Item Response Theory for NLP

EACL2024 Tutorial, 21st March 2024

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

https://eacl2024irt.github.io/

In this session

Introduction

Improving Model Training

Finding Annotation Error

Evaluation Metrics

Introduction

IRT for NLP

Overview of IRT Applications:

- Dataset Construction
- Model Training
- Evaluation

Assumptions for IRT + NLP

Basic assumptions of the data and parameterization we have:

- A dataset with items indexed by i.
- A set of subjects indexed by j.
- Responses r_{ij} from graded responses of subjects to each item.
- An IRT parameterization, e.g., one with item difficulty β_i, discriminability γ_i, and ability θ_j might assume:

$$p(r_{ij} = 1|\beta_i, \theta_j) = \frac{1}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

IRT Applications: Example of Model Behavior

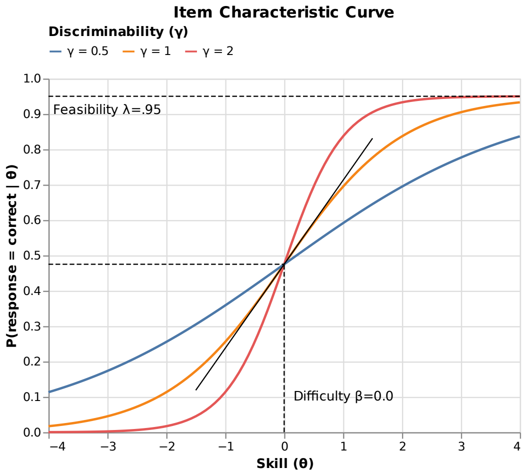
Likelihood of correct answer for subject j on item i.

$$\overbrace{p(y_{ij} = 1 | \gamma_i, \beta_i, \lambda_i, \theta_j)}^{\text{Likelihood of correct answer for subject } j \text{ on item } i.} = \frac{\lambda_i}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

Discriminability of item i

Ability of subject j

Difficulty of item i



What IRT Yields

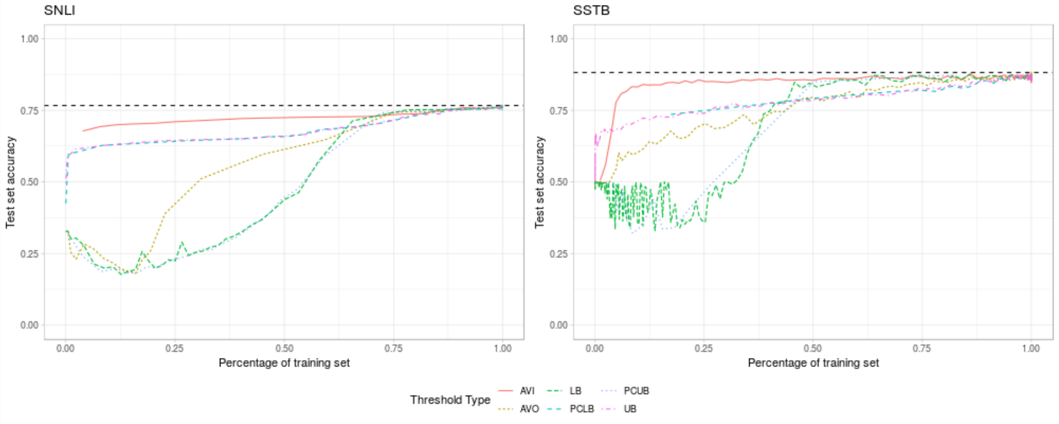
Given the previous information, IRT will yield estimates for chosen parameters, i.e.: item difficulty β_i, discriminability γ_i, and ability θ_j.

Consider two scenarios:

- What if the dataset is the training data?
- What if the dataset is a test set?

Improving Model Training

Data set filtering



- AVI: $|b_i| < \tau$
- UB: $b_i < \tau$
- PCUB: $pc_i < \tau$
- AVO: $|b_i| > \tau$
- LB: $b_i > \tau$
- PCLB: $pc_i > \tau$

Source: Lalor et al. (2019)

Biggest Differences

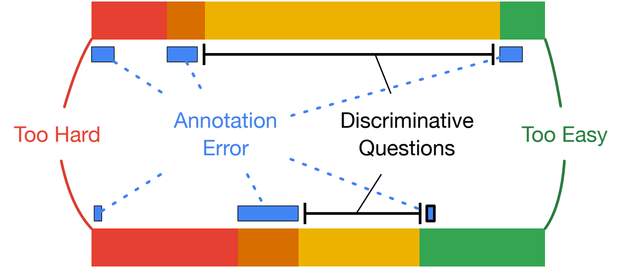
Task	Label	Item Text	Difficulty ranking		
			Humans	LSTM	NSE
SNLI	Con.	*P:* Two dogs playing in snow. *H:* A cat sleeps on floor	168	1	5
	Ent.	*P:* A girl in a newspaper hat with a bow is unwrapping an item. *H:* The girl is going to find out what is under the wrapping paper.	55	172	176
SSTB	Pos.	Only two words will tell you what you know when deciding to see it: Anthony. Hopkins.	9	103	110
	Neg.	...are of course stultifyingly contrived and too stylized by half. Still, it gets the job done–a sleepy afternoon rental.	128	46	41

Source: Lalor et al. (2019)

Finding Annotation Error

IRT Applications: Finding Annotation Error

Test examples can be: too hard, discriminative, too easy, or erroneous [1]



How can we use IRT to identify each example type?

[1] Boyd-Graber and Börschinger (2020)

IRT Applications: Finding Annotation Error

What makes examples bad?

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong
- Example: Correctness is a coinflip

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong
- Example: Correctness is a coinflip
- Non-Example: Difficult example few models get correct

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong
- Example: Correctness is a coinflip
- Non-Example: Difficult example few models get correct
- What parameter could identify this?

IRT Applications: Finding Annotation Error

What makes examples bad?

- Examples that do not discriminate between good and bad subjects
- Example: Bad label $\rightarrow$ all models get wrong
- Example: Correctness is a coinflip
- Non-Example: Difficult example few models get correct
- What parameter could identify this?
- We can use IRT discriminability γ_i to find bad examples!

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/Skill $\sim U(-4, 4)$

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$
- Items have a 5% of being invalid

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$
- Items have a 5% of being invalid
- Responses for valid items: $r_{ij} = sigmoid(\theta_j - \beta_i) > u, u \sim U(0, 1)$

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$
- Items have a 5% of being invalid
- Responses for valid items: $r_{ij} = sigmoid(\theta_j - \beta_i) > u, u \sim U(0, 1)$
- Responses for invalid items: $r_{ij} = u > .5, u \sim U(0, 1)$

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$
- Items have a 5% of being invalid
- Responses for valid items: $r_{ij} = sigmoid(\theta_j - \beta_i) > u, u \sim U(0, 1)$
- Responses for invalid items: $r_{ij} = u > .5, u \sim U(0, 1)$

IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:
- 10 Subjects, Ability/Skill $\sim U(-4, 4)$
- 1000 Items, Difficulty $\sim U(-4, 4)$
- Items have a 5% of being invalid
- Responses for valid items: $r_{ij} = sigmoid(\theta_j - \beta_i) > u, u \sim U(0, 1)$
- Responses for invalid items: $r_{ij} = u > .5, u \sim U(0, 1)$

Then, train a 3PL IRT model with py-irt

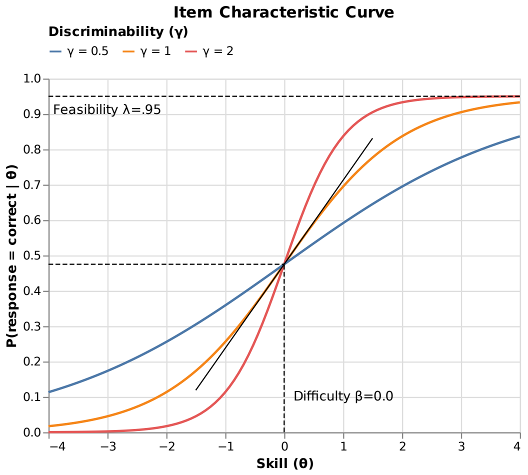
IRT Applications: 3PL Model

$$\underbrace{p(y_{ij} = 1 \mid \gamma_i, \beta_i, \lambda_i, \theta_j)}_{\text{Likelihood of correct answer for subject } j \text{ on item } i.} = \frac{\lambda_i}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

Discriminability of item i

Ability of subject j

Difficulty of item i



IRT Applications: Setup for Finding Annotation Error

IRT Parameters

- Item Difficulty: $\beta_i \sim$ Normal
- Item Discriminability: $\gamma_i \sim$ LogNormal
- Subject Ability $\theta_j \sim$ Normal

IRT Model

$$p(r_{ij} = 1 | \beta_i, \gamma_i, \theta_j) = \frac{1}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

IRT Applications: Setup for Finding Annotation Error

IRT Parameters

- Item Difficulty: $\beta_i \sim$ Normal
- Item Discriminability: $\gamma_i \sim$ LogNormal
- Subject Ability $\theta_j \sim$ Normal

IRT Model

$$p(r_{ij} = 1 | \beta_i, \gamma_i, \theta_j) = \frac{1}{1 + e^{-\gamma_i(\theta_j - \beta_i)}}$$

Note:

- Why $\gamma_i \sim$ LogNormal? Following Vania et al. (2021), forces γ_i to be non-negative.
- Other variables are zero centered.

IRT Applications: Sample Code for Finding Errors

Sample Code

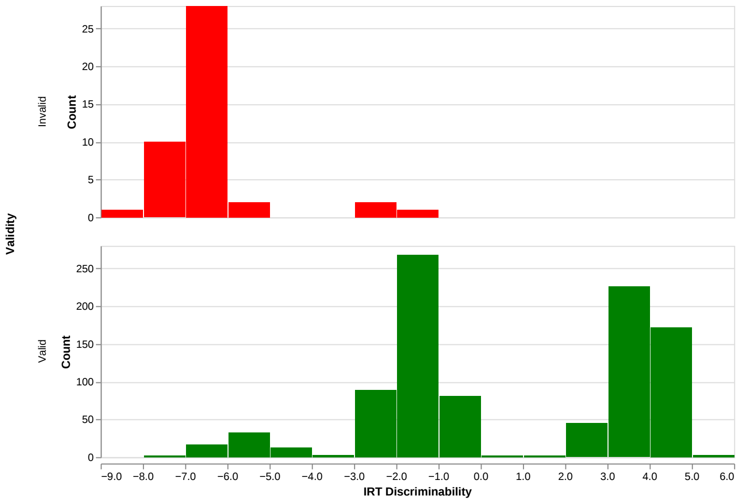
```
dataset = Dataset.from_jsonlines("/tmp/irt_dataset.jsonlines")
config = IrtConfig(
  model_type='tutorial', log_every=500, dropout=.2
)
trainer = IrtModelTrainer(
  config=config, data_path=None, dataset=dataset
)
trainer.train(epochs=5000, device='cuda')
```

IRT Applications: Simulation Results

Can we distinguish valid from invalid items based on discriminability γ_i?

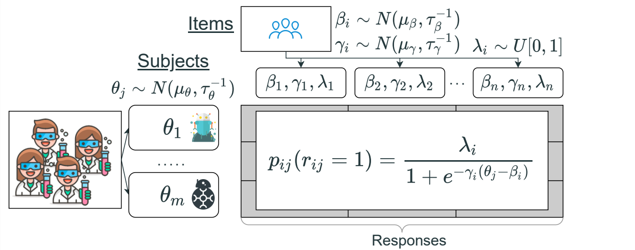
IRT Applications: Simulation Results

Can we distinguish valid from invalid items based on discriminability γ_i?



IRT Applications: Finding Annotation Error

In Rodriguez et al. (2021), we used a slightly different model to do this for SQuAD:

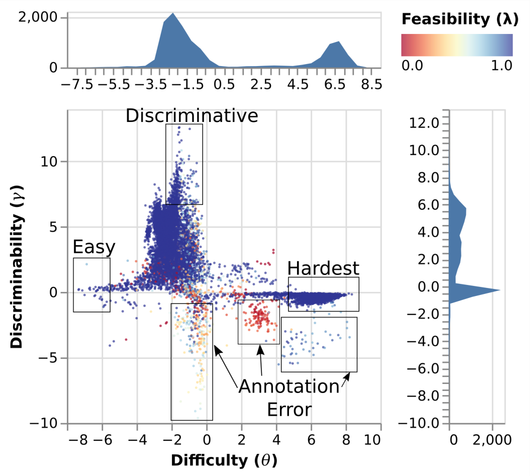


Differences

- Discriminability γ_i could be negative, which is inconvenient.
- Feasibility λ_i.

IRT Applications: Finding Annotation Error

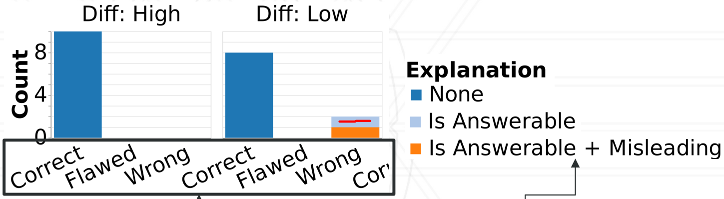
Plotting IRT parameters:



IRT Applications: Finding Annotation Error

Use IRT parameters to find partitions of data with annotation errors

Was the example correct?
- Question makes sense
- Answer is correct
- No ambiguity
- …

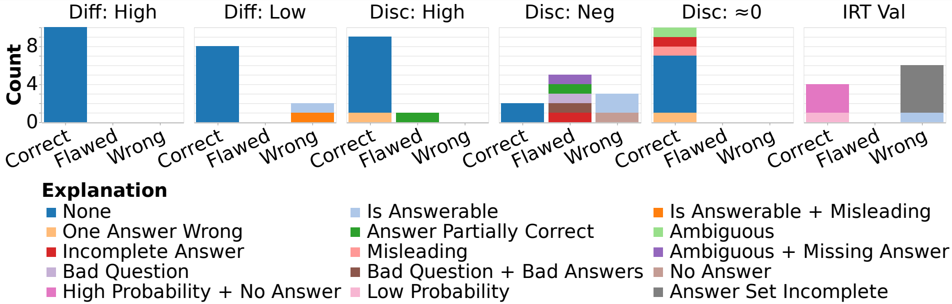


If the example is wrong, then why?
- It is *“Wrong/Flawed”* because it *“Explanation”*

Example:
One low difficulty questionwas wrong, because although the label says it is not answerable, it is answerable

IRT Applications: Finding Annotation Error

Use IRT parameters to find partitions of data with annotation errors



Things to note:

- Negative discriminability identifies errors

IRT Applications: Finding Annotation Error

Example of bad example identified by IRT

discriminability: -9.63 **Difficulty**: -0.479 **Feasibility**: 0.614 **Mean Exact Match**: 0.472
Wikipedia Page: Economic inequality **Question ID:** 572a1c943f37b319004786e3
Question: Why did the demand for rentals decrease?
Official Answer: demand for higher quality housing
Context: A number of researchers (David Rodda, Jacob Vigdor, and Janna Matlack), argue that a shortage of affordable housing – at least in the US – is caused in part by income inequality. David Rodda noted that from 1984 and 1991, the number of quality rental units decreased as the demand for higher quality housing increased (Rhoda 1994:148). Through gentrification of older neighbourhoods, for example, in East New York, rental prices increased rapidly as landlords found new residents willing to pay higher market rate for housing and left lower income families without rental units. The ad valorem property tax policy combined with rising prices made it difficult or impossible for low income residents to keep pace.

Evaluation Metrics

IRT Applications: Evaluation Metrics

Simple Idea: Instead of accuracy, use subject ability θ_j to rank.

IRT Applications: Evaluation Metrics

Simple Idea: Instead of accuracy, use subject ability θ_j to rank.

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before
- As before, 1,000 Test Examples

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before
- As before, 1,000 Test Examples
- A set of 800 easy examples $\sim U(-4, 0)$, Validity Rate 95%

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before
- As before, 1,000 Test Examples
- A set of 800 easy examples $\sim U(-4, 0)$, Validity Rate 95%
- A set of 150 moderate examples $\sim U(0, 3)$, Validity Rate 90%

IRT Applications: Evaluation Metrics Example

Suppose the following:

- 10 Subjects, similar setup as before
- As before, 1,000 Test Examples
- A set of 800 easy examples $\sim U(-4, 0)$, Validity Rate 95%
- A set of 150 moderate examples $\sim U(0, 3)$, Validity Rate 90%
- A set of 50 hard examples $\sim U(3, 4)$, Validity Rate 80%

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall
  - Easy subset

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall
  - Easy subset
  - Moderate subset

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
  - Overall
  - Easy subset
  - Moderate subset
  - Hard subset

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

In table we show:

- Subjects sorted by True Ability
- IRT Inferred Ability
- Accuracy:
    - Overall
    - Easy subset
    - Moderate subset
    - Hard subset
- What does the data show?

Ability		Accuracy			
True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100

IRT Applications: Evaluation Metrics Example

The data shows:

- Variation in true/inferred ability and accuracy by subset $\rightarrow$ Asking the right question matters!

True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100
-3.000	-7.61	0.256	0.301	0.066	0.100
-2.645	-4.88	0.325	0.380	0.093	0.140
-1.214	0.348	0.543	0.650	0.113	0.120
-1.156	1.40	0.560	0.667	0.120	**0.160**
-0.748	2.68	0.602	0.712	0.146	**0.200**
-0.455	3.36	0.631	0.746	0.193	0.100
0.232	5.76	0.729	0.848	0.293	0.120
2.16	11.1	**0.865**	0.956	**0.586**	**0.240**
2.50	14.2	**0.897**	0.971	**0.686**	**0.340**

IRT Applications: Evaluation Metrics Example

The data shows:

- Variation in true/inferred ability and accuracy by subset $\rightarrow$ Asking the right question matters!
- Fewer hard examples $\rightarrow$ noisier subset.

True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100
-3.000	-7.61	0.256	0.301	0.066	0.100
-2.645	-4.88	0.325	0.380	0.093	0.140
-1.214	0.348	0.543	0.650	0.113	0.120
-1.156	1.40	0.560	0.667	0.120	**0.160**
-0.748	2.68	0.602	0.712	0.146	**0.200**
-0.455	3.36	0.631	0.746	0.193	0.100
0.232	5.76	0.729	0.848	0.293	0.120
2.16	11.1	**0.865**	0.956	**0.586**	**0.240**
2.50	14.2	**0.897**	0.971	**0.686**	**0.340**

IRT Applications: Evaluation Metrics Example

The data shows:

- Variation in true/inferred ability and accuracy by subset $\rightarrow$ Asking the right question matters!
- Fewer hard examples $\rightarrow$ noisier subset.
- Accuracy difference between best two subjects is not large.

True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100
-3.000	-7.61	0.256	0.301	0.066	0.100
-2.645	-4.88	0.325	0.380	0.093	0.140
-1.214	0.348	0.543	0.650	0.113	0.120
-1.156	1.40	0.560	0.667	0.120	**0.160**
-0.748	2.68	0.602	0.712	0.146	**0.200**
-0.455	3.36	0.631	0.746	0.193	0.100
0.232	5.76	0.729	0.848	0.293	0.120
2.16	11.1	**0.865**	0.956	**0.586**	**0.240**
2.50	14.2	**0.897**	0.971	**0.686**	**0.340**

IRT Applications: Evaluation Metrics Example

The data shows:

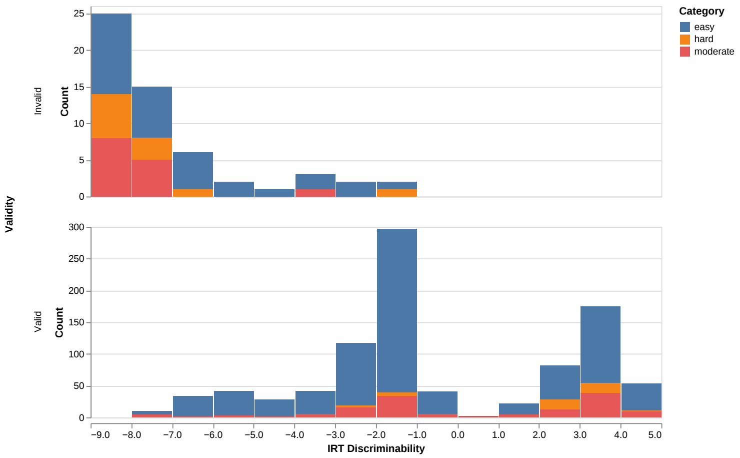
- Variation in true/inferred ability and accuracy by subset → Asking the right question matters!
- Fewer hard examples → noisier subset.
- Accuracy difference between best two subjects is not large.
- IRT is well suited to this type of data.

True	IRT	Overall	Easy	Mod	Hard
-3.506	-12.1	0.194	0.218	0.093	0.100
-3.000	-7.61	0.256	0.301	0.066	0.100
-2.645	-4.88	0.325	0.380	0.093	0.140
-1.214	0.348	0.543	0.650	0.113	0.120
-1.156	1.40	0.560	0.667	0.120	**0.160**
-0.748	2.68	0.602	0.712	0.146	**0.200**
-0.455	3.36	0.631	0.746	0.193	0.100
0.232	5.76	0.729	0.848	0.293	0.120
2.16	11.1	**0.865**	0.956	**0.586**	**0.240**
2.50	14.2	**0.897**	0.971	**0.686**	**0.340**

IRT Applications: Discounting Bad Examples

What do we see?

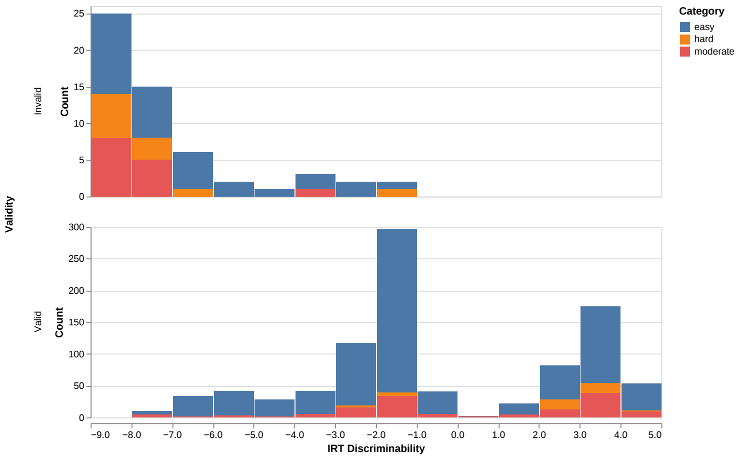
- Invalid examples sorted down



IRT Applications: Discounting Bad Examples

What do we see?

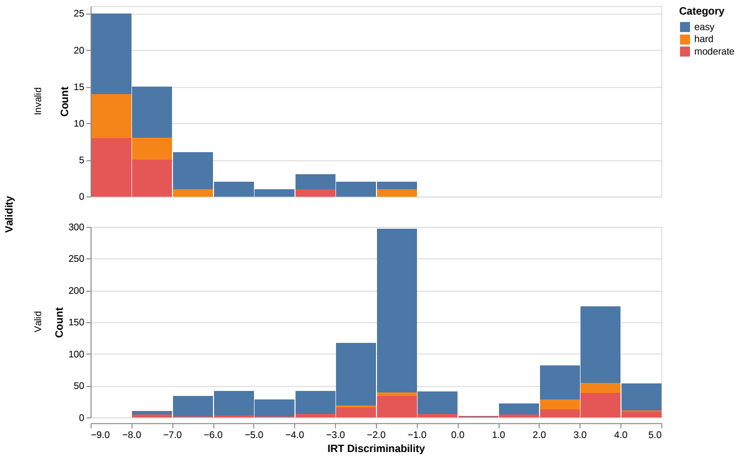
- Invalid examples sorted down
- Proportion of invalid examples represented



IRT Applications: Discounting Bad Examples

What do we see?

- Invalid examples sorted down
- Proportion of invalid examples represented
- Valid Hard examples are more discriminating



IRT Applications: Discounting Bad Examples

Why does this matter?

- Noisy examples $\rightarrow$ noisy metrics

IRT Applications: Discounting Bad Examples

Why does this matter?

- Noisy examples $\rightarrow$ noisy metrics
- Noise metrics $\rightarrow$ noisy rankings

IRT Applications: Discounting Bad Examples

Why does this matter?

- Noisy examples $\rightarrow$ noisy metrics
- Noise metrics $\rightarrow$ noisy rankings
- IRT is one way to mitigate the effect of noisy examples by directly modeling them!

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models
- We want to:

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models
- We want to:
    - Minimize annotation cost

IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models
- We want to:
  - Minimize annotation cost
  - Maximize correlation to ranking if fully annotate

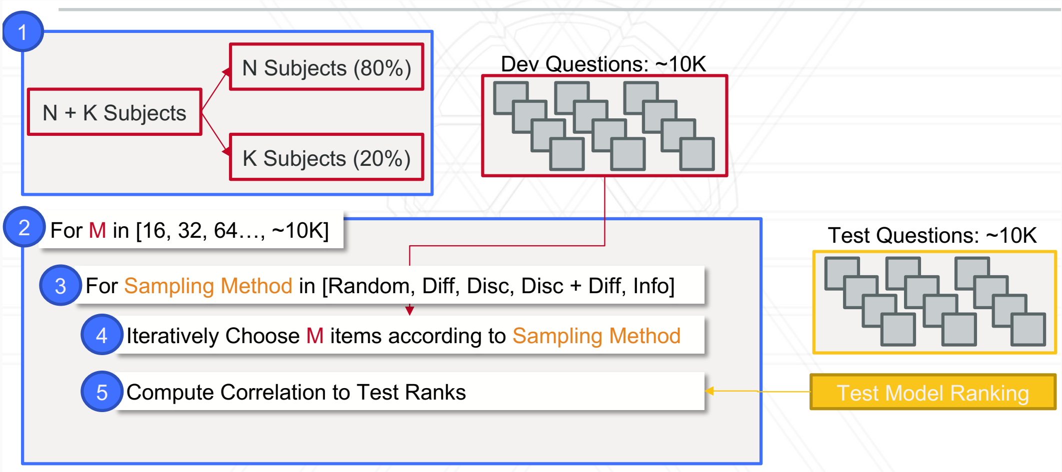
IRT Applications: Rank Reliability in Evaluation Metrics

In Rodriguez et al. (2021), we examined a case where:

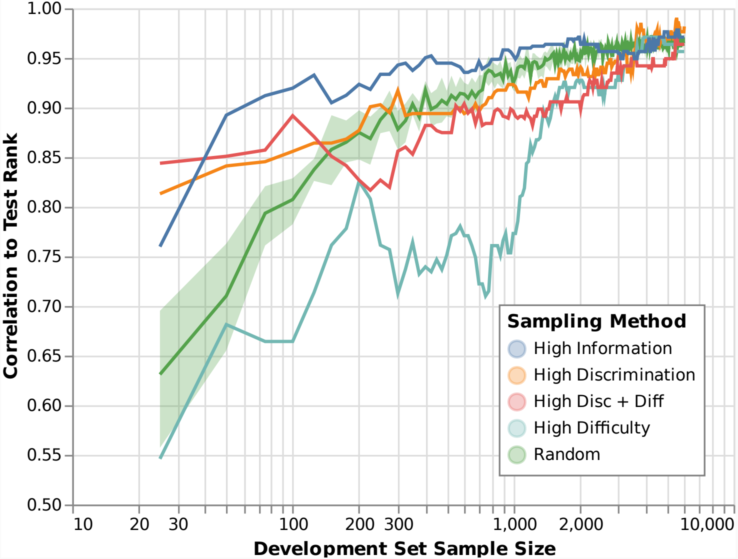
- The cost of annotation model responses is high.
- Pre-existing leaderboard data (i.e., response matrix).
- A new set of subjects/models
- We want to:
  - Minimize annotation cost
  - Maximize correlation to ranking if fully annotate
- Experiment: What method for selecting subset to annotate is best?

IRT Applications: Rank Reliability in Evaluation Metrics

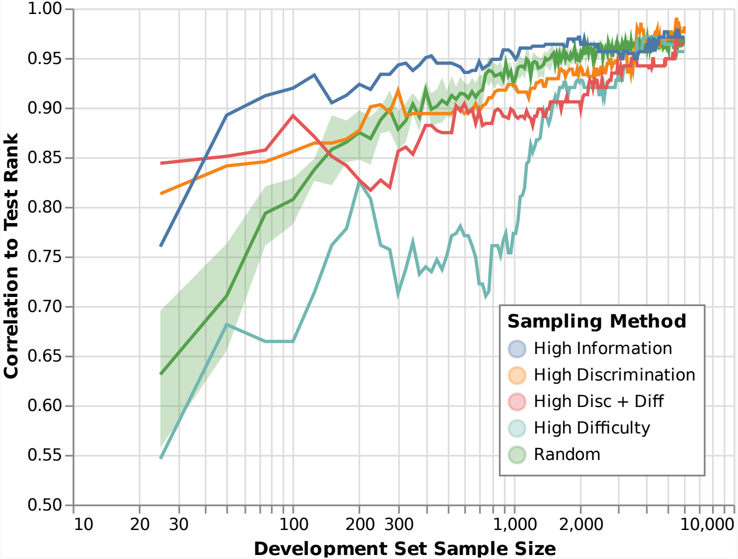
We test this setup with SQuAD leaderboard data:



IRT Applications: Rank Reliability in Evaluation Metrics



IRT Applications: Rank Reliability in Evaluation Metrics



Overall best method: pick item that maximizes Fisher information content, i.e.,

$$I_i(\theta_j) = \gamma_i^2 p_{ij}(1 - p_{ij})$$

$$Info(i) = \sum_j I_i(\theta_j)$$

Additional Work

- Adaptive Language-based Mental Health Assessment with Item-Response Theory (Varadarajan et al., 2023)
- Alternate Evaluation Metrics, e.g., Subject ability θ_j (Lalor et al., 2018)
- Anchor Points: Benchmarking Models with Much Fewer Examples (Vivek et al., 2024)
- tinyBenchmarks: evaluating LLMs with fewer examples (Polo et al., 2024)
- Comparing Test Sets with Item Response Theory (Vania et al., 2021)
- IRT for Efficient Human Evaluation of Chatbots (Sedoc and Ungar, 2020)

Break!

- Back in 15 minutes
- Next section: Advanced Topics

References

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, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716, Brussels, Belgium. 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.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples.

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.

João Sedoc and Lyle Ungar. 2020. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, Online. Association for Computational Linguistics.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing test sets with item response theory. 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 1141–1158, Online. Association for Computational Linguistics.

Vasudha Varadarajan, Sverker Sikström, Oscar NE Kjell, and H Andrew Schwartz. 2023. Adaptive language-based mental health assessment with item-response theory. *arXiv preprint arXiv:2311.06467*.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. 2024. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1576–1601, St. Julian's, Malta. Association for Computational Linguistics.