15, p = 300$)
- Constructs are mutually independent and items within the same construct have correlation of ± 0.7
- All items use the following answer probabilities:

$\mathbb{P}[X_j = 1]$	$\mathbb{P}[X_j = 2]$	$\mathbb{P}[X_j = 3]$	$\mathbb{P}[X_j = 4]$	$\mathbb{P}[X_j = 5]$
0.05	0.25	0.40	0.25	0.05

- Items are in random order

$\longrightarrow$ Inspired by the NEO PI-R (Costa and McCrae, 1992)

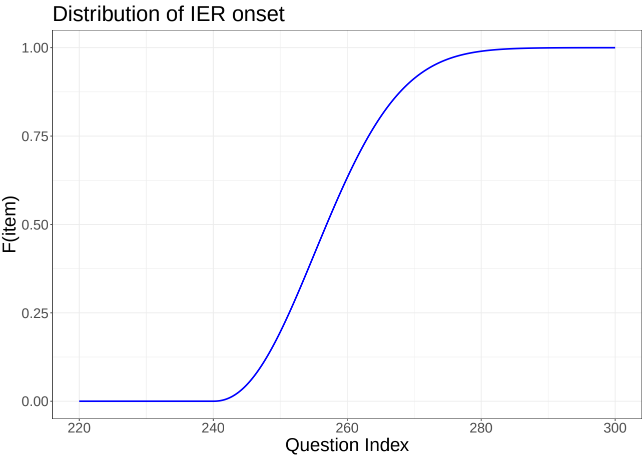
Simulation setup: IER onset

- Sample the item index at which IER onsets as i.i.d. draws from a Weibull distribution with location 240 (80% of all items), shape 2.2, and scale 20

$\longrightarrow$ Inspired by Bowling et al. (2021a)

Setup: IER onset

⟶ 80% probability that IER starts before having answered 90% of all items (based on estimates in Bowling et al., 2021a)



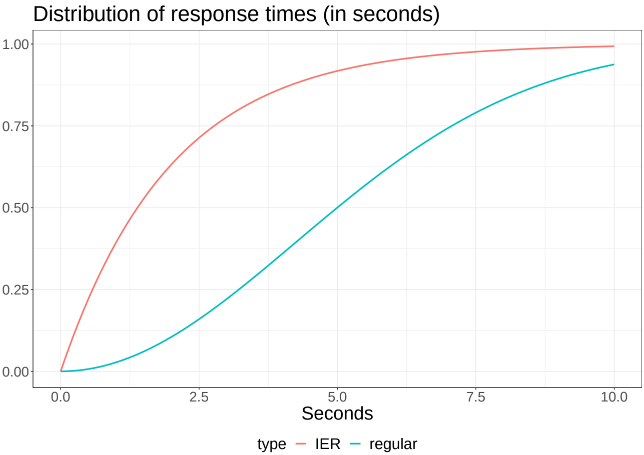
Simulation setup: Response times

- Regular: Sample per-item response times (in seconds) as i.i.d. draws from a Weibull distribution with scale 6 and shape 2
  - $\longrightarrow$ Mean ≈ 5.3 and variance ≈ 2.8
  - $\longrightarrow$ Emulates empirical observations in Schroeders et al. (2022)

- IER: Sample per-item response times (in seconds) as i.i.d. draws from a Weibull distribution with scale 2 and shape 1
  - $\longrightarrow$ Mean and variance of 2
  - $\longrightarrow$ Inspired by Bowling et al. (2021b)

Setup: Response times

→ Emulates empirical response times in Schroeders et al. (2022), also inspired by Bowling et al. (2021b)



Setup: Adding IER

- Vary the prevalence of IER: $\{0\%, 20\%, \ldots, 100\%\}$
- Four types of IER: Random responding, straightlining, imperfect straightlining, pattern responding
- Each respondent who starts IER gets randomly assigned one of the four IER types

$\longrightarrow$ Each type is present in each contaminated dataset

Setup: Types of IER

- Random responding: Respond completely at random
- Straightlining: Respond always to the same, randomly determined answer category
- Imperfect straightlining: Straightlining, but with random variation of up to ± 1 answer category
- Pattern responding (based on Schroeders et al., 2022): Respond according to a fixed, randomly determined pattern

  $\longrightarrow$ For example, 1-2-3-1-2-3, 4-5-4-5, ...

Evaluation measure

Adjusted Rand Index (ARI; Hubert and Arabie, 1985):

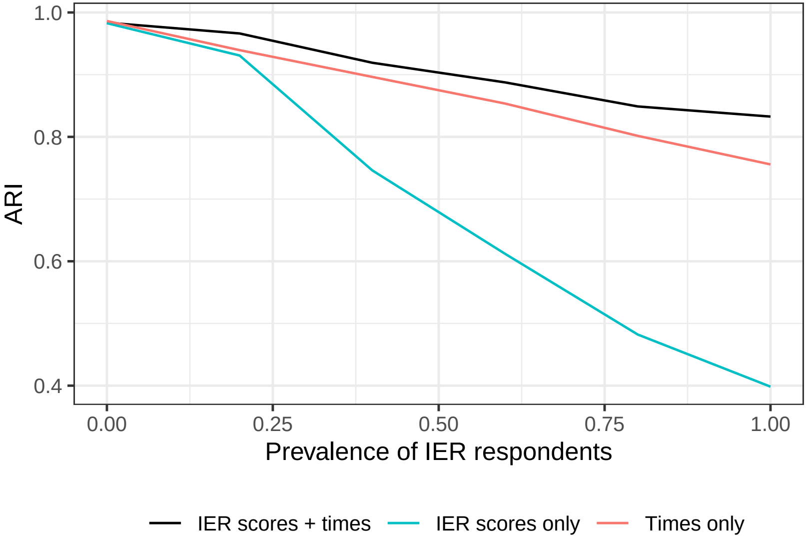
- ARI is a measure of classification performance
- Bounded between 0 (random classification) and 1 (perfect classification)

$\longrightarrow$ For each item, a simulated respondent is either IER or not

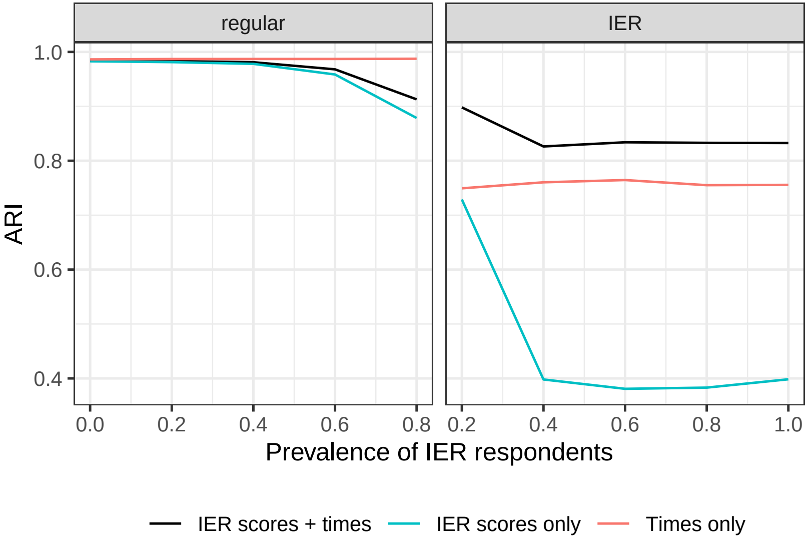
$\longrightarrow$ Calculate ARI of each respondent's series of per-item responses

$\longrightarrow$ Average the ARI over respondents

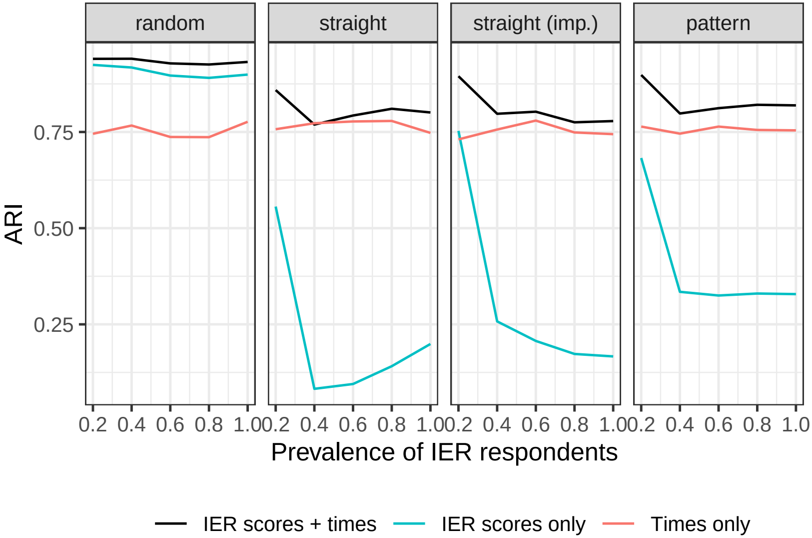
Results: ARI at level $\alpha = 0.025$



Results: ARI at level $\alpha = 0.025$

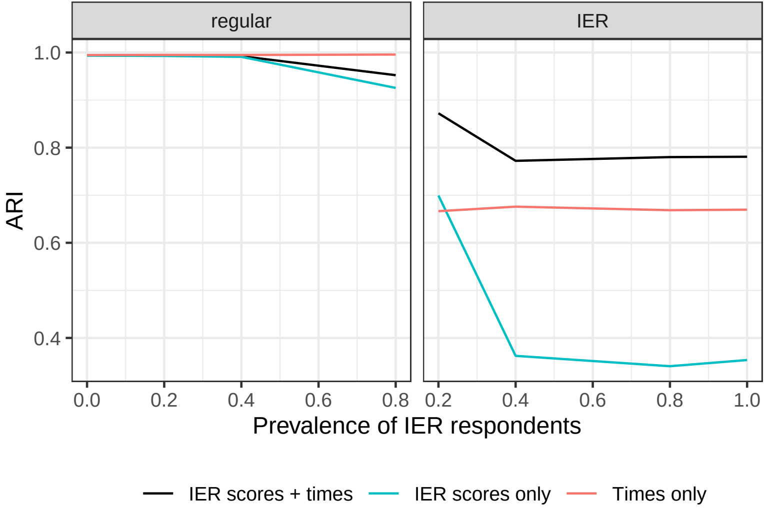


Results: ARI at level $\alpha = 0.025$



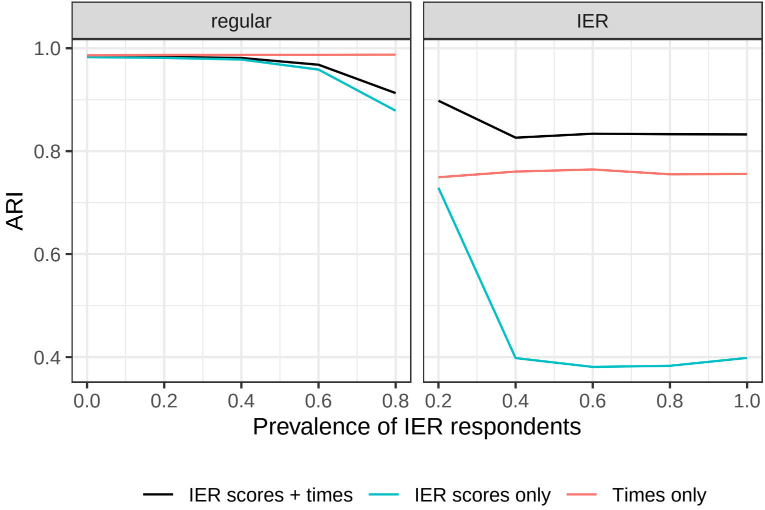
Simulation study: Size control

Results with size $\alpha = 0.01$:



Simulation study: Size control

Results with size $\alpha = 0.025$:



Simulation study: Size control

Results with size $\alpha = 0.05$: