Realistic Simulation of Item Difficulties

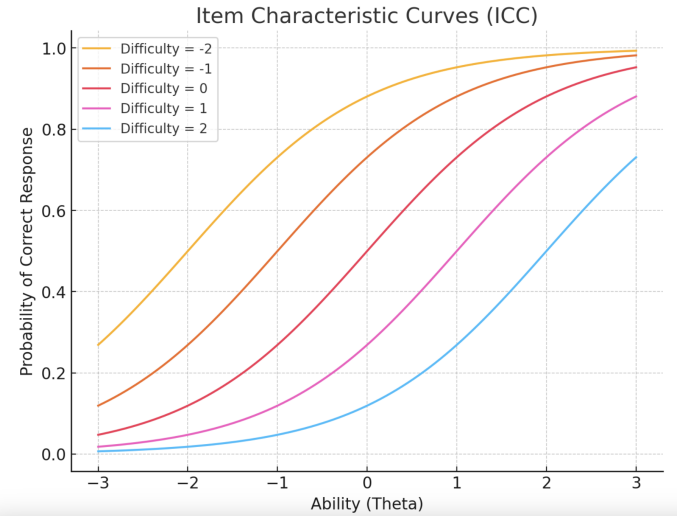
Lijin Zhang[1], Yiqing Liu[1], Dylan Molenaar[2],
Joshua Gilbert[3], Klint Kanopka[4], Ben Domingue[1]

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

Nov 11, 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

- Simulation provides 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.

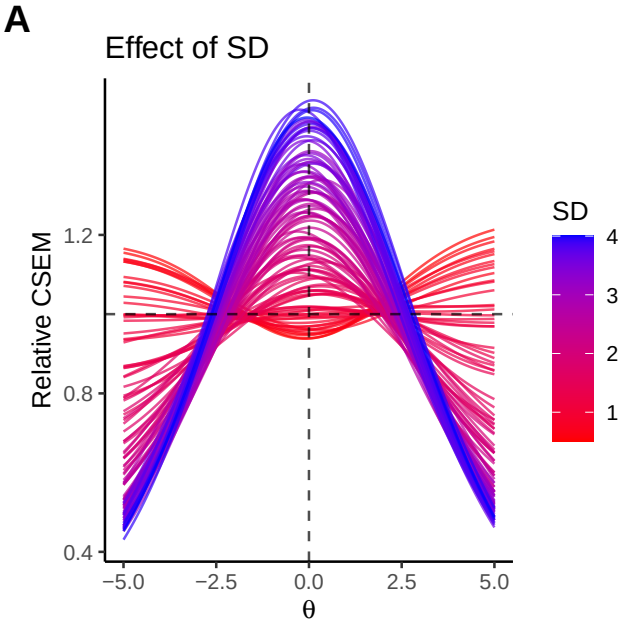
Simulation Study

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 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.

Aim

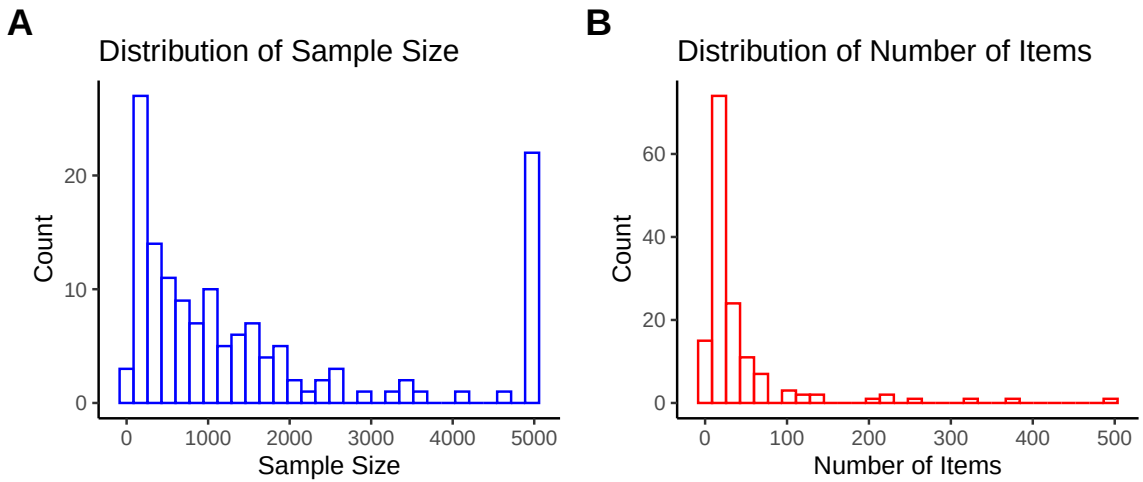
- Understand the key moments (e.g., mean, standard deviation, skewness, kurtosis) of item difficulty distributions in real-world datasets.
- Propose a new method for generating more realistic item difficulties in simulations.

Data

Item Response Warehouse (Domingue et al., 2025).
145 datasets were retained after excluding those that:

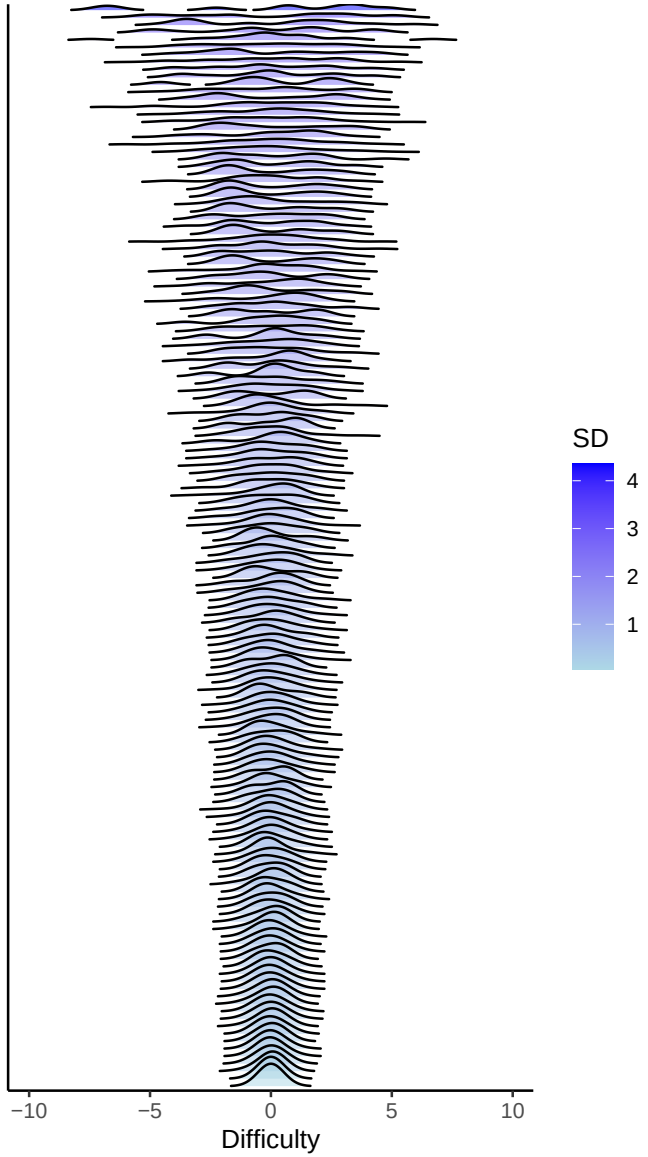
- with duplicate responses, non-binary-response, or >50% missing responses.

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



Applied the Rasch model to obtain difficulty parameters.

Results



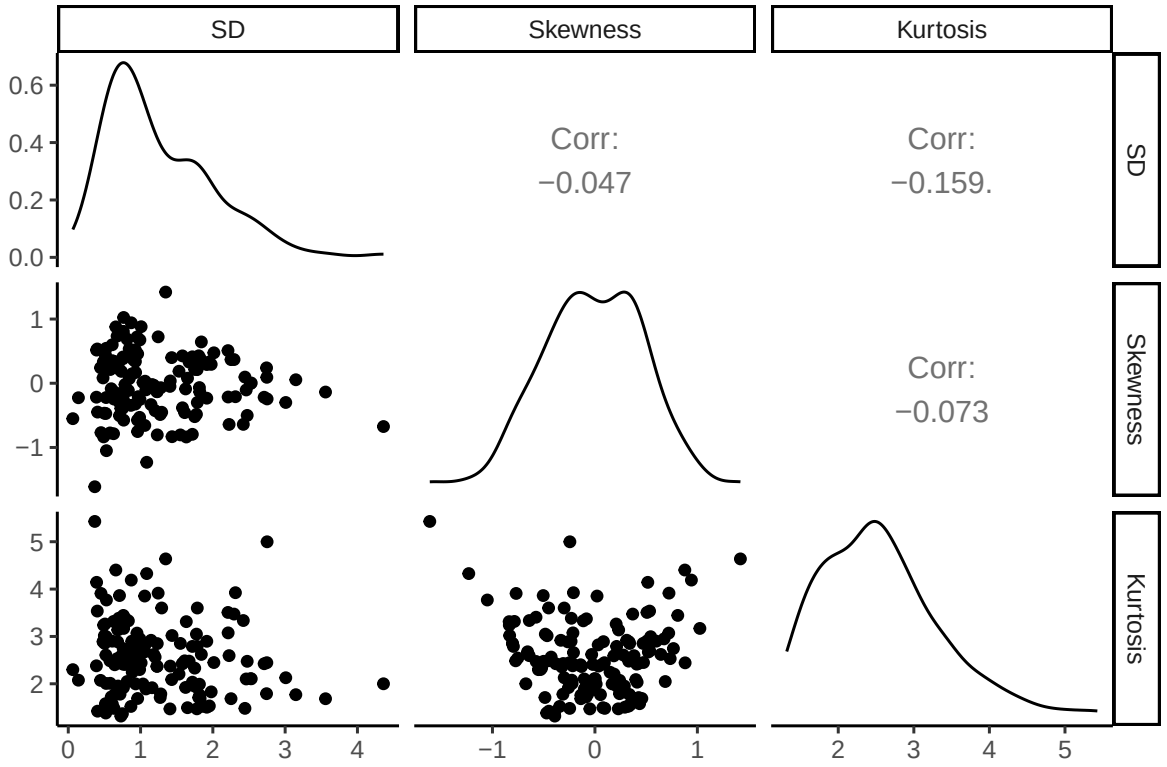
145 Datasets:

- 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.

Skewness is around zero. Kurtosis is mostly <3.



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

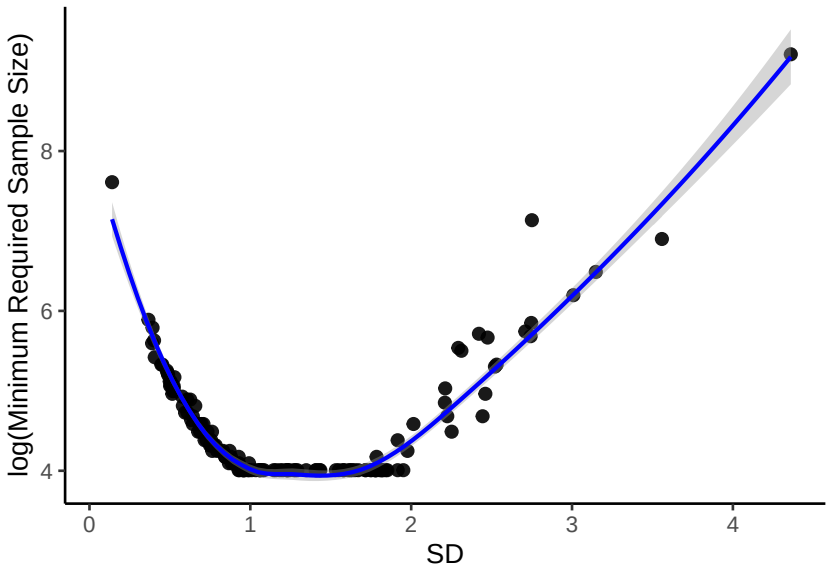
Sample Size Simulation

We ran a small simulation to show why the item difficulty distributions matter. For each dataset in the pool, the procedure consisted of the following steps:

- Generate item difficulties: We sampled 50 items according to the empirical difficulty distribution from that dataset, and repeated this process 100 replications.
- Simulate response data and estimate parameters for a given sample size using a Rasch model.

Sample Size Simulation

- Search for the minimal sufficient sample size: We used a binary search between $= 50$ and 10,000.
- The search terminated once the $\text{cor}(\beta_{est}, \beta_{true}) > 0.95$ and the gap between the lower and upper bounds of the search interval was less than 10.



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

```
# download the package
devtools::install_github("itemresponsewarehouse/Rpkg")

# load the difficulty pool from the package
data("diff_long", package = "irw")

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

# usage of the simu_item_diff() function
simulated_difficulties <- irw::irw_simu_diff(
  num_items = 25,
  num_replications = 100,
  difficulty_pool = diff_long
)
# num_items and num_replications define the number of items and
# the number of simulation replications, respectively
```

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.
- This method serves as a framework for researchers who wish to explore how diverse, empirically grounded item structures may influence their findings, thereby enhancing the realism and applicability of simulation results.

Preprint

Thank you!

Contact: lijinzhang.com

Preprint: