

# 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/>

## 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  $\beta_i$ , discriminability  $\gamma_i$ , and skill  $\theta_j$  might assume:

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

# What IRT Yields

Given the previous information, IRT will yield estimates for chosen parameters, i.e.: item difficulty  $\beta_i$ , discriminability  $\gamma_i$ , and skill  $\theta_j$ .

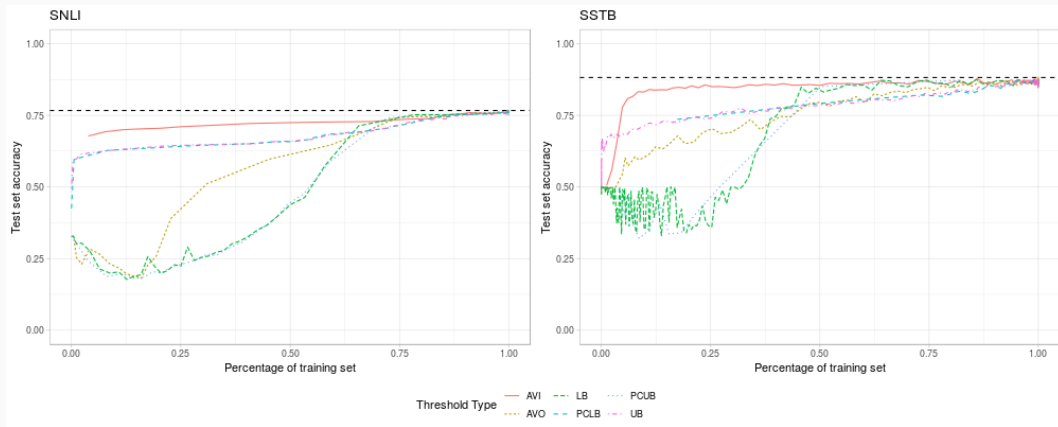
Consider two scenarios:

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

## IRT for 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:  $pc_i > \tau$

## MT-DNN Results

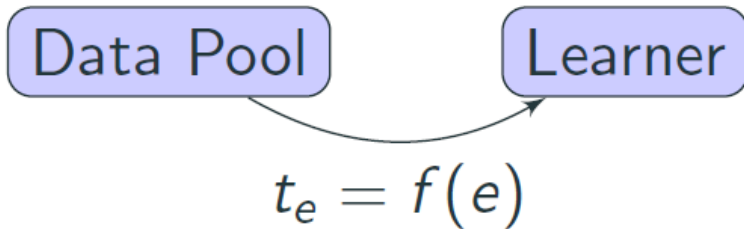
Strategy	% of Training Data		
	0.1%	1%	10%
Random (reported)	82.1	85.2	<b>88.4</b>
Random (small batch)	81.79	84.90	88.32
Lower-bound	43.68	41.56	39.89
Upper-bound	81.62	80.46	79.06
AVI	<b>82.44</b>	<b>85.44</b>	86.73
AVO	43.60	42.05	40.81

# Biggest Differences

Task	Label	Item Text	Difficulty ranking		
			Humans	LSTM	NSE
SNLI	Con.	<i>P</i> : Two dogs playing in snow. <i>H</i> : A cat sleeps on floor	168	1	5
	Ent.	<i>P</i> : A girl in a newspaper hat with a bow is unwrapping an item. <i>H</i> : 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

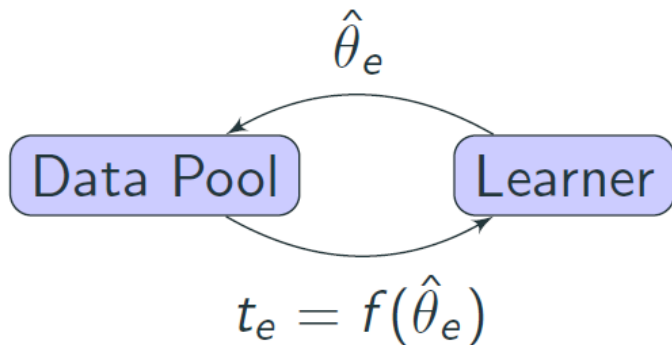


# Traditional Curriculum Learning



- Example difficulty based on heuristics
  - Replace heuristic with IRT difficulty
- Strategy is static
- Competence-based CL:  $t_e = f(e, c_0)$

Platanios, E., et al. "Competence-based Curriculum Learning for Neural Machine Translation." NAACL 2019.



- Example difficulty is learned
- Training set *dynamically selected* as a function of model ability

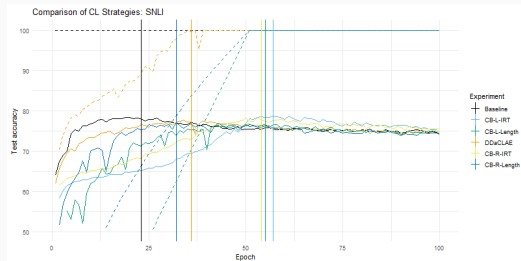
Gather responses from model  $j$  for items with known difficulties

$$\begin{aligned}Z_j &= \forall_{y \in \mathcal{Y}} \mathbf{I}[y_i = \hat{y}_i] \\L(\theta_j | Z_j) &= p(Z_j | \theta_j) \\\hat{\theta}_j &= \arg \max_{\theta_j} \prod_{i=1}^I p(z_{ij} = y_{ij} | \theta_j)\end{aligned}$$

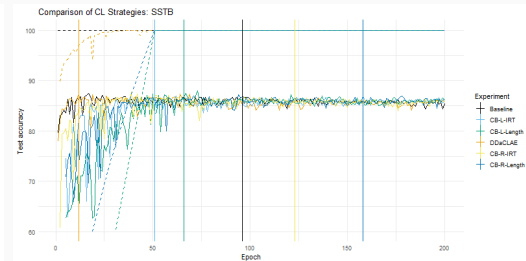
## Dynamic Data selection for Curriculum Learning via Ability Estimation

- At each epoch  $e$ :
  - Label all data:  $\hat{Y}$
  - Estimate  $\hat{\theta}_e$ :  $score(Y, \hat{Y}, B)$
  - Select training data:  $b_i \leq \hat{\theta}_e$

# Results



(a) SNLI



(b) SSTB

# Results

Metric	Experiment	MNIST	CIFAR	SSTB	SNLI
% $\Delta$ Train Size	Baseline	0	0	0	<b>0</b>
	DDaCLAE	<b>-9.37</b>	<b>-53.71</b>	<b>-88.68</b>	33.51
	CB Lin	-8.22	-21.56	-73.17	38.07
	CB Root	11.29	-22.63	10.23	60.08
% $\Delta$ Accuracy	Baseline	<b>0</b>	0	0	0
	DDaCLAE	-0.17	<b>0.66</b>	<b>0.45</b>	-1.08
	CB Lin	-0.01	-0.90	-0.18	<b>0.69</b>
	CB Root	-0.06	0.13	-0.38	-0.37

# Results

Label	Review	$\Delta_d$
Pos	Heart	67342
Pos	The year's greatest adventure, and Jackson's limited but enthusiastic adaptation has made literature literal without killing its soul – a feat any thinking person is bound to appreciate.	67334
Pos	Hip	67332
Neg	Exit	67346
Neg	There's an admirable rigor to Jimmy's relentless anger, and to the script's refusal of a happy ending, but as those monologues stretch on and on, you realize there's no place for this story to go but down.	67330

# Results

Label	Premise	Hypothesis	$\Delta_d$
Con.	Two men in a jogging race on a black top street, one man wearing a black top and pants and the other is dressed as a nun with bright red tennis shoes, while onlookers stand in a grassy area and watch from behind a waist high metal railing.	There is no metal railing.	549179
Ent.	Two dogs in the water.	They are swimming	549180
Neut.	Male musicians are playing a gig with one on the drums and the other on the guitar, with a backdrop of purple graphics apart of the light show.	Male musicians with long hair are playing a gig with one on the drums and the other on the guitar, with a backdrop of purple graphics apart of the light show.	549184
Neut.	A dog in a lake.	A dog is swimming.	549183



- Correlation between parameters between human and machine IRT models
- Downstream effectiveness of difficulty
- Qualitative check of learned parameters
- What about  $\theta$ ?

# IRT Applications in NLP Evaluations

---

- Find Bad Evaluation Items (Rodriguez et al., 2021)
- Alternate Evaluation Metrics, e.g., Subject skill  $\theta_j$  (Lalor et al., 2018; Rodriguez et al., 2021)
- Estimate Longevity of Tasks (Vania et al., 2021)

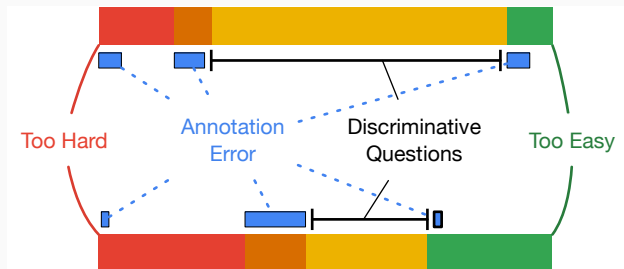


## **IRT Applications: Finding Annotation Error**

---

# IRT Applications: Finding Annotation Error

Test examples can be: too hard, discriminative, too easy, or erroneous <sup>1</sup>



How can we use IRT to identify each example type?

---

<sup>1</sup>Boyd-Graber and Börschinger (2020)

What makes examples bad?

What makes examples bad?

- Examples that do not discriminate between good and bad subjects



## What makes examples bad?

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

## 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

## 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 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?

## 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  $\gamma_i$  to find bad examples!

## IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

- Run a simulation where:

Then, train a 3PL IRT model with

## IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

Then, train a 3PL IRT model with

## IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

Then, train a 3PL IRT model with



## IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

Then, train a 3PL IRT model with

## IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

Then, train a 3PL IRT model with

## IRT Applications: Setup for Finding Annotation Error

Can follow along in notebook! Setup/Assumptions:

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

Then, train a 3PL IRT model with

# IRT Applications: Setup for Finding Annotation Error

## IRT Parameters

- Item Difficulty:  $\beta_i \sim \text{Normal}$
- Item Discriminability:  $\gamma_i \sim \text{LogNormal}$
- Subject Skill  $\theta_j \sim \text{Normal}$

## 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:  $\beta_i \sim \text{Normal}$
- Item Discriminability:  $\gamma_i \sim \text{LogNormal}$
- Subject Skill  $\theta_j \sim \text{Normal}$

## IRT Model

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

## Note:

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

## IRT Applications: Sample Code for Finding Errors

### Sample Code

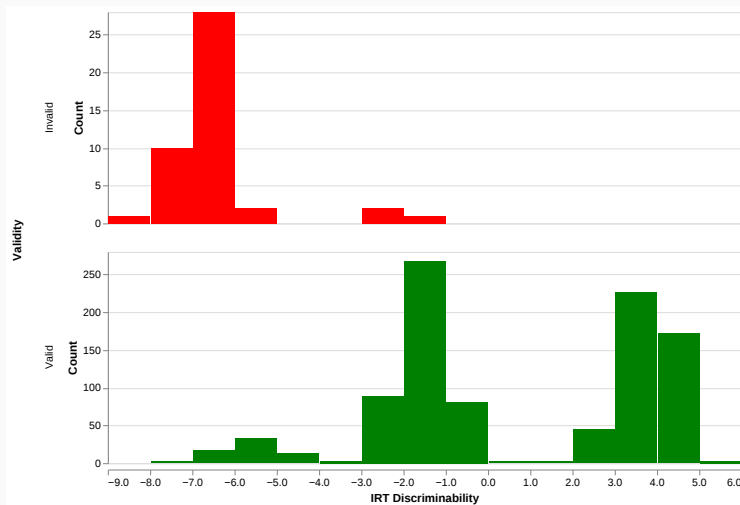
```
dataset = Dataset.from_jsonlines("/tmp/irt_dataset.jsonlines")
config = IrtConfig(
    model_type='tutorial', log_every=500, dropout=.2
)
trainer = IrtModelTrainer(
    config=config, data_path=None, dataset=dataset
)
trainer.train(epochs=5000, device='cuda')
```

## IRT Applications: Simulation Results

Can we distinguish valid from invalid items based on discriminability  $\gamma_i$ ?

# IRT Applications: Simulation Results

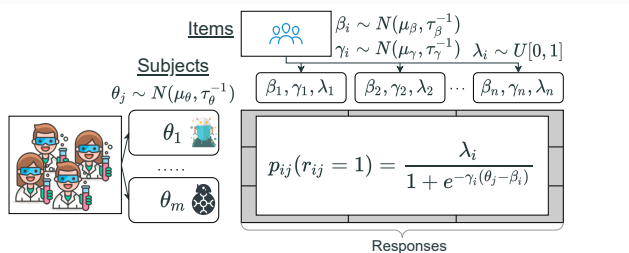
Can we distinguish valid from invalid items based on discriminability  $\gamma_i$ ?





# IRT Applications: Finding Annotation Error

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

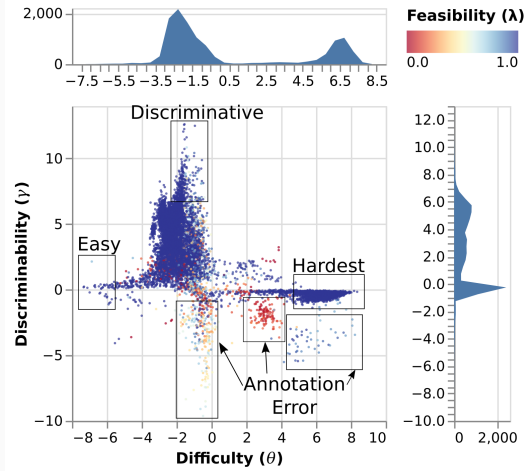


## Differences

- Discriminability  $\gamma_i$  could be negative, which is inconvenient
- Feasibility  $\lambda_i$  more difficult to control

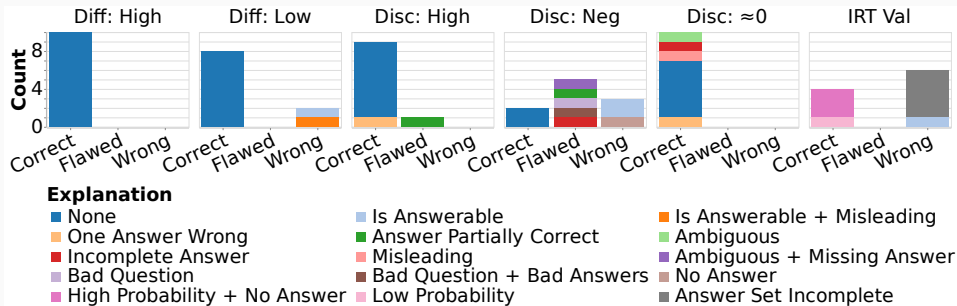
# IRT Applications: Finding Annotation Error

Plotting IRT parameters:



# IRT Applications: Finding Annotation Error

Use IRT parameters to find partitions of data with annotation errors



Things to note:

- Difficulty can be high or low, not an issue itself
- Negative discriminability identifies errors

## IRT Applications: Evaluation Metrics

---

Simple Idea: Instead of accuracy, use subject skill  $\theta_j$  to rank.

Simple Idea: Instead of accuracy, use subject skill  $\theta_j$  to rank.

What are the tradeoffs?

## IRT Applications: Evaluation Metrics Example

Suppose the following:

- As before, 1,000 Test Examples
- A set of 800 easy examples
- A set of 150 moderate examples
- A set of 50 hard examples
- 10 Subjects, similar setup as before

## IRT Applications: Evaluation Metrics Example

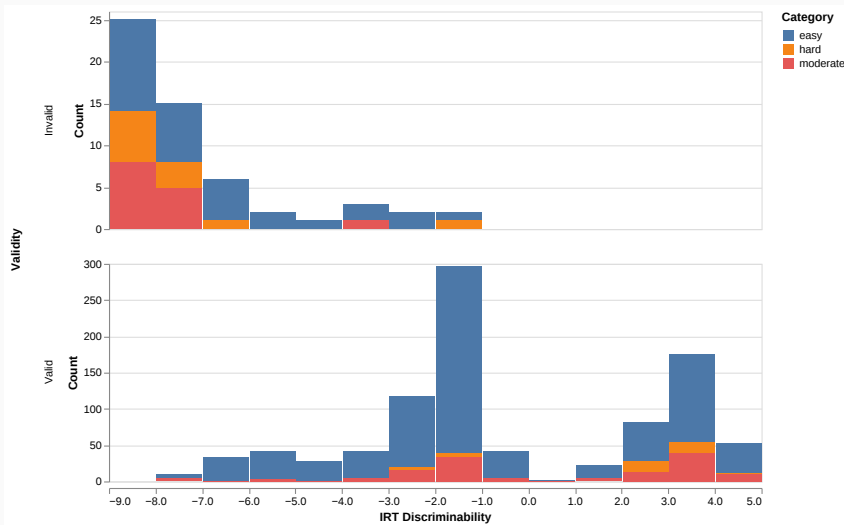
- Subjects sorted by True skill
- Accuracy gaps vary
- IRT can account for some of this variability

Skill			Accuracy		
True	IRT	Total	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

- Invalid examples sorted down
- Harder examples tend to be more discriminating



# IRT Applications: Rank Reliability in Evaluation Metrics

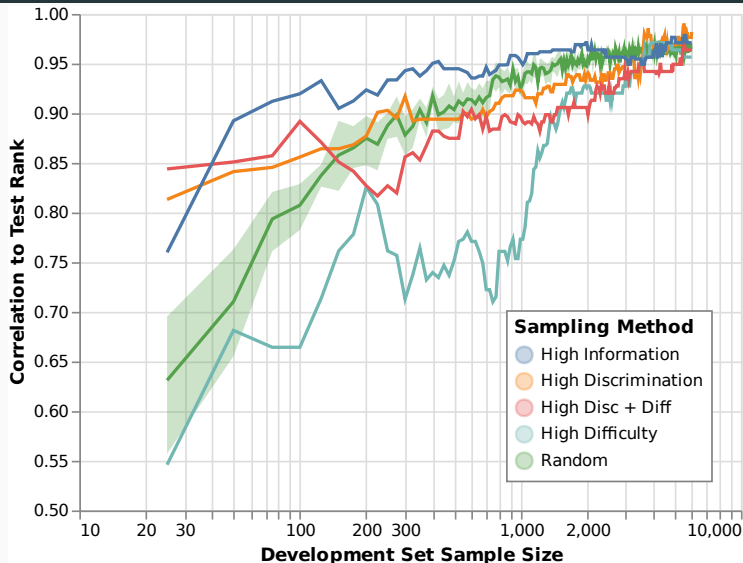
In Rodriguez et al. (2021), we examined a case like the SAT where we have:

- Pre-existing set of annotated responses for subjects/items
- Have a set of subjects (i.e., new models), same items.
- We want to minimize the number of subject responses to annotate, while maximizing the reliability of the resulting ranking.
- Baseline: Random sample
- IRT Methods: Sample based on different parameters

# IRT Applications: Rank Reliability in Evaluation Metrics

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

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



## IRT in NLP: Current and Future Work

---

# 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 *EMNLP*.
- 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 *ACL*.
- 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.