# DO MODELS SEE CORRUPTION AS WE SEE? AN ITEM RESPONSE THEORY BASED STUDY IN COMPUTER VISION

**Charchit Sharma, Ayan Kumar Pahari, Vineeth N Balasubramanian**
Indian Institute of Technology, Hyderabad, India
charchit.sharma@cse.iith.ac.in
{cs21mtech14003,vineethnb}@iith.ac.in

**Deepak Vijaykeerthy**
IBM Research, Bangalore, India
{deepakvij}@in.ibm.com

## ABSTRACT

On a given dataset, some models perform better than others. Can we examine this performance w.r.t. different strata of the dataset rather than just focusing on an aggregate metric (such as accuracy)? Given that noise and corruption are natural in real-world settings, can we study model failure under such scenarios? For a particular corruption type, do some classes become more difficult to classify than others? To answer such fine-grained questions, in this paper, we explore the use of Item Response Theory (IRT) in computer vision tasks to gain deeper insights into the behavior of models and datasets, especially under corruption. We show that incorporating IRT can provide instance-level understanding beyond what classical metrics (such as accuracy) can provide. Our findings highlight the ability of IRT to detect changes in the distribution of the dataset when it is perturbed through corruption, using latent parameters derived from IRT models. These latent parameters can effectively suggest annotation errors, informative images, and class-level information while highlighting the robustness of different models and dataset classes under consideration.

## 1 INTRODUCTION

In the past two decades, remarkable advancements have been made in the field of computer vision resulting in super-human performance on various well-known benchmarks such as Imagenet (Russakovsky et al., 2015), CIFAR10/CIFAR100 (Krizhevsky et al., 2009) & MNIST(LeCun et al., 1995) courtesy of deep learning models, especially with the advent of Vision Transformers (Liu et al., 2022), (Dosovitskiy et al., 2021). However, studies have shown that classical measures of progress like accuracy are not reliable as they overestimate the capabilities of the system (Sagawa et al., 2020; Bras et al., 2020; Menon et al., 2021; Shankar et al., 2021; Patel et al., 2021). One such study, (Shankar et al., 2021) showed that model accuracy drops due to the inability to generalize to slightly "harder" images than "easy" ones, which raises the question of whether these models are uniformly better across all instances. Thus, we argue that an instance-level analysis of evaluation data becomes necessary, especially in risk-sensitive applications.

In computer vision, researchers have tried to obtain global properties of images, including saliency, memorability, photo quality, and object importance (Liu et al., 2011; Russakovsky et al., 2015). Additionally, extrinsic anthropomorphic approaches have also been explored, such as estimating image difficulty based on the time required for human segmentation (Vijayanarasimhan & Grauman, 2009). However, in this work, we focus on assessing the difficulty of an image from the perspective of multiple models and also understand the change in the difficulty of images as they get corrupted at different severity levels. We study models and datasets from a multi-dimensional perspective. In particular, we draw upon the statistical framework of Item Response Theory (IRT) (Baker & Kim, 2004), which enables us to study model performance in relation to latent traits of

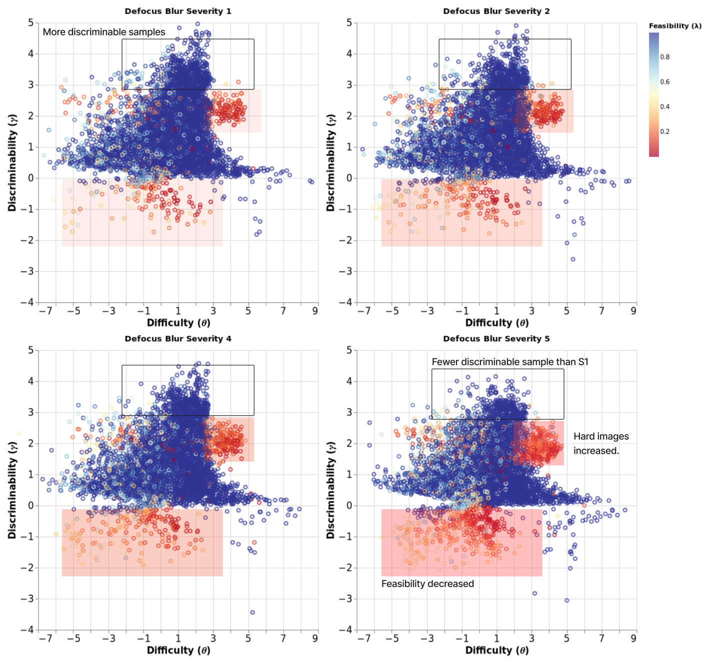


Figure 1: This figure shows how IRT captures instance level changes as we go from Severity 1(S1) to Severity 5(S5) of corruption in the case of Defocus Blur on the CIFAR10 dataset. As we move from S1-S5, discriminability, and feasibility decrease, suggesting that images are getting ambiguous at higher severity levels.

samples. IRT is widely accepted in educational assessment to evaluate test items (Birnbaum, 1968). The IRT framework learns latent variables corresponding to model performance, sample difficulty, and discriminability. The variable corresponding to model performance is referred to as the *ability score*. Based on the ability score, three measures are used to represent samples (questions in educational assessment, images in our context): *difficulty*, *discriminability*, and *feasibility*. *Difficulty* measures the item's hardness for the classification task. *Discriminability* measures the effectiveness of a sample in differentiating between similar models. *Feasibility* of a sample can be viewed as capturing a measure of inherent ambiguity in a sample (e.g., in educational assessment, a computer-based assessment that uses complex visual symbols may not be feasible for individuals with limited access to technology or with certain disabilities). We discuss specifics of the IRT framework used in our work in Section 2 and Figure 3.

Our key contributions and inferences in this work are summarized below:

- We estimate the latent parameters like difficulty in well-known computer vision datasets and models using classical Item Response Theory framework. We idenfity robust and sucecptible classes when perturbed to corruption in these datasets using the calculated latent parameters. For example, we find that **Airplane** is most **robust**, and **Frog** is most **susceptible** in case of the Defocus-Blur corruption in CIFAR10.

- We observe that the detection of samples with low discriminability can aid in uncovering $\approx 30-40\%$ of annotation errors across CIFAR10, CIFAR100, and Imagenet datasets.

- Our class-level analysis sheds light on the ease of distinguishability of each class. For example, in CIFAR10, we observe that **cat** and **dog** classes are the **easiest** to distinguish, whereas **ships** and **automobiles** are the **hardest**.

- Top-performing models are those that can classify a high number of difficult and discriminative samples. Unsurprisingly, we find through our studies that pretrained **Vision Transformers** are able to classify hard samples in general.

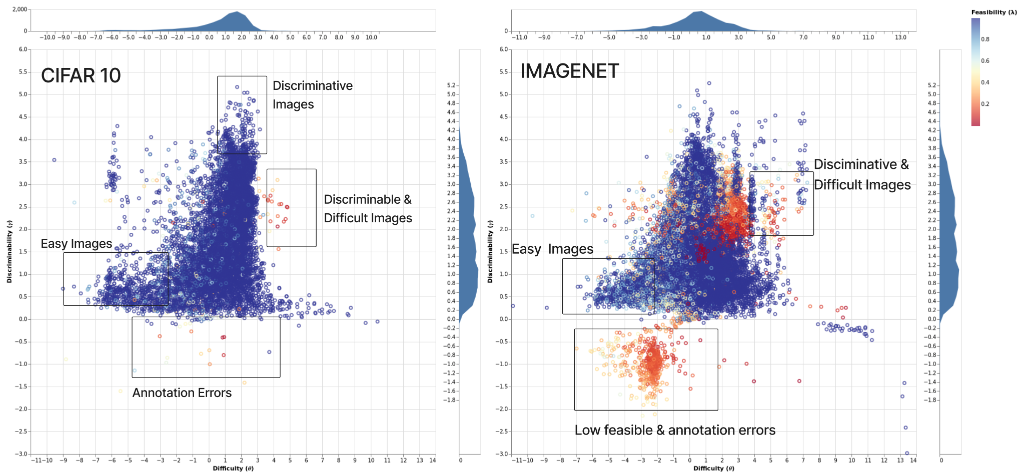


Figure 2: Discriminability vs. Difficulty Chart: Low feasibility and discriminability suggest annotation errors. 3.1

## 2 Item Response Theory

Item Response Theory (IRT) is a widely used psychometric methodology that provides a comprehensive framework for estimating the latent traits of both items (such as exam questions) and subjects (such as test-takers). Developed as a substitute to traditional summary statistics (Edgeworth, 1888), IRT has become a widely used tool in educational testing (Birnbaum, 1968). Furthermore, its applications have extended to machine learning, with recent studies exploring its use for model and data analysis (Martínez-Plumed et al., 2016) , (Martínez-Plumed et al., 2019), (Martínez-Plumed et al., 2016), (Vania et al., 2021), evaluating test sets (Lalor et al., 2016), (Lalor & Yu, 2020), (Rodriguez et al., 2021a).

In IRT, subjects are human test-takers, and items are assessment questions. For dichotomous data, a subject's answers to a set of items are evaluated as either correct or incorrect. Given n test items $T$ = $(T_1, T_2, ..., T_n)$ and m subjects $S$ = $(S_1, S_2, ...., S_m)$, we create a binary response matrix $Z^{n \times m}$ where each row in the matrix represents subject $m's$ responses to all the items in the dataset, ie $z_{ij} \in [0, 1]$, where $z_{ij}$ = $1$ indicates a correct response and $z_{ij}$ = $0$ indicates an incorrect response. Appendix Sec B talks about type of IRT models and how to estimate the latent parameters.

## 3 Experimental Results and Analysis

A key driving factor behind IRT is the observation that different deep learning models (equivalent to individuals with different abilities) exhibit different error patterns. E.g., some models perform extremely well on particular strata of the dataset, whereas some perform reasonably well on all strata. IRT framework considers this phenomenon (analogous to the bandwidth-fidelity trade-off (Liu et al., 2011) in educational assessment), enabling richer comparisons between models.

The following sections highlight the results and inferences of our study on different benchmark corrupted datasets - CIFAR10-C, and CIFAR100-C to provide a compelling use case for using Item Response Theory.

### 3.1 Identify Annotation Errors?

Low feasible and discriminable items indicate ambiguous images. Thus, we use this information to check for annotation errors in CIFAR10 and Imagenet datasets. (Northcutt et al., 2021) report annotation errors in CIFAR10 and Imagenet datasets, which we use as ground truth for validating our annotation error detection performance. We find that low discriminable and feasible items account for $\approx 30 - 40$ percent of annotation errors in the dataset. This shows that IRT can help identify and remove incorrect items and enhance the model's overall performance. 2

### 3.2 Analysis on Corrupted Dataset

| Corruption | Susceptible Classes | Robust Classes |
|---|---|---|
| Defocus Blur | Frog, Truck | Airplane, Horse |
| Glass Blur | Automobile, Frog | Horse, Cat |
| Motion Blur | Bird, Truck | Ship, Dog |
| Zoom Blur | Frog, Deer | Horse, Ship |
| Gaussian Blur | Truck, Dog | Horse, Airplane |
| Elastic Transform | Automobile, Deer | Airplane, Horse |
| JPEG Compression | Cat, Automobile | Airplane, Dog |
| Pixelate | Automobile, Deer | Horse, Airplane |
| Gaussian Noise | Cat, Ship | Automobile, Horse |
| Impulse Noise | Cat, Ship | Bird, Deer |

Table 1: Susceptible and robust classes for various corruptions.

We analyzed the effect of corruptions on CIFAR10-C, CIFAR100-C, and Imagenet-C (Hendrycks & Dietterich, 2019) datasets at the instance and class level by examining the image's difficulty and discriminability scores. We sorted the images based on these scores and divided them into four categories ("easy," "medium-easy," "medium-hard," and "hard"). The robustness of a class refers to the consistency of the score achieved across different levels of corruption. Thus, we use the standard deviation of the score computed across five severity levels of corruption. We perform this operation for all of the above categories ("easy," "medium-easy," "medium-hard," and "hard") and compute the average of these standard deviation values. This becomes the robustness score for a particular class. We use this score to find the most robust (high robustness score) and most susceptible classes (low robustness score) w.r.t. a given corruption. For e.g., in the case of blur corruption on CIFAR10, 'Frog' and 'Truck' were the most susceptible, while 'Horse' and 'Ship' were the most robust. We report more such interesting results in table 1 for ten different corruptions.

The above-proposed robustness score is helpful for understanding which classes need more attention to make the model robust. For e.g., consider Fig 5, where we study the effect of brightness corruption on different classes. Here, we report the number of "easy" category images for each class for different severity levels. It can be observed that for some classes, the number of easy images decreases. For some, it remains the same, and for others, it increases. Thus, an ML practitioner could focus only on those classes that become difficult to classify as the severity of corruption increases.

### 3.3 Difficulty and Discriminability Analysis

Here we perform a class-level analysis on CIFAR10, CIFAR100, and Imagenet datasets. We used the difficulty $\beta$ and discriminability $\gamma$ parameters using the 4PL model as shown in Fig. 3 to perform this analysis. IRT helps us analyze the model by combining discriminability and difficulty parameters. To study this, we divided the difficult and discriminable images into four segments, i.e., Easy, Med-Easy, Med-Hard, and Hard. We checked how many images in these segments a model can classify. We observe that high-accuracy models can correctly classify more high-difficulty images. Table 3 shows ViT16, a pretrained model that classifies more hard samples. Figure 4 shows classes with low and high difficulty and discriminability for CIFAR10. We find that in the case of CIFAR10: cat, dog, and bird are low discriminable and difficult classes. Similarly, we can see ships, trucks, and automobiles as highly discriminable and difficult classes.

## 4 Conclusions

In this paper, we advocate using Item Response Theory (IRT) for evaluation in computer vision tasks instead of using the classical model accuracy. In particular, we used IRT to get a detailed and instance-level understanding of the behavior of models and datasets, particularly in the presence of noise and corruption - beyond what classical metrics such as accuracy can provide. We used the difficulty $\beta$ and discriminability $\gamma$ parameters of IRT to identify robust and susceptible classes w.r.t. different corruption types. We further exploit the feasibility $\lambda$ parameter to detect up to $40\%$ of the annotation errors across CIFAR10, CIFAR100, and Imagenet datasets. We also conducted a class-level analysis, which helped shed light on the ease of distinguishability for each class. Finally, we also observed that the top-performing models could classify many difficult and discriminative samples. In summary, our findings highlight the importance of considering the impact of noise and corruption on model performance and the benefits of using IRT to analyze and improve computer vision models. We believe by incorporating IRT into computer vision tasks, researchers and practitioners can gain a more nuanced understanding of model behavior and identify potential areas for improvement in data annotation and model robustness. We hope our work can inspire further research in this area and contribute to developing more robust and reliable computer vision systems.

## References


F.B. Baker and S.H. Kim. *Item Response Theory: Parameter Estimation Techniques, Second Edition*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2004. ISBN 9780824758257. URL https://books.google.co.in/books?id=y-Q_Q7pasJ0C.

Allan Birnbaum. Some latent trait models. *Statistical theories of mental test scores*, 1968.

R Darrell Bock and Murray Aitkin. Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, 46(4):443–459, 1981.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases, 2020. URL https://openreview.net/forum?id=H1g8p1BYvS.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Francis Ysidro Edgeworth. The statistics of examinations. *Journal of the Royal Statistical Society*, 51(3):599–635, 1888.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Learning in graphical models*, pp. 105–161, 1998.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

John P Lalor and Hong Yu. Dynamic data selection for curriculum learning via ability estimation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2020, pp. 545. NIH Public Access, 2020.

John P Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, pp. 648. NIH Public Access, 2016.

John P Lalor, Hao Wu, and Hong Yu. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2019, pp. 4240. NIH Public Access, 2019.

Yann LeCun, Lawrence D. Jackel, Léon Bottou, Corinna Cortes, John S. Denker, Harris Drucker, I. Ramadass Subramanian, Urs Muller, E. Sackinger, Patrice Y. Simard, and Vladimir Naumovich Vapnik. Learning algorithms for classification: A comparison on handwritten digit recognition. 1995.

Dingding Liu, Yingen Xiong, Kari Pulli, and Linda Shapiro. Estimating image segmentation difficulty. In Petra Perner (ed.), *Machine Learning and Data Mining in Pattern Recognition*, pp. 484–495, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. ISBN 978-3-642-23199-5.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, June 2022.

Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Making sense of item response theory in machine learning. In *ECAI 2016*, pp. 1140–1148. IOS Press, 2016.

Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. Item response theory in ai: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271(C):18–42, 2019. doi: 10.1016/j.artint.2018.09.004.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=37nvvqkCo5`.

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

Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL `https://openreview.net/forum?id=XccDXrDNLek`.

Kanil Patel, William H. Beluch, Bin Yang, Michael Pfeiffer, and Dan Zhang. Multi-class uncertainty calibration via mutual information maximization-based binning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=AICNpd8ke-m`.

Georg Rasch. Probabilistic models for some intelligence and attainment tests. *The SAGE Encyclopedia of Research Design*, 1981.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P Lalor, Robin Jia, and Jordan Boyd-Graber. 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)*, pp. 4486–4503, 2021a.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan L. Boyd-Graber. Evaluation examples are not equally informative: How should that change nlp leaderboards? In *Annual Meeting of the Association for Computational Linguistics*, 2021b.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/sagawa20a.html`.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9661–9669, October 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R Bowman. Comparing test sets with item response theory. *arXiv preprint arXiv:2106.00840*, 2021.

Sudheendra Vijayanarasimhan and Kristen Grauman. What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2262–2269, 2009.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

# Appendix

## A Acknowledgement

We thank John Laylor, Pedro Rodriguez and the anonymous reviewers for their valuable feedback.

## B IRT Models

We now briefly discuss existing models based on the Item Response Theory framework. IRT-Base is called a Rasch Rasch (1981) or a 1PL model in psychometry. The 1-PL estimates a latent ability parameter $\theta_j$ for subjects and a latent difficulty parameter $\beta_i$ for items. A high-performing subject will have a higher ability score. However, we can only be sure about the top-performing model being good if one keeps evaluating them on varying difficulty samples. Like our previous example, Model B should have been better in accuracy. IRT handles this with ability $\theta$ and difficulty $\beta$ scores. More challenging items have higher difficulty $\beta$. So, the more significant the gap between the subject skill $\theta_j$ and item difficulty $\beta_i$, the more likely the subjects will respond correctly.

More complex IRT models like 2PL 1, 3PL 2 and 4PL 3 introduce additional item-specific latent parameters.

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

$$p(y_{ij} = 1|\alpha_i, \beta_i, \gamma_i, \theta_j) = \gamma_i + \frac{1 - \gamma_i}{1 + e^{-\alpha_i(\theta_j - \beta_i)}} \tag{2}$$

$$p(y_{ij} = 1|\alpha_i, \beta_i, \lambda_i, \theta_j) = \frac{\lambda_i}{1 + e^{-\alpha_i(\theta_j - \beta_i)}} \tag{3}$$

Our study uses the 4PL Model 3, which introduces the discriminability parameter $\alpha_i$ and feasibility parameter $\lambda_i$. A discriminable item is challenging but can be answered by a strong subject. For example, if Model A's ability is higher than most items' difficulty, ($\theta_j$ - $\beta_i$) will be large. Discriminability multiplies this gap by $\gamma_i$, normalizing a strong model's ability. In the context of computer vision, images with low $\beta_i$ are generally easy to classify for the model. or the model is unclear about them. Feasibility captures the ambiguity of images; the 4PL model (Fig 3) imposes a cap on the probability of correctly answering an item if a significant portion of subjects has answered it incorrectly. This cap, represented by $\lambda_i$, ensures that the probability remains feasible despite many incorrect responses. Our study found that some low-feasibility images had annotation errors (Fig 2).

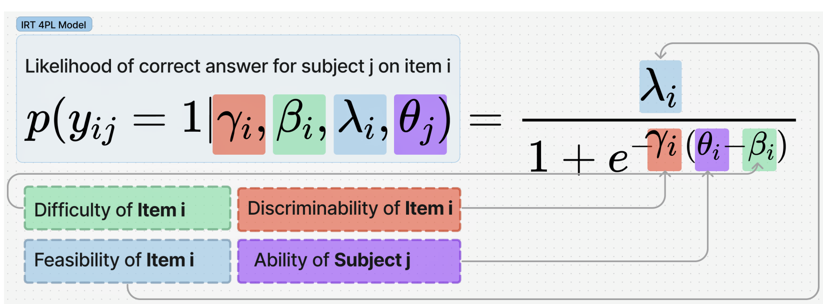


Figure 3: The likelihood of the correct response is modeled as a relationship between the difficulty $\beta_i$ of an item, its discriminability $\gamma_i$, feasibility $\lambda_i$, and the subject ability $\theta_i$.

### B.1 Estimating parameters using Variational Inference

We use mean field variational inference to infer IRT parameters from model responses. Natesan et al. (2016) shows that the variational inference Jordan et al. (1998) works better than traditional methods like maximum marginal likelihood estimation Bock & Aitkin (1981) to calculate the latent parameters when we scale to a larger dataset or a more significant number of subjects. Lalor et al. (2019) have shown the effectiveness of this method on NLP datasets like Bowman et al. (2015).

Given the data response matrix $Z^{n \times m}$ as discussed in the main section, we approximate the joint probability of the parameters $p(.)$ with a variational posterior $q_\phi(.)$ 4 which is a mean-field distribution. For each parameter, we choose the distribution mentioned in Eq. 5

$$q_\phi(\theta, \beta, \gamma, \mu, \sigma) = q(\mu)q(\sigma)\prod_{ij} q(\theta_j)q(\beta_i)q(\gamma_i) \tag{4}$$

$$\begin{aligned} \theta_j &\sim \mathcal{N}(\mu_\theta, \tau_\theta^{-1}) \\ \beta_i &\sim \mathcal{N}(\mu_\beta, \tau_\beta^{-1}) \\ \gamma_i &\sim \mathcal{N}(\mu_\gamma, \tau_\gamma^{-1}) \\ \lambda_i &\sim U[0,1] \end{aligned} \tag{5}$$

Means $\mu_\theta$, $\mu_\beta$, $\mu_\gamma$ are drawn from $\mathcal{N}(0, 10^6)$ and $\tau_\theta$, $\tau_\beta$, $\tau_\gamma$ from a $\Gamma(1,1)$ prior, as recommended by Natesan et al. (2016). Also, note that each latent $z$ is drawn from $\mathcal{N}(\mu_\theta, \tau_\theta^{-1})$ whose parameters are also latent variables. For these variables, we follow Lalor et al. (2019) Rodriguez et al. (2021b) and use the variational distributions $q(\mu_z) = \mathcal{N}(u_{\mu_z}, t_{\mu_z}^{-1})$ and $q(\tau_z) = \Gamma(a_{\tau_z}, b_{\tau_z})$. We then fit the posterior parameters by maximizing the evidence lower bound (ELBO).

### B.2 Reliability of IRT Models

The IRT (Item Response Theory) test stands apart from other tests in two fundamental ways. Firstly, it recognizes that certain items are more informative than others. Secondly, it assumes that each item's degree of informativeness depends on the subject's skill level (in our case, the model's performance).

To test for the reliability of the IRT models, we compute the Kendall rank correlation between the ability score ranking and the classical ranking of the models. We also evaluate how well the IRT models predict the classical model's responses on a held-out test set. In both cases, IRT models have a high correlation. Further, Table 2 shows that the IRT model achieves high correlation even under different types of corruption.

| **Corruption** | **Correlation** |
|---|---|
| Elastic Transform | 0.9709 |
| Frost | 0.9729 |
| Zoom Blur | 0.9698 |
| Motion Blur | 0.9776 |
| Brightness | 0.9729 |
| Defocus Blur | 0.9750 |
| Snow | 0.9735 |
| Fog | 0.9784 |
| Spatter | 0.9753 |

Table 2: Kendall rank correlation between model accuracy and ability score for multiple corruptions(at severity level 1) on the CIFAR10-C dataset.

| **Models** | **Easy** | **Med_Easy** | **Med_Hard** | **Hard** |
|---|---|---|---|---|
| Mnetv3_E1 | 392 | 281 | 173 | 157 |
| AlexNet | 1323 | 575 | 329 | 255 |
| Mnetv3_E2 | 246 | 206 | 246 | 302 |
| Mnetv3_E3 | 1028 | 633 | 378 | 335 |
| ViTb16_E2 | 2283 | 2431 | 2313 | 2144 |
| SwinBase_E2 | 2284 | 2428 | 2336 | 2169 |
| R101Best | 2285 | 2415 | 2252 | 1578 |
| R101_E2 | 2287 | 2400 | 2219 | 1384 |

Table 3: Multiple models classifying varying levels of difficult samples in case of CIFAR10 Brightness noise in no particular order. E1,E2,E3 denotes intermediate checkpoints in increasing order.)

## C EXPERIMENTS SETTINGS

This section talks about the models and datasets we considered for our study.

**Datasets:**

We experiment with CIFAR-10, CIFAR-100, and Imagenet-1K along with their corrupted counterparts, which include 19 types of algorithmically created corruptions from noise, blur, weather, and digital categories for CIFAR-10 and CIFAR-100, and 18 types of corruptions for Imagenet-1K.

Each type of corruption has five severity levels resulting in 95 distinct corruptions in CIFAR-10 and CIFAR-100 and 90 in Imagenet-1K. We focused primarily on testing the reliability of IRT on various easily accessible datasets and the robustness and susceptibility of models in the presence of corrupted datasets.
*Custom Test Splits* - Imagenet-1K and its corrupted counterpart with 18 noises do not have publicly available labeled test examples. For this, we create a new custom split by randomly sampling $50\%$ of the validation examples as a new test set and keeping the rest for validation (25 samples per class for test and 25 samples per class for validation, for a total of 1000 classes in Imagenet-1K, this comes as 25000 samples for validation and 25000 samples for test data).

**Models**: For our investigation, we used various models, ranging from CNN-based to pre-trained Transformer-based models. The use of some weaker CNN-based models was justified since utilizing solely high-performing models could result in a poor IRT model fit (Martínez-Plumed et al., 2019). We use 15 deep learning models: ViTb16, ViTb32 ((Dosovitskiy et al., 2021)), Swin-Base (Liu et al., 2021), Convnext-Base, Convnext-Tiny ((Liu et al., 2022)), EfficientnetV2-Small (Tan & Le, 2019), MobilenetV2 ((Sandler et al., 2018)), MobilenetV3-Large ((Howard et al., 2019)), Alexnet (Krizhevsky et al., 2012), VGG16, VGG19 ((Simonyan & Zisserman, 2014)), Densenet121 ((Huang et al., 2017)), Resnet-18, Resnet-50, Resnet-101 ((He et al., 2015)). For all the models, we evaluate five different checkpoints—at $1\%$, $10\%$, $30\%$, $50\%$, and $70\%$ of the maximum epochs as well as the best checkpoint on the validation set, which need not be one of the other five. This yields a total of 90 model predictions for each test example of CIFAR10 and CIFAR100.
However, with Imagenet-1K we included a few more pre-trained models' best checkpoints to achieve a satisfactory IRT model fit. We train the models using timm scripts (Wightman, 2019)

## D IRT SAMPLES AFTER INFERENCE

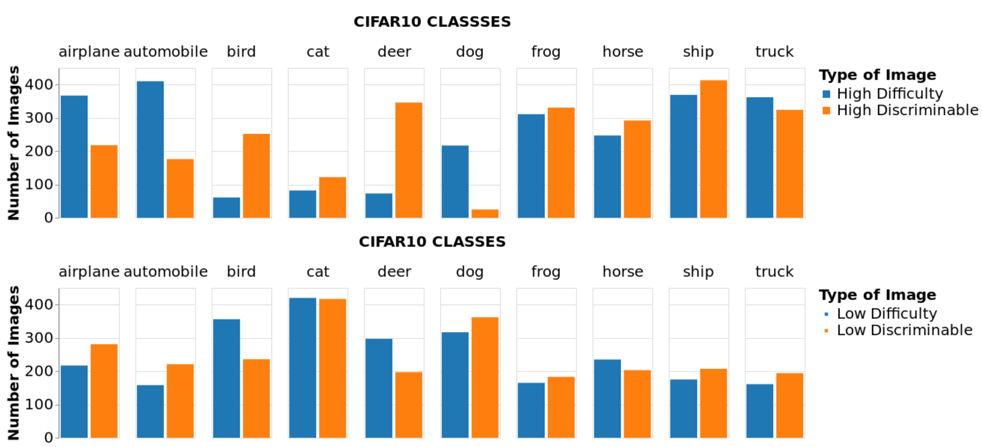


Figure 4: **Top Image**: This image shows a visualization of the top $25\%$ data in descending order of discriminable $\gamma$ and difficult $\beta$ samples using 4PL model in the CIFAR10 dataset. The classes were determined based on each image's discriminative $\gamma$ and difficulty $\beta$ scores. We can see that **cat** images are fewer and **ships** images are more in this segment.
**Bottom Image**: This image shows a visualization of the bottom $25\%$ data in descending order of discriminable $\gamma$ and difficult $\beta$ samples using 4PL model in the CIFAR10 dataset. The classes were determined based on each image's discriminative $\gamma$ and difficulty $\beta$ scores.In this quartile **cat** images are more, and **ships** are fewer.

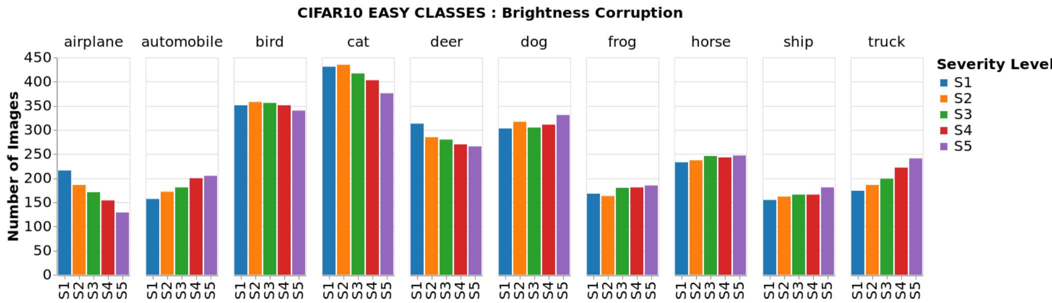


Figure 5: Exploring the Impact of Corruption on Easy Samples in CIFAR10-Brightness Corruption: Surprising Findings as Easy images persist in some classes only.

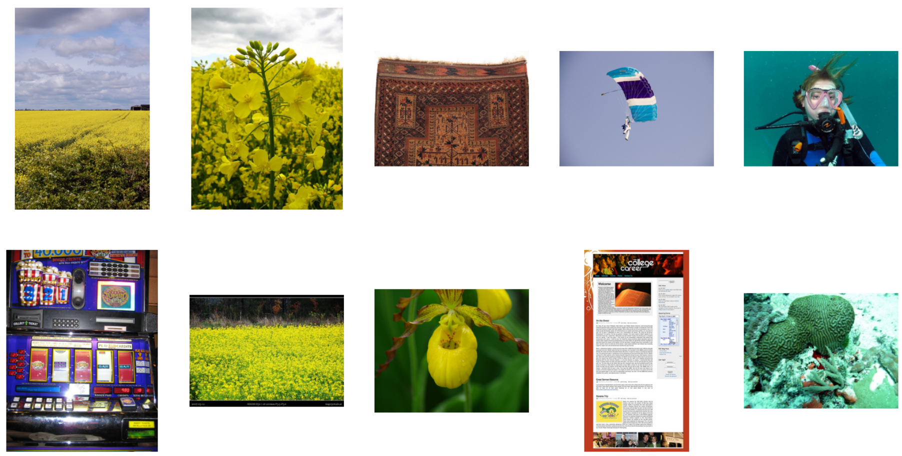

Figure 6: Hard images based on difficulty scores $\beta$ for the Imagenet dataset.

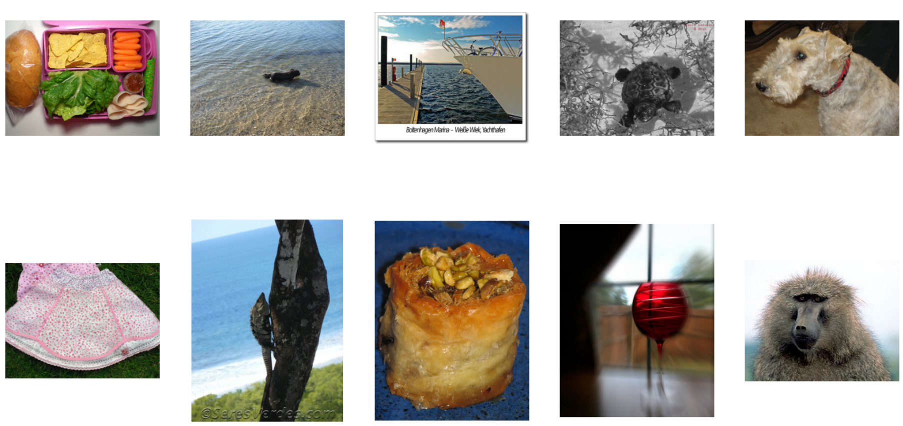

Figure 7: Easy images based on difficulty scores $\beta$ for the Imagenet dataset.

# E EFFECT OF CORRUPTION ON CIFAR10

This section gives more examples mentioned in Section 3.2

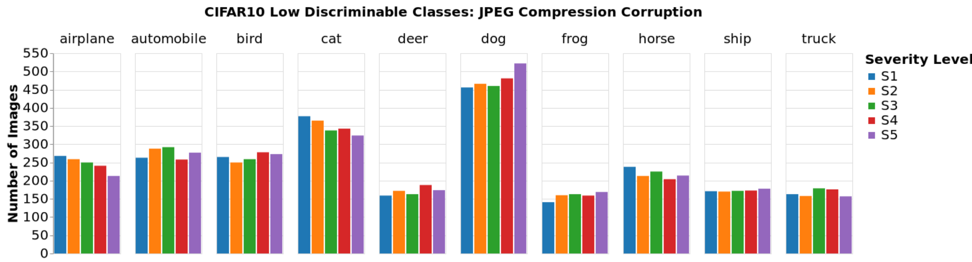


Figure 8: Low discriminable class-level samples changes in CIFAR10 dataset when exposed to the JPEG Corruption. For Eg. Dog class shows an increase in low discriminable samples, while airplane class shows a decrease in samples as we increase the severity level.

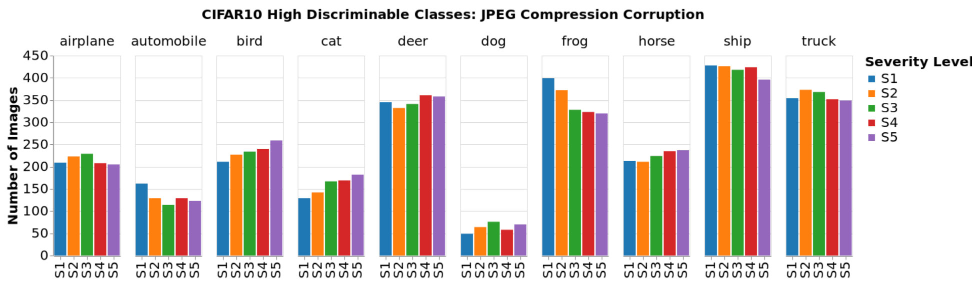


Figure 9: High discriminable class-level samples changes in CIFAR10 dataset when exposed to the JPEG Corruption. For eg. Bird class shows an increase in high discriminable samples, while frog class shows a decrease in samples as we increase the severity level.

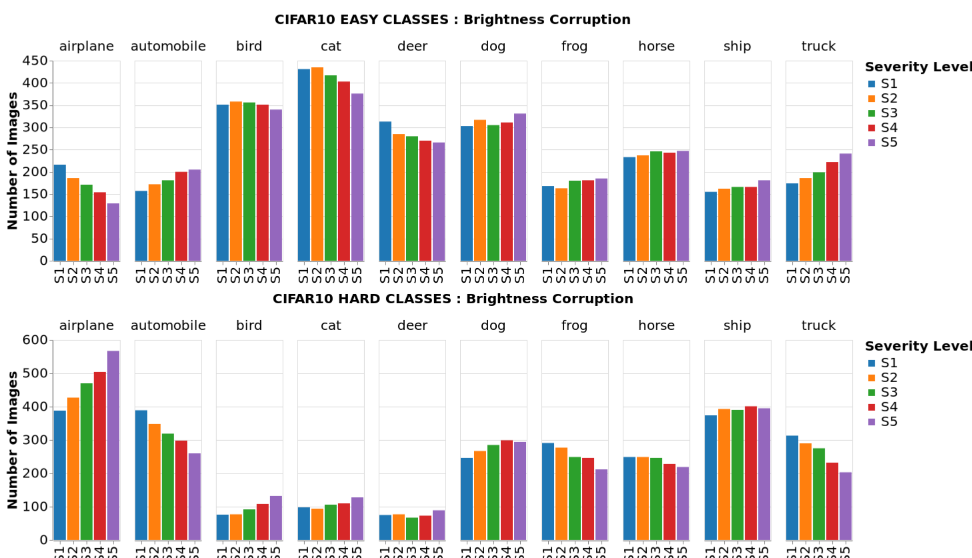


Figure 10: Class-level hardness changes in CIFAR10 when perturbed to Brightness Corruptions. **Top Image** : Easy classes changes, for eg. we can see airplane instances are getting hard to classify while truck is getting easy as we increase the severity level.
**Bottom Image**: Hard classes changes, for eg. airplance instances are getting hard, truck classes are getting easy. For class like **deer** we could see it can sustain effect of corruption.