

# Item Response Theory for NLP

EACL2024 Tutorial, 21<sup>st</sup> March 2024

John P. Lalor, Pedro Rodriguez, João Sedoc, Jose Hernandez-Orallo

<https://eacl2024irt.github.io/>

## In this session

Motivation

Introducing IRT

IRT Models with Artificial Crowds

The py-irt Package

## Motivation

---

## Natural language inference (NLI)

Premise	Hypothesis	Label	Difficulty
A little girl eating a sucker	A child eating candy	Entailment	<i>easy</i>
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	<i>hard</i>
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	<i>easy</i>

## Sentiment analysis (SA)

Phrase	Label	Difficulty
The stupidest, most insulting movie of 2002's first quarter.	Negative	<i>easy</i>
Still, it gets the job done - a sleepy afternoon rental.	Negative	<i>hard</i>
An endlessly fascinating, landmark movie that is as bold as anything the cinema has seen in years.	Positive	<i>easy</i>
Perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions.	Positive	<i>hard</i>

# Leaderboards

## 🤖 Open LLM Leaderboard

🚩 The 🤖 Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

🤖 Submit a model for automated evaluation on the 🤖 GPU cluster on the "Submit" page! The leaderboard's backend runs the great [Eleuther AI Language Model Evaluation Harness](#) - read more details in the "About" page!

🏆 LLM Benchmark  Metrics through time  About  Submit here!

🔍 Search for your model (separate multiple queries with ';' and press ENTER...

Select columns to show

- Average  ARC  HellaSwag  MMLU  TruthfulQA  Winogrande  
 GSM8K  DROP  Type  Architecture  Precision  Hub License  
 #Params (B)  Hub  Available on the hub  Model sha

Show gated/private/deleted models

Model types

- pretrained   fine-tuned   instruction-tuned   RL-tuned  ?

Precision

- float16  bfloat16  8bit  4bit  GPTQ  ?

Model sizes (in billions of parameters)

- ?  -1.5  -3  -7  -13  -35  -60  70+

T	Model	Average	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	DROP
🔹	<a href="#">TigerResearch/tigerbot-70b-chat-v2</a>	69.76	87.03	82.83	66	75.4	79.16	46.02	51.9
🔸	<a href="#">bhenrym14/platypus-yi-34b</a>	68.96	68.43	85.21	78.13	54.48	84.06	47.84	64.55
🟢	<a href="#">@1-ai/Yi-34B</a>	68.68	64.59	85.69	76.35	56.23	83.03	50.64	64.2
🟢	<a href="#">chargoddard/Yi-34B-Llama</a>	68.4	64.59	85.63	76.31	55.6	82.79	49.51	64.37
🔸	<a href="#">MayaPH/GozillaLa2-70B</a>	67.01	71.42	87.53	69.88	61.54	83.19	43.21	52.31

# Differences in Questions

Compare Two Systems

Question



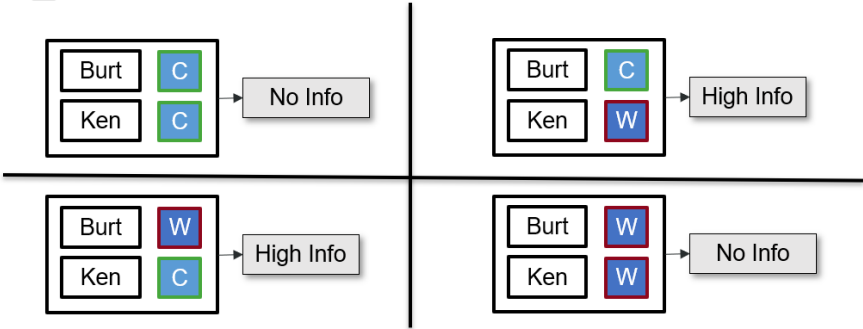
Burt



Ken

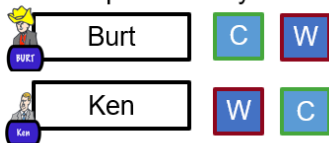


**Question:** Who did the Normans team up with in Anatolia?



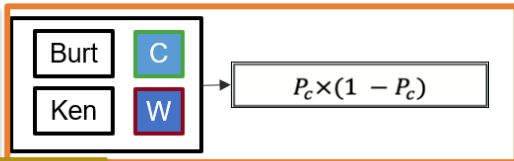
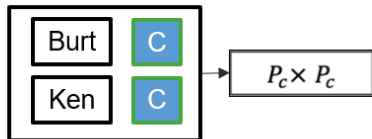
## Differences in Questions

### Compare Two Systems

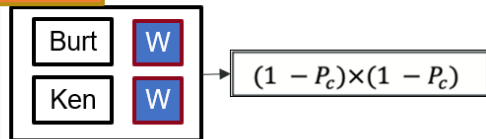
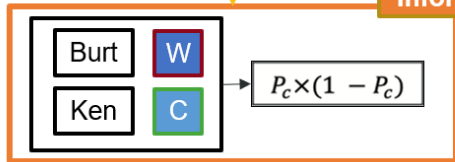


$P_c$  = Correct Probability,  $P_w$  = Wrong Probability

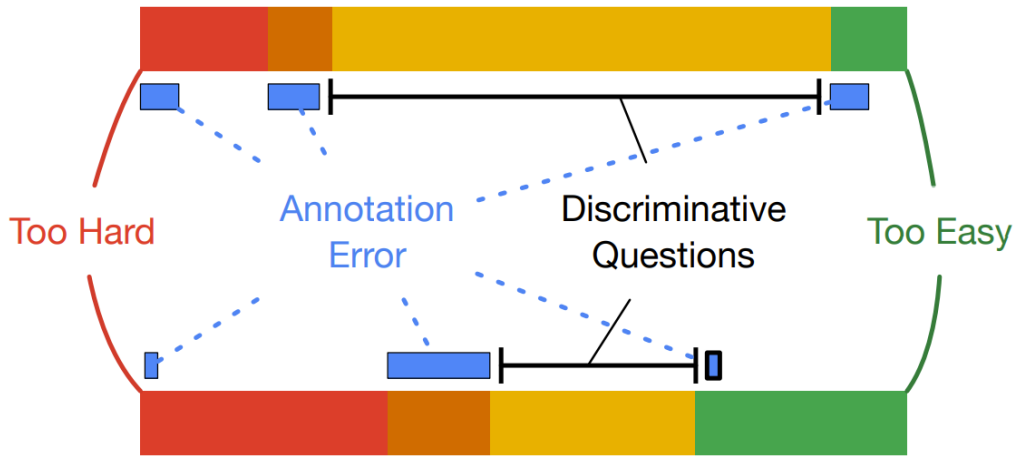
$$P_w = 1 - P_c$$



We're Informed Here



## Differences in Questions





## Introducing IRT

---

Psychometrics: study of quantitative measurement practices

- Building instruments for measurement (**standardized tests**)
- Development of theoretical approaches to measurement

Item Response Theory (IRT): measure latent traits of test-takers and test questions (“items”)



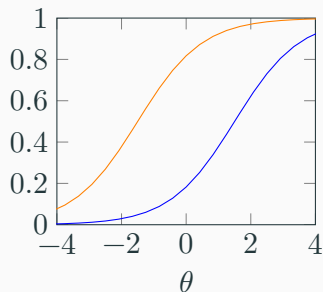
## IRT: 1 Parameter Logistic Model (1PL)

Also known as *Rasch model*

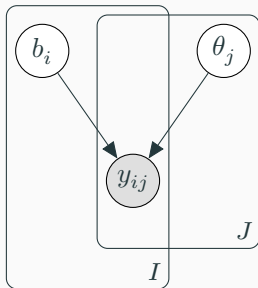
$$p(y_{ij} = 1 | b_i, \theta_j) = \frac{1}{1 + e^{-(\theta_j - b_i)}}$$

$\theta_j$ : latent ability

$b_i$ : difficulty



## 1PL Plate Notation



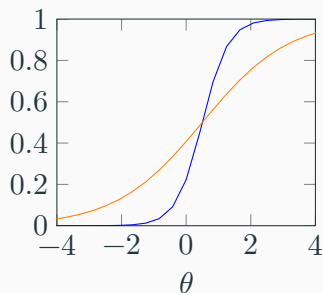
## IRT: Other Examples (2PL)

$$p(y_{ij} = 1 | a_i, b_i, \theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$

$\theta_j$ : latent ability

$b_i$ : difficulty

$a_i$ : discriminability



## IRT: Other Examples (3PL)

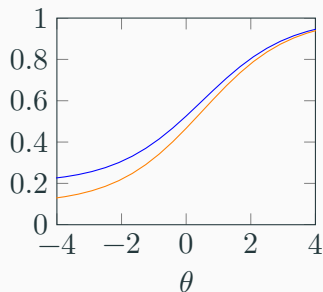
$$p(y_{ij} = 1 | a_i, b_i, c_i, \theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

$\theta_j$ : latent ability

$b_i$ : difficulty

$a_i$ : discriminability

$c_i$ : guessing



## IRT: Other Examples (Feasibility)

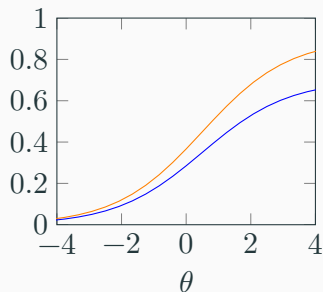
$$p(y_{ij} = 1 | a_i, b_i, c_i, \theta_j) = \frac{\gamma_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

$\theta_j$ : latent ability

$b_i$ : difficulty

$a_i$ : discriminability

$\gamma_i$ : feasibility



$$p(y_{ij} = 1|b_i, \theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$
$$p(y_{ij} = 0|b_i, \theta_j) = 1 - p(y_{ij} = 1|b_i, \theta_j)$$

$$L = \prod_{j=1}^J \prod_{i=1}^I p(Y_{ij} = y_{ij}|b_i, \theta_j)$$
$$q(\Theta, B) = \prod_j \pi_j^\theta(\theta_j) \prod_i \pi_i^b(b_i)$$

- $p(Y|B, \Theta)$  – model
- $q(\Theta, B)$  – guide (variational distribution)



Let's look at the code

Intro to IRT notebook 1 – 2\_IntroToIrt.ipynb

## Evaluating DNN Performance with IRT

Premise	Hypothesis	Label	Difficulty
A little girl eating a sucker	A child eating candy	Entailment	-2.74
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	0.51
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	-1.92
Nine men wearing tuxedos sing	Nine women wearing dresses sing	Contradiction	0.08

Phrase	Label	Difficulty
The stupidest, most insulting movie of 2002's first quarter.	Negative	-2.46
Still, it gets the job done - a sleepy afternoon rental.	Negative	1.78
An endlessly fascinating, landmark movie that is as bold as anything the cinema has seen in years.	Positive	-2.27
Perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions.	Positive	2.05

## IRT for NLP: Human Annotations

Item Set	Ability Score	Percentile	Test Acc.
<b>“Easier”</b>			
Entailment	-0.133	44.83%	96.5%
Contradiction	1.539	93.82%	87.9%
Neutral	0.423	66.28%	88%
<b>“Harder”</b>			
Contradiction	1.777	96.25%	78.9%
Neutral	0.441	67%	83%

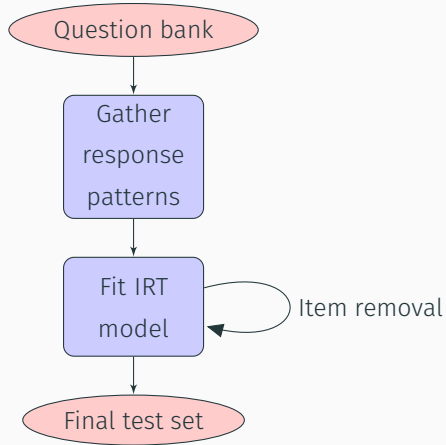
Source: Lalor et al. (2016)

- Gathering human response patterns is expensive
- Can we use ensembles of models to gather response patterns?
- Even if we can, should we?

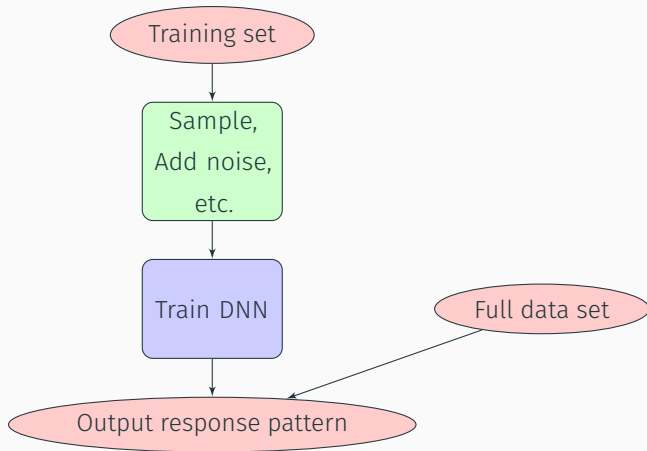
## IRT Models with Artificial Crowds

---

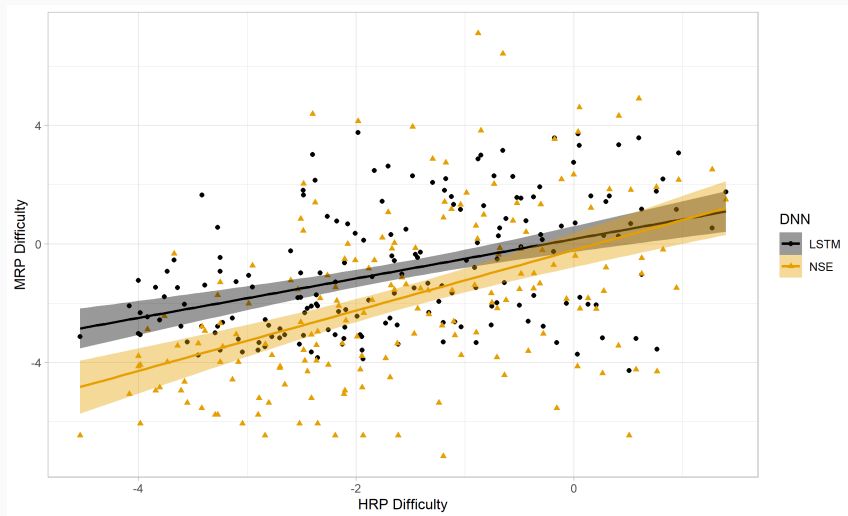
## Building IRT Test Sets



# Artificial Crowd Construction



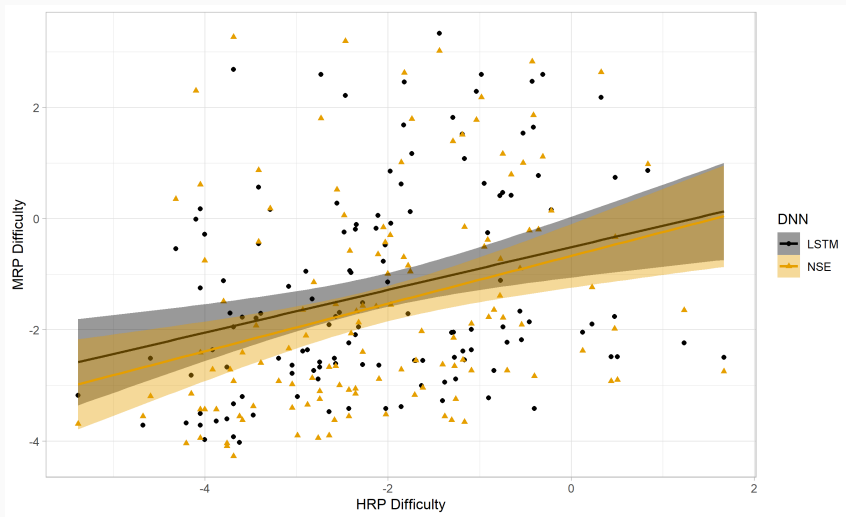
# Human-Machine Correlation



- Spearman  $\rho$  (NLI): 0.409 (LSTM) and 0.496 (NSE) (Lalor et al., 2019)

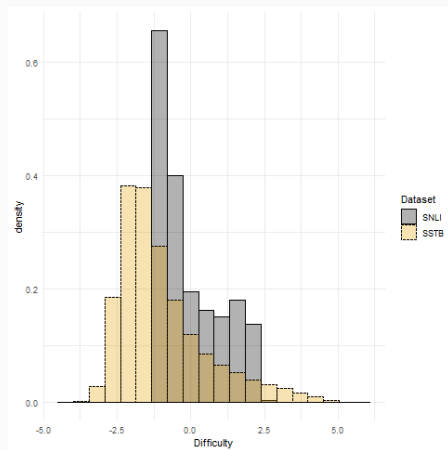


# Human-Machine Correlation



- Spearman  $\rho$  (SA): 0.332 (LSTM) and 0.392 (NSE) (Lalor et al., 2019)

# Difficulty Distribution



Source: Lalor et al. (2019)

# IRT for Leaderboards (SQuAD)

System Developer



Runnable System



What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

SQuAD 2.0 contains the 100,000 questions in SQuAD 1.1 with over 10,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD 2.0, systems must not only answer questions when possible, but also determine when the answer is supported by the paragraph and decline from answering.

[SQuAD 2.0 Leaderboard v1.1](#)

[SQuAD 2.0 Leaderboard v1.1](#)

SQuAD 1.1, the previous version of the SQuAD dataset, contains 100,000 question-answer pairs on 100k articles.

[SQuAD 1.1 Leaderboard v1.1](#)

[SQuAD 1.1 Leaderboard v1.1](#)

Leaderboard

SQuAD 2.0 tests the ability of a system to not only answer reading comprehension questions, but also detect when provided with a question that cannot be answered based on the provided paragraph.

Rank	Model	EM	F1
	Human Performance	88.9%	91.4%
	Stanford University		
	<a href="#">Rajpuria &amp; So et al. '18</a>		
1	DA Net on Albert (ensemble)	86.7%	91.8%
	CDLNet		
2	DA Net V2 (ensemble)	86.4%	91.8%
	CDLNet		
3	Baino Reader (ensemble)	86.0%	91.8%
	Shanghai Jiao Tong University		
	<a href="#">http://arxiv.org/abs/1808.09942</a>		
4	AT&T-P4 (ensemble)	86.0%	91.8%
	ETH Zurich		
5	ELECTRA-ALBERT-ensemblePlus (ensemble)	86.0%	91.8%
	DFKI, DLRG		
6	ELECTRA-ALBERT-ensemblePlus (ensemble)	86.0%	91.7%
	DFKI, DLRG		
7	ALBERT + DAAR + Reader (ensemble)	86.0%	91.7%
	PROGON Open Studio		

Dev Questions



Test Questions



Runnable System

Dev Predictions



Test Predictions



SQuAD Scoring Script

Dev Scores



Test Scores



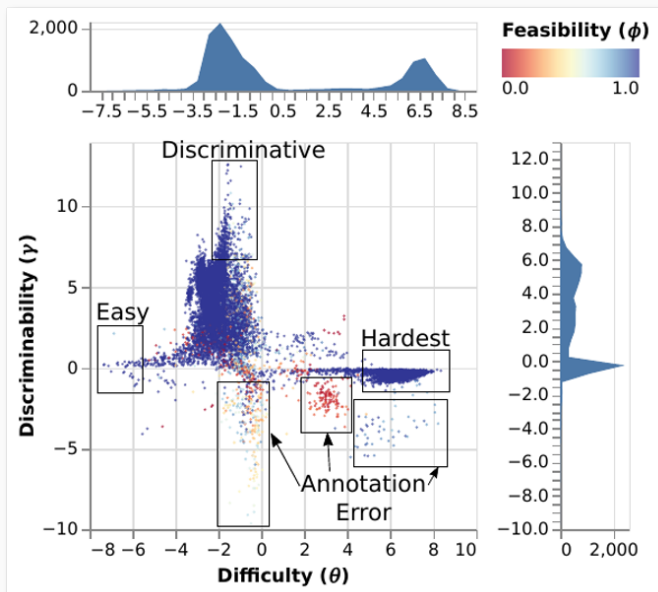
This is our data

66%

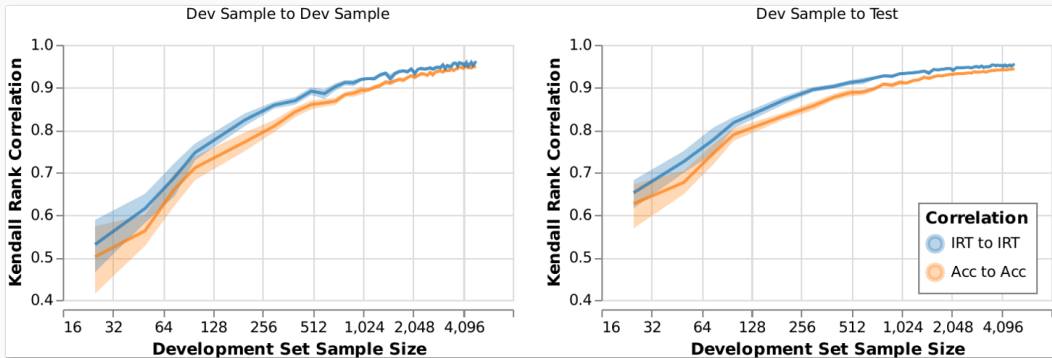
33%

- 1.9 million subject-item pairs

# IRT for SQuAD



# Ranking Performance



Source: Rodriguez et al. (2021)

## The py-irt Package

---

## IRT in Python: py-irt

```
{"subject_id": "pedro", "responses": {"q1": 1, "q2": 0, "q3": 1, "q4": 0}}
{"subject_id": "pinguino", "responses": {"q1": 1, "q2": 1, "q3": 0, "q4": 0}}
{"subject_id": "ken", "responses": {"q1": 1, "q2": 1, "q3": 1, "q4": 1}}
{"subject_id": "burt", "responses": {"q1": 0, "q2": 0, "q3": 0, "q4": 0}}
```

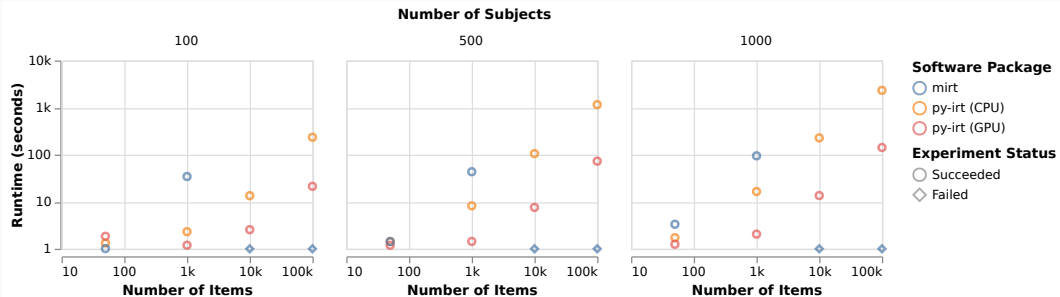
```
py-irt train 1pl data/data.jsonlines output/1pl/
```

```
{
  "ability": [
    -1.7251124382019043,
    -0.06789101660251617,
    1.6059941053390503,
    -0.20248053967952728
  ],
  "diff": [
    0.008014608174562454,
    9.654741287231445,
    -5.5452165603637695,
    -0.2792229950428009
  ],
  1,

```

```
"irt_model": "1pl",
"item_ids": {
  "0": "q2",
  "1": "q4",
  "2": "q1",
  "3": "q3"
},
"subject_ids": {
  "0": "burt",
  "1": "pinguino",
  "2": "ken",
  "3": "pedro"
}
}
```

# IRT in Python: py-irt



Search projects

## py-irt 0.6.0

```
pip install py-irt
```

Bayesian IRT models in Python

## Contributors 6



<https://github.com/nd-ball/py-irt>

Lalor and Rodriguez (2022)



Let's look at the code

Intro to IRT notebook 2 – 2\_pyirt\_example.ipynb

# References

Frank B Baker. 2001. *The basics of item response theory*. ERIC.

Jordan Boyd-Graber and Benjamin Börschinger. 2020. What question answering can learn from trivia nerds. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7422–7435, Online. Association for Computational Linguistics.

John P. Lalor and Pedro Rodriguez. 2022. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*.

John P. Lalor, Hao Wu, and Hong Yu. 2016. Building an evaluation scale using item response theory. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 648–657, Austin, Texas. Association for Computational Linguistics.

John P. Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259, Hong Kong, China. Association for Computational Linguistics.

Prathiba Natesan, Ratna Nandakumar, Tom Minka, and Jonathan D Rubright. 2016. Bayesian prior choice in irt estimation using mcmc and variational bayes. *Frontiers in psychology*, 7:1422.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4486–4503, Online. Association for Computational Linguistics.

# Break!

- Back in 15 minutes
- Next section: IRT in NLP