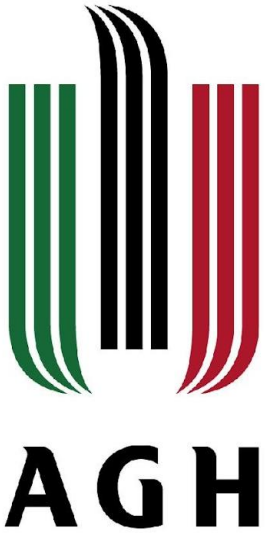


Generalised Score Distribution

A Two-Parameter Discrete Distribution Accurately Describing Responses from Quality of Experience Subjective Experiments (arXiv ID: 2202.02177)

Jakub Nawała <jnawala@agh.edu.pl>, Lucjan Janowski, Bogdan Ćmiel, Krzysztof Rusek, Pablo Pérez

VQEG SAM F2F Meeting. May 10, 2022. INSA Rennes, France.

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie | AGH University of Science and Technology

Background: General Objective

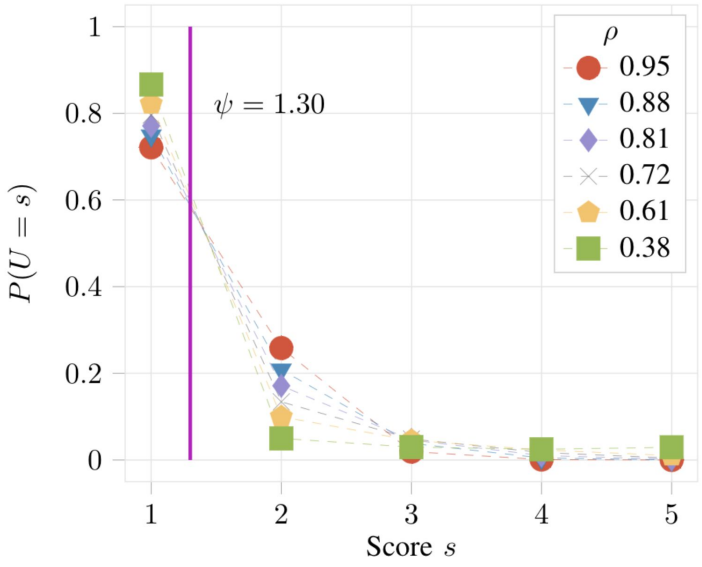
- Model the response generation process.
- Get as much information as possible from subjective responses.
- Reflect the discrete nature of responses.
- Something better than the MOS, but less complicated than 4-parameter multinomial distribution.
- We focus on the 5-level ACR scale.

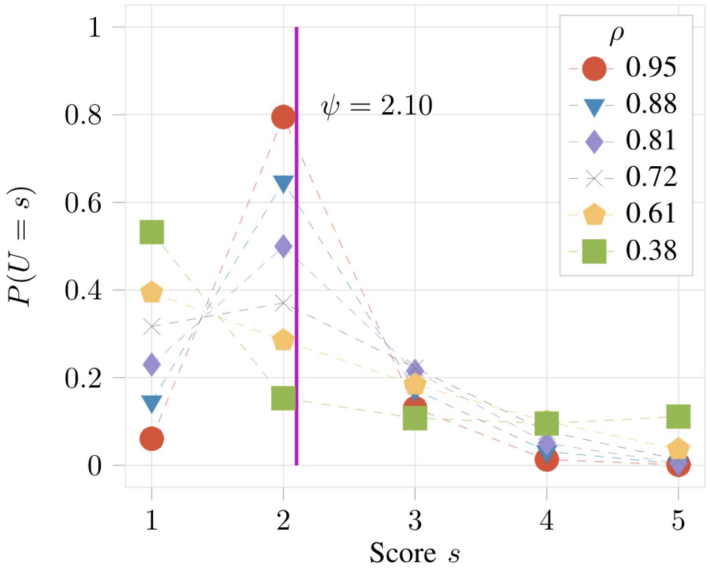
Background: The GSD

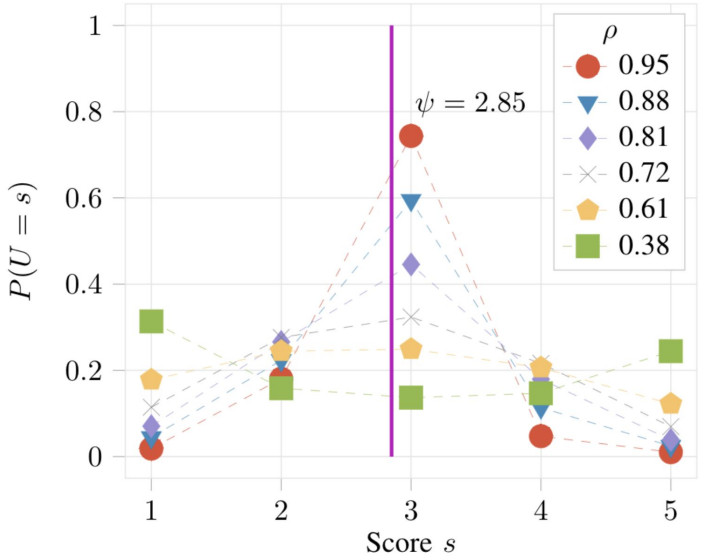
- Family of two-parameter discrete distributions.
- ψ — "true quality"; central tendency of responses.
- ϱ — response spread; confidence parameter.

$$U \sim \mathrm{GSD}(\psi, \rho)$$

Background: The GSD cont'd







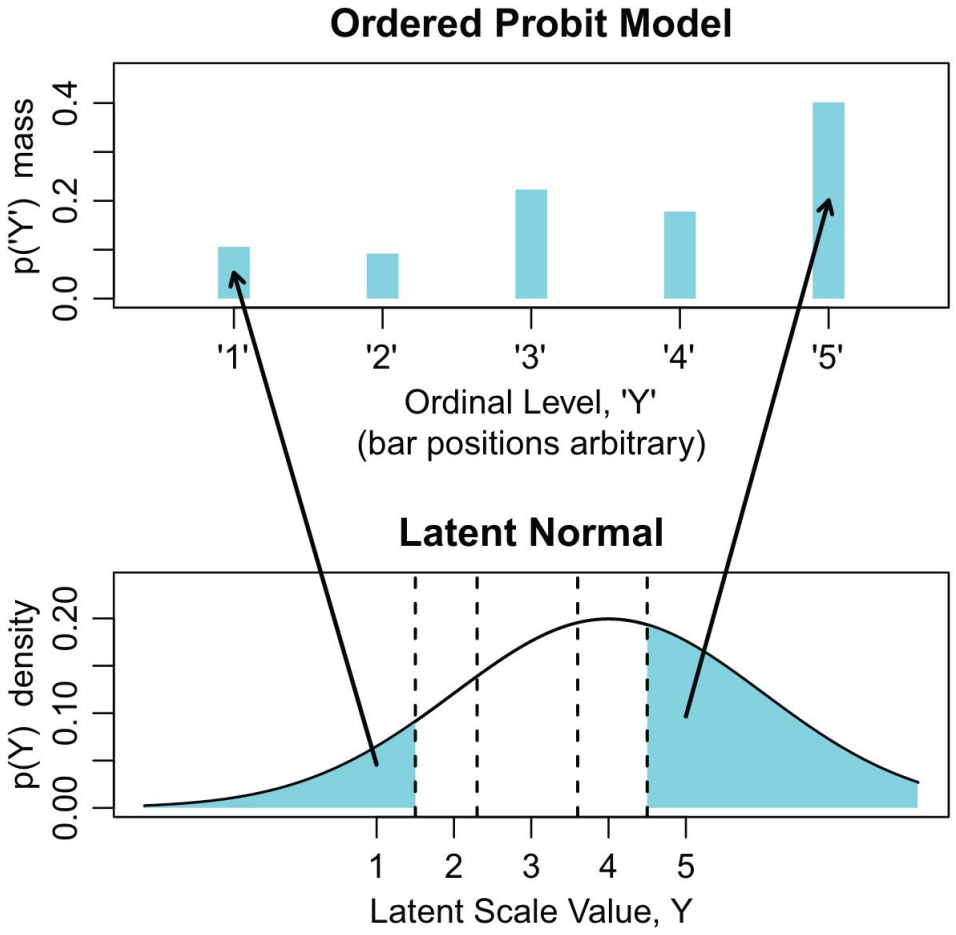
Background: The GSD

$$U \sim \psi + \epsilon$$

$$\epsilon \sim H(\rho)$$

- H satisfies the following requirements:
  - mean equals zero,
  - variance is **linearly** dependent on ϱ,
  - variance is a decreasing function of ϱ,
  - ϱ defines distribution shape,
  - H models the complete range of variance for a discrete process with limited support.

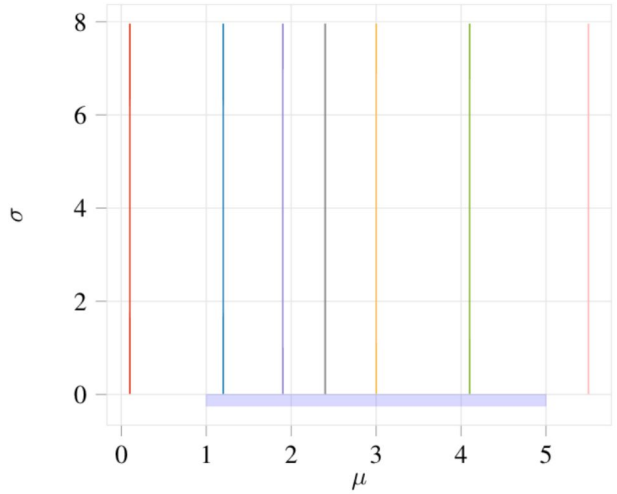
Background: Ordered Probit

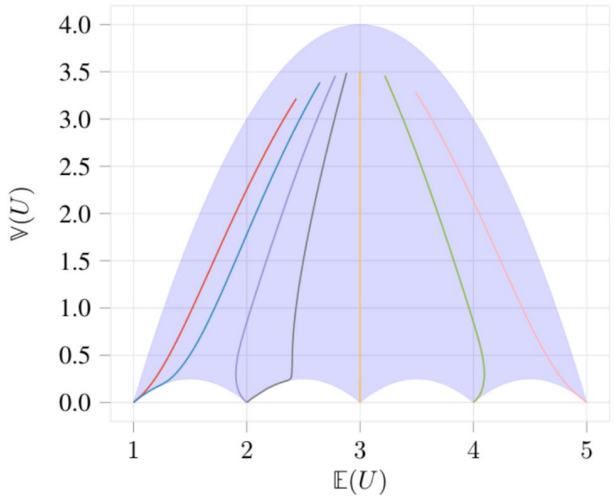


- Latent continuous normal distribution $\mathcal{N}(\mu, \sigma^2)$ mapped to a discrete distribution.
- We control μ and σ^2 to change the shape of the resultant discrete distribution.
- Unlike in the figure, we assume constant thresholds.

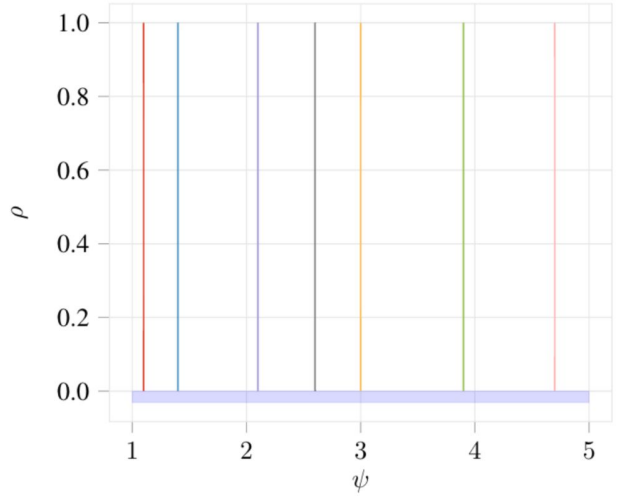
Figure taken from Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*(November 2017), 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

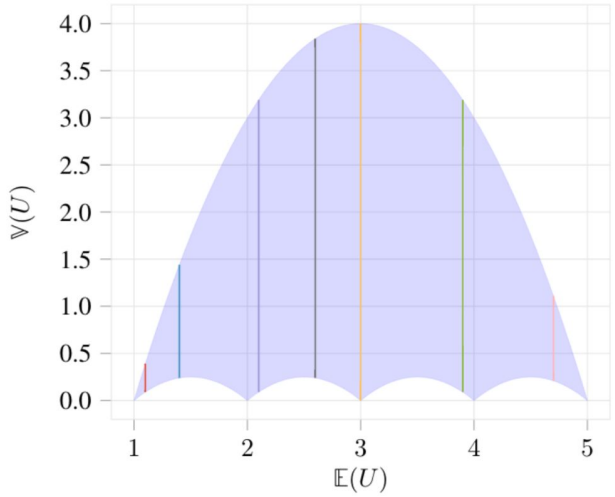
Results: Interpretable Parameters





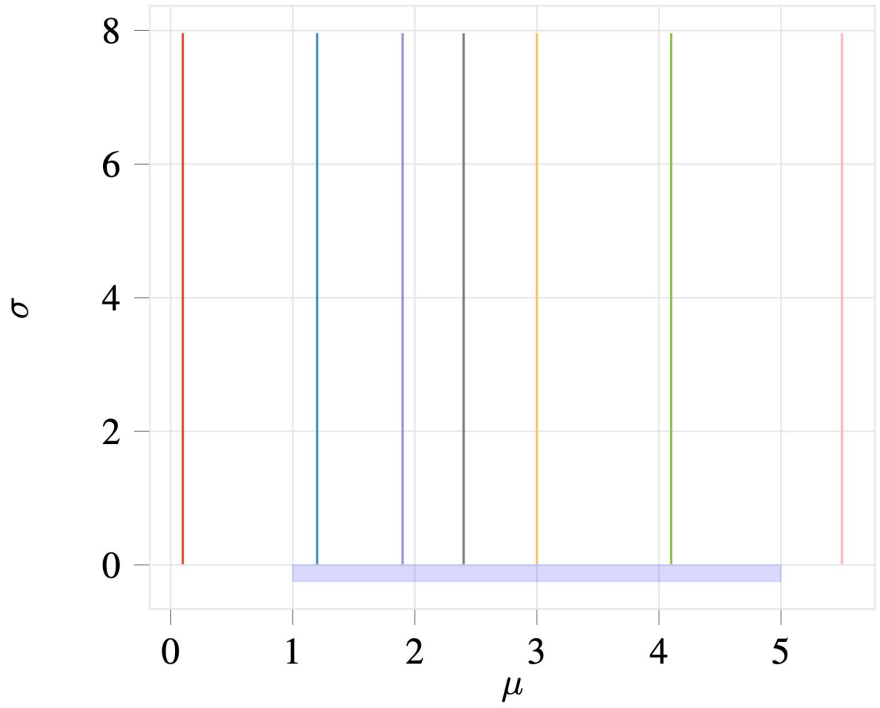
Ordered Probit

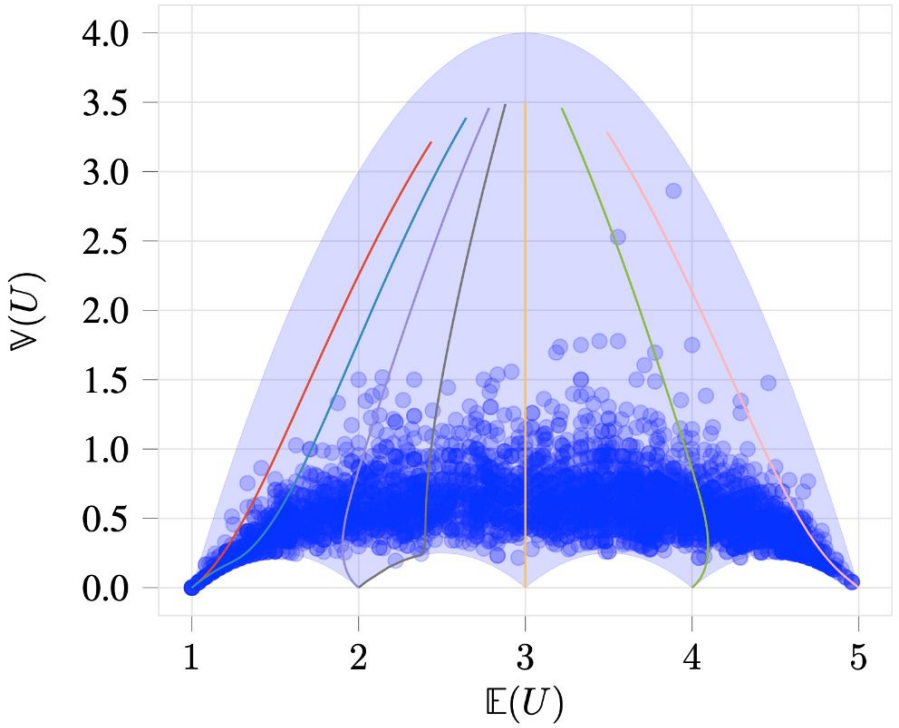




GSD

Mean-Variance Scatter Plot





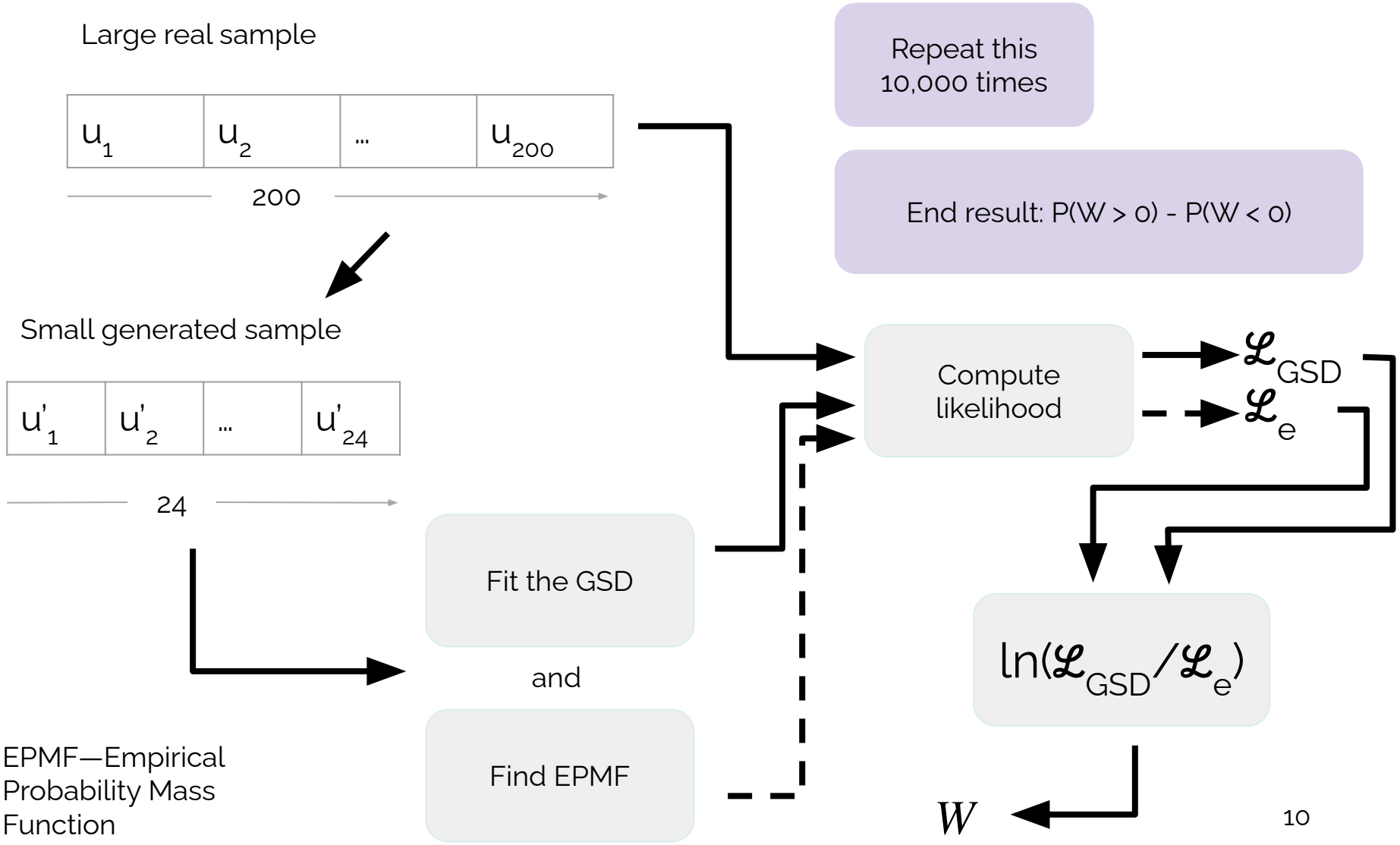
Left: The space of Ordered Probit parameters (μ and σ). Right: Mean and variance of per stimulus response distributions taken from typical MQA experiments.

Results: Bootstrapping

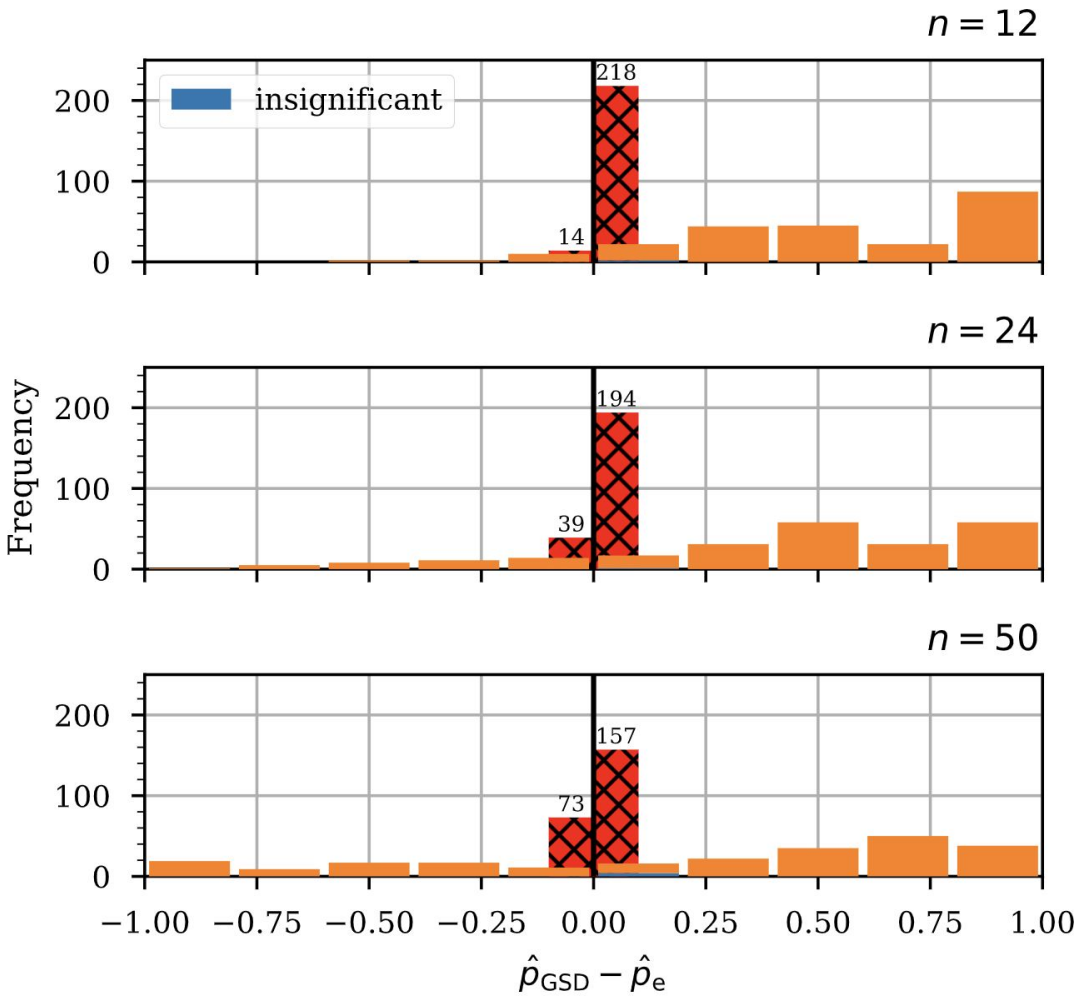
- To test generalisability potential.
- Inferring large sample distribution by observing only a small sample.
- We use selected stimuli from four QoE studies.
- Three small sample sizes: 12, 24, and 50.

No. of Responses	No. of Stimuli	Study
144	24	HDTV
200	40	NFLX
213	60	MM2
228	110	ITERO

Results: Bootstrapping



Results: Bootstrapping



$$\hat{p}_{\mathrm{GSD}} - \hat{p}_{\mathrm{e}} \approx P(W > 0) - P(W < 0)$$

Results: Bootstrapping

Conclusions

- The GSD should be preferred over the empirical distribution when resampling responses from MQA subjective experiments.
- This opens up a possibility of generating data sets of subjective responses of a size allowing to use machine learning techniques.

Discussion & Conclusions

- GSD's parameterisation is intuitive and less error-prone than Ordered Probit's one.
- GSD should be preferred over the empirical distribution when resampling subjective responses.
- GSD properly describes response distributions observed in typical MQA experiments.
- GSD outperforms the SOTA model both in terms of goodness-of-fit and bootstrapping capabilities.

Mention shortcomings of the GSD model

Thank You

- Netlifx, Inc.
- Anush Krishna Moorthy (Netflix, Inc.).
- Zhi Li (Netflix, Inc.).
- Norwegian Financial Mechanism 2014–2021
- PL-Grid Infrastructure.

Jakub Nawała

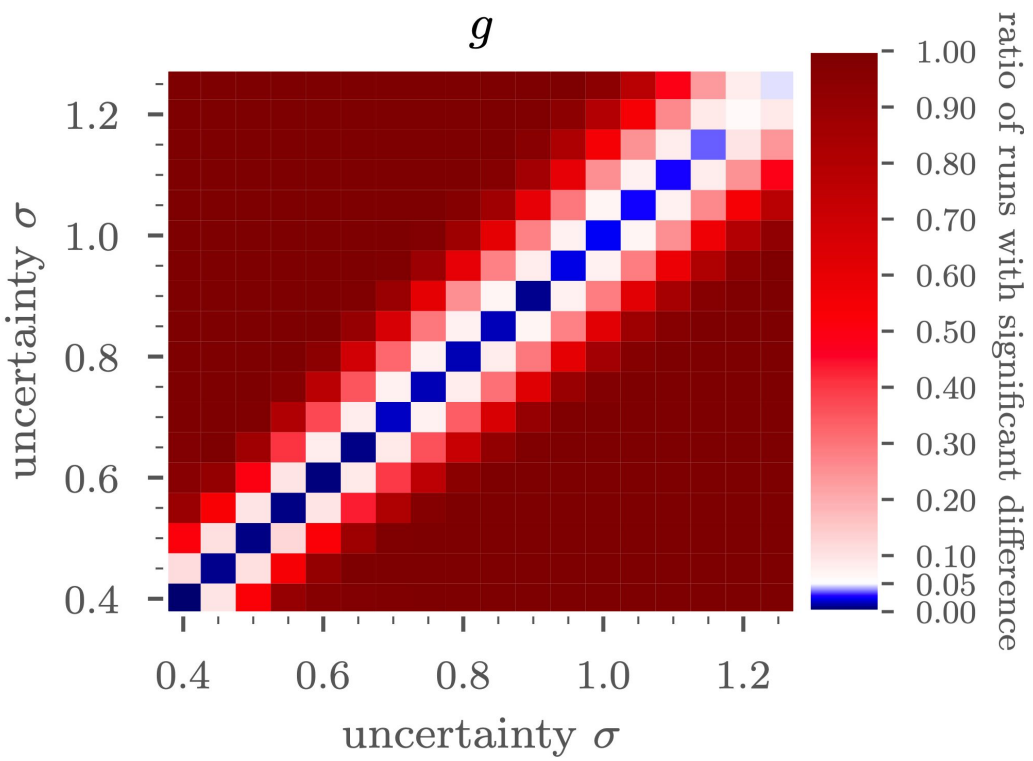
jnawala@agh.edu.pl

AGH University of Science and Technology

Institute of Telecommunications

Extra Slides

GSD as an Experiment Precision Measure



- A large-scale simulation study.
- We introduce the concept of experiment precision.
- We test how well three data modelling approaches perform in assessing experiment precision.
- The three models tested are: GSD, Li2020 (a.k.a. AP2 Model), and SOS a (a.k.a. HSE α).

[NawalaHossfeld2022] Nawała, J., Hoßfeld, T., Janowski, L., & Seufert, M. (2022). *Systematic Analysis of Experiment Precision Measures and Methods for Experiments Comparison*. https://doi.org/10.48550/arXiv.2204.07131

T-MM Reviewers' Comments

- Compare the GSD with more models.
  - Are you aware of any?
- Show practical examples presenting how GSD's parameterisation helps avoid drawing erroneous conclusions.
  - There are numerous examples of this in [Liddell2018] (e.g., see Fig. 4).

[Liddell2018] Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*(November 2017), 328–348. https://doi.org/10.1016/j.jesp.2018.08.009

T-MM Reviewers' Comments cont'd

- Are there approaches to bootstrapping NOT based on the empirical distribution?
  - Could those be used to assess GSD's performance?
- The literature review should be more comprehensive.
  - Are you aware of works related to ours?

AP2 Model

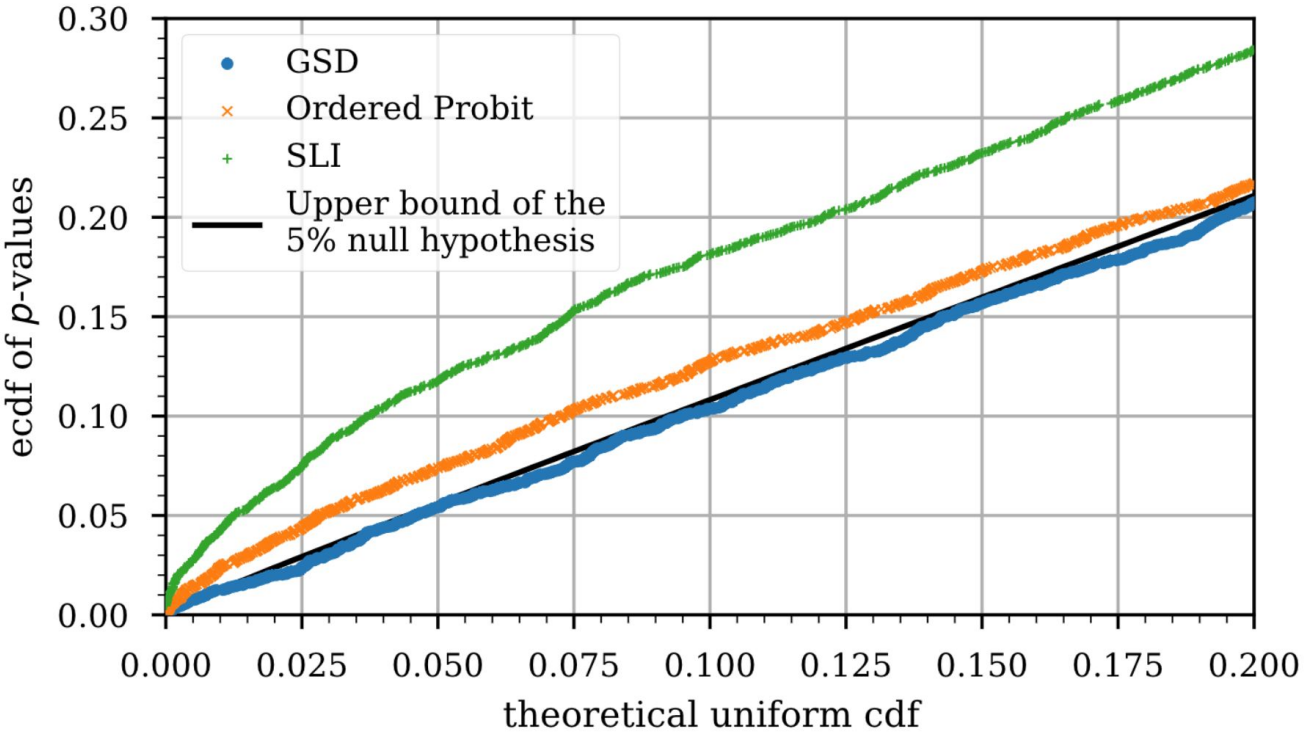


Fig. 4. p-Value P–P plot for typical MQA experiments. p-Values come from the G-test of goodness-of-fit applied to the GSD, Ordered Probit and Simplified Li2020 (SLI) models, fitted to responses from typical MQA experiments. CDF stands for cumulative distribution function and ECDF for empirical cumulative distribution function.

AP2 Model cont'd