Realistic Simulation of Item Difficulties

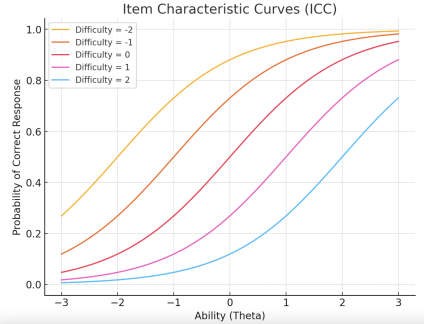
Lijin Zhang[1], Yiqing Liu[1], Dylan Molenaar[2],
Joshua Gilbert[3], Klint Kanopka[4], Ben Domingue[1]

[1] Stanford University
[2] University of Amsterdam
[3] Harvard University
[4] New York University

April 24, 2025

Item Difficulties

- Item difficulty is a key parameter in psychometric models
- A well-distributed range of item difficulties ensures assessments are fair, reliable, and effective across a wide ability spectrum



Simulation Study

- Simulations provide a controlled environment to evaluate psychometric methods and assumptions.
- In the vast majority of such studies in psychometrics, some assumption must be made about the distribution of item difficulties during data generation.
- Item difficulty is often not treated as a central focus.
- Common distributions (e.g., $N(0, 1)$, Uniform$[-2, 2]$) are used for convenience.

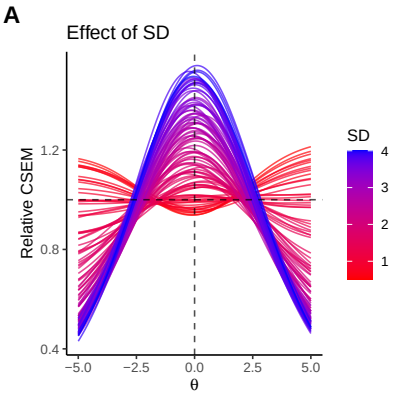
Simulation Study

A fundamental assumption is that the parameters employed to generate response data should closely resemble those observed in actual assessments.

- Normal assumption implies equal probability of extreme difficulties, which might be challenged with real-world testing conditions
- Uniform distribution may overlook common patterns in item difficulties (e.g., higher concentration near the mean).
- Real tests often exhibit asymmetry, varying standard deviations, skewness, or kurtosis.
- Variability in item difficulty distributions affects reliability and standard error of estimates.

Motivating Example

How item difficulty variability (SD, skewness, kurtosis) shapes relative Conditional Standard Error of Measurement (CSEM) across the latent trait scale



‘

- Baseline: $\beta_j \sim N(0, 1)$
- Manipulation of SD: 0.5 - 4.

A relative CSEM value greater than 1 indicates reduced measurement precision compared to the baseline, while values less than 1 indicate improved precision.

Goals

- Understand the key moments (e.g., mean, standard deviation, skewness, kurtosis) of item difficulty distributions in real-world datasets.
- Propose a new method for generating more realistic item difficulties in simulations.

Data

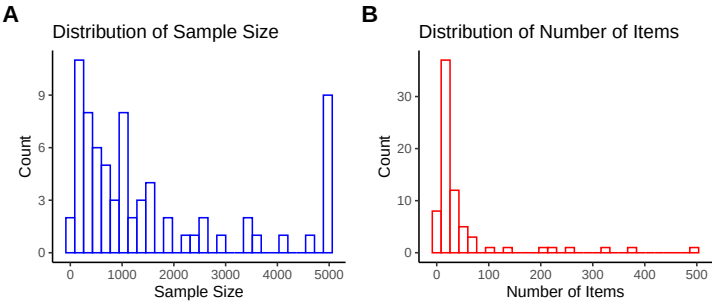
- Used the Item Response Warehouse (IRW v11.12) to examine item difficulty distributions in real-world contexts.
- Applied the Rasch model to selected datasets to estimate item difficulty parameters.
- Rescaled item difficulties for cross-dataset comparability.

Data

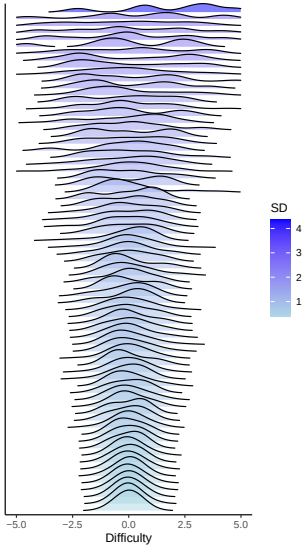
73 datasets were retained after excluding those that:

- with duplicate responses, non-binary-response, or >50% missing responses.

Subsampled datasets with >5,000 respondents to 5,000 to reduce computation while maintaining representativeness.



Results



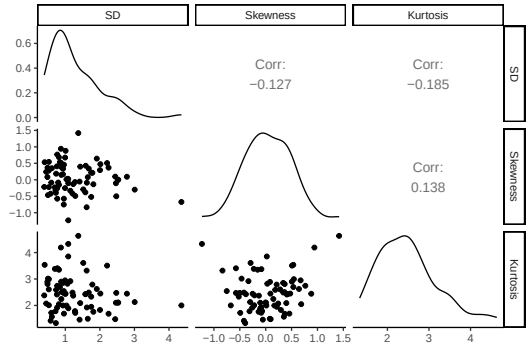
73 Dataset:

- Some distributions are centered, flat, and a few exhibit bimodal or multimodal patterns.
- Most are relatively symmetric.
- Some distributions have imbalanced difficulty.

Results

Minimal correlation among these metrics. Low SD overall.

Skewness is around zero. Kurtosis is mostly <3.



Comparison

- Real-world SDs vary widely, unlike fixed SDs in simulations (e.g., 1 for $N(0,1)$, 1.333 for $U(-2,2)$).
- Most datasets show low SD, with item difficulties tightly clustered.
- Some distributions are asymmetric or multimodal, challenging symmetry assumptions.
- Kurtosis is generally < 3, suggesting flatter distributions than $N(0,1)$.

Fixed simulation distributions cannot fully capture real-world item difficulty patterns.

A New Way to Generate Difficulty Parameters

To improve realism and generalizability, simulations could incorporate empirical insights from actual datasets.

A New Way to Generate Difficulty Parameters

Dataset Pool Preparation and Random Dataset Selection.

Item Difficulty Distribution Construction:

- Each difficulty estimate is modeled as $\mathcal{N}(\beta_{est}, \text{SE})$.
- Form a mixture distribution by combining the normal distributions of all difficulty estimates.
- Estimate a smooth probability density function (PDF) and derive the cumulative distribution function (CDF).

Item Difficulty Sampling:

- Draw uniform random values (0–1) and map them to difficulty values via the inverse CDF.

Repetition for Replications.

A New Way to Generate Difficulty Parameters

Generating Realistic Item Difficulties

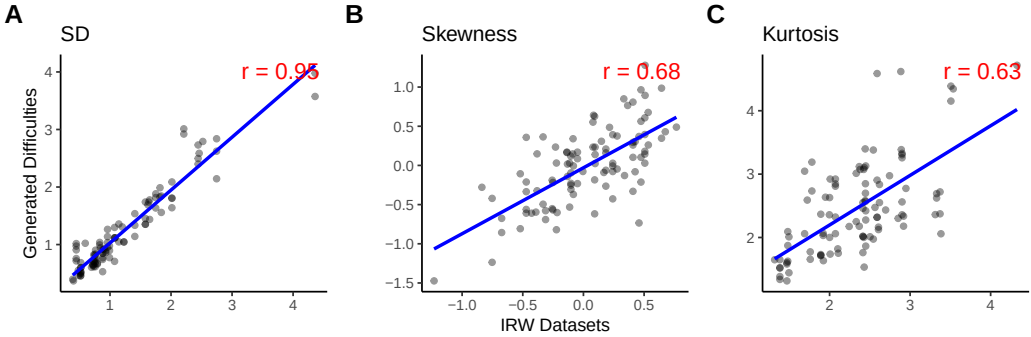
```r
# difficulty parameters from IRW datasets
difficulty = read.csv('diff_long.csv')

# seed setting for simulation reproducibility
set.seed(1)

# usage of the simu_item_diff() function
simulated_difficulties <- simu_item_diff(
  difficulty_data = difficulty,
  num_items = 25,
  num_replications = 100
)
# difficulty_data indicated the difficulty parameter pool
# num_items and num_replications define the number of items and
# the number of simulation replications, respectively
# min_items serves as a threshold for dataset inclusion
# excluding those with an item count below the specified value
```

- available at `https://anonymous.4open.science/r/irw_diff-2E1F`

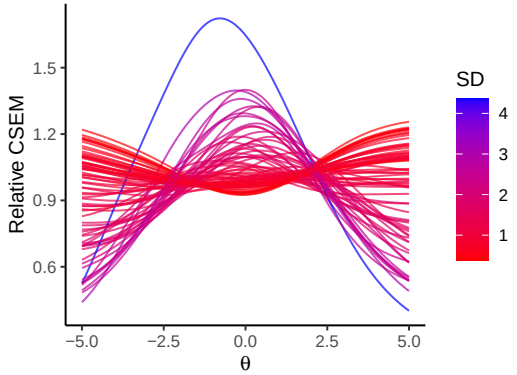
Correlation



The generated difficulties closely mirror key features of real-world distributions.

Variation

Relative CSEM varies when item difficulties are sampled from real IRW datasets versus a standard normal distribution:



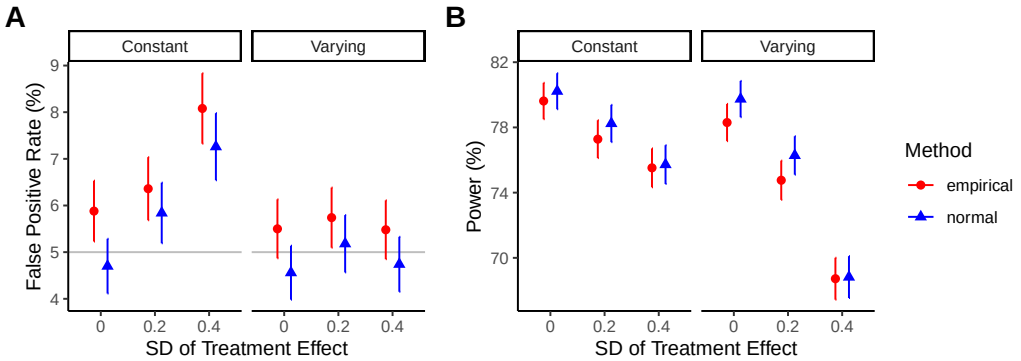
Simulation

Gilbert et al., (2023) examined how constant versus varying treatment effect models perform in estimating item-level heterogeneous treatment effects:

- Average treatment effect ($\tau = 0$ vs. $\tau = 0.4$)
- Standard deviation of item-level treatment effects ($\sigma_\tau = 0, 0.2, 0.4$)
- Model: constant vs. varying treatment effects
- Item difficulty generation method ($N(0, 1)$ vs. empirical)

Each condition: 5000 replications, 500 individuals, 20 items.

Simulation



- The realistic item difficulty condition tended to yield slightly higher false positive rates and lower power.

Takeaways

- Simulations often assume fixed normal or uniform distributions (e.g., $N(0,1)$, $U(-2,2)$), which poorly reflect real-world item difficulty variability.
- We analyzed many real-world datasets from the IRW to document actual difficulty distributions—highlighting variability in SD, skewness, and kurtosis.
- We proposed a new approach for generating realistic item difficulties.
- We encourage future simulation studies to adopt it for more accurately evaluating the performance of psychometric models in real-world settings.

Preprint

Thank you!

Contact: lijinzhang.com

Preprint: