

Ordinality in Discrete-level Question Difficulty Estimation: Introducing Balanced DRPS and OrderedLogitNN

Arthur Thuy^{1,2,*}, Ekaterina Loginova³ and Dries F. Benoit^{1,2}

¹Ghent University, Tweeckerkenstraat 2, 9000 Ghent, Belgium

²CVAMO Core Lab Flanders Make, Tweeckerkenstraat 2, 9000 Ghent, Belgium

³Dedalus Healthcare, Roderveldlaan 2, 2600 Antwerp, Belgium

Abstract

Recent years have seen growing interest in Question Difficulty Estimation (QDE) using natural language processing techniques. Question difficulty is often represented using discrete levels, framing the task as ordinal regression due to the inherent ordering from easiest to hardest. However, the literature has neglected the ordinal nature of the task, relying on classification or discretized regression models, with specialized ordinal regression methods remaining unexplored. Furthermore, evaluation metrics are tightly coupled to the modeling paradigm, hindering cross-study comparability. While some metrics fail to account for the ordinal structure of difficulty levels, none adequately address class imbalance, resulting in biased performance assessments. This study addresses these limitations by benchmarking three types of model outputs—discretized regression, classification, and ordinal regression—using the balanced Discrete Ranked Probability Score (DRPS), a novel metric that jointly captures ordinality and class imbalance. In addition to using popular ordinal regression methods, we propose OrderedLogitNN, extending the ordered logit model from econometrics to neural networks. We fine-tune BERT on the RACE++ and ARC datasets and find that OrderedLogitNN performs considerably better on complex tasks. The balanced DRPS offers a robust and fair evaluation metric for discrete-level QDE, providing a principled foundation for future research.

Keywords

Question Difficulty Estimation, Natural language processing, Ordinal regression, Ordered logit, Fine-tuning

1. Introduction

Question Difficulty Estimation (QDE), also known as question calibration, aims to predict a question's difficulty directly from its textual content and its answer options. This task plays a central role in personalized learning tools such as computerized adaptive testing [1] and dynamic online learning platforms, which aim to present questions aligned with a learner's proficiency level. Selecting questions that are either too easy or excessively difficult can reduce student motivation and hinder learning outcomes [2]. Reliable estimation of question difficulty is therefore essential.

Traditionally, QDE has relied on manual calibration [3] or pretesting [4], both of which are time-consuming and costly. To address these limitations, research has explored the use of natural language processing (NLP) techniques. These approaches train machine learning models to infer difficulty from the question text, allowing for rapid and scalable calibration of new questions without the need for manual intervention.

Difficulty levels in QDE are represented either as continuous scores or as discrete categories, with discrete levels being attractive for their ease of use. While continuous difficulty estimation is generally framed as a regression task, discrete-level QDE is essentially an ordinal regression problem, given the inherent ordering of difficulty levels from easiest to hardest.

However, the discrete-level QDE literature has neglected the ordinal nature of the task. Instead, existing work relies exclusively on classification and discretized regression methods, both of which are oversimplifications of the problem structure. Classification models disregard ordinal relationships

EvaLLAC'25: 2nd Workshop on Automatic Evaluation of Learning and Assessment Content, July 26, 2025, Palermo, Italy

*Corresponding author.

✉ arthur.thuy@ugent.be (A. Thuy); ekaterina.d.loginova@gmail.com (E. Loginova); dries.benoit@ugent.be (D. F. Benoit)

ORCID 0000-0001-9107-5646 (A. Thuy); 0000-0002-0952-9213 (E. Loginova); 0000-0002-9421-8566 (D. F. Benoit)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

Table 1

Related work on discrete-level QDE

Paper	Year	Difficulty	Model output format			Metric
			Regression	Classification	Ordinal	
[8]	2018	IRT	—	✓	—	(Adjacent) accuracy
[9]	2018	Manual	—	✓	—	Accuracy
[10]	2019	Manual	—	✓	—	Accuracy
[11]	2019	Manual, IRT	✓	—	—	(Adjacent) accuracy
[12]	2020	CTT	—	✓	—	Accuracy, F_1 -score
[13]	2021	Manual	—	✓	—	Accuracy
[14]	2023	Manual	✓	—	—	RMSE, R^2 , Spearman’s ρ
[15]	2024	Manual	✓	—	—	RMSE
Current	2025	Manual	✓	✓	✓	Balanced DRPS

altogether, and while discretized regression methods preserve ordinality, they implicitly assume equal spacing between levels—an assumption often violated in real-world data. As such, specialized ordinal regression techniques remain unexplored in this context. Moreover, no prior studies have systematically compared these competing approaches. Compounding the issue, studies typically only report performance metrics aligning with their chosen modeling paradigm, making cross-study comparisons difficult. The metrics also fail to account for class imbalances, a prevalent issue in this setting. As a result, there is no consensus on the most effective evaluation metric or modeling approach for discrete-level QDE.

This study addresses the literature gaps in discrete-level QDE by proposing the balanced Discrete Ranked Probability Score (DRPS), a novel evaluation metric that jointly captures ordinality and class imbalance. It also provides a direct way to compare deterministic predictions to probabilistic ones, which are especially valuable for downstream decision-making. We benchmark three types of model outputs—discretized regression, classification, and ordinal regression—using the balanced DRPS. Moreover, we propose a novel ordinal regression model OrderedLogitNN, extending the ordered logit model from econometrics to neural networks (NNs). Our work is the first to (i) introduce the balanced DRPS metric, (ii) compare classification and discretized regression models for QDE, and (iii) investigate specialized ordinal regression techniques, including the novel OrderedLogitNN. We conduct experiments by fine-tuning the Transformer model BERT on the RACE++ and ARC datasets.

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 3 introduces the balanced DRPS metric. Section 4 discusses the novel OrderedLogitNN model and section 5 outlines existing methods for ordinal tasks. Section 6 describes our experimental setup, and Section 7 presents the results and discussion. We conclude in Section 8. The source code is available on GitHub.¹

2. Related Work

Building on the survey by [5], we further investigate studies in QDE that utilize datasets with discrete difficulty levels. Question difficulty is defined using one of three main approaches: (i) Classical Test Theory (CTT) [6], (ii) Item Response Theory (IRT) [7], and (iii) manual calibration. In the case of manual calibration with expert annotators, difficulty is almost exclusively assigned in discrete levels due to its ease of use. While difficulty scores derived from CTT and IRT are continuous by nature, they are often discretized in practical applications to facilitate interpretation. Table 1 provides an overview of related work in discrete-level QDE.

¹<https://github.com/arthur-thuy/qde-ordinality>

2.1. Output types

Early work on discrete-level QDE employed classification models, such as support vector machines and Bayesian NNs [8, 9, 10]. Subsequently, [11] proposed a discretized regression approach using an LSTM-based NN, where difficulty was predicted as a continuous value between 0 and 1 and mapped to discrete intervals (e.g., [0.0; 0.2), [0.2; 0.4), etc.).

[12] introduced a multi-task BERT-based model that leverages shared representations across datasets, using a classification head to predict difficulty levels. To reduce the reliance on large labeled datasets in supervised methods, [13] proposed an unsupervised QDE method. They leverage the uncertainty in pre-trained question-answering models as a proxy for human-perceived difficulty, computed as the variance over the predictions from an ensemble of classification models.

[14] conducted a benchmarking study comparing traditional machine learning methods and end-to-end NNs across datasets with both discrete and continuous difficulty labels. Their results show that fine-tuned Transformer-based models such as BERT and DistilBERT consistently outperform classical methods. However, these models exclusively employed a discretized regression approach to address the ordinal nature of the labels.

More recently, [15] investigated active learning for QDE, demonstrating that comparable performance to fully supervised models can be achieved by labeling only a small fraction of the training data. Yet again, only a discretized regression modeling strategy was considered. As summarized in Table 1, prior work has not directly compared discretized regression and classification approaches, and specialized ordinal regression methods remain entirely unexplored.

2.2. Metrics

Accuracy is the most widely used evaluation metric for discrete-level QDE, particularly in studies employing classification models, as it aligns with the cross-entropy loss typically used during training. [11] is the only study using a discretized regression approach that also reports accuracy. However, accuracy fails to account for the ordinal structure of difficulty levels: all misclassifications are treated equally, regardless of their distance from the true label. Even if a prediction is incorrect, it should still be as close as possible to the true difficulty level. Additionally, accuracy is a threshold-based metric, relying solely on the final predicted label rather than the full output distribution—an issue that also applies to metrics such as the F_1 -score. To partially address this, [8] and [11] report adjacent accuracy, defined as the proportion of predictions within k levels of the true label. However, the choice of k is often arbitrary and dataset-dependent, complicating comparisons across studies.

The most recent works, which adopt discretized regression approaches, report RMSE as their primary evaluation metric, treating difficulty levels as integers. While RMSE reflects the ordinal structure to some extent, it assumes uniform spacing between levels—a condition rarely met in real-world data. For example, in primary school, there is a non-linear increase in difficulty over the years [16]. This limitation also applies to metrics such as R^2 and Spearman’s rank correlation. Furthermore, since RMSE is used as the loss function in discretized regression models, these models benefit from being evaluated on the same objective they were optimized for, giving them an unfair advantage and introducing a potential bias in comparative evaluations.

In addition to the ordinality aspect, these commonly used metrics fail to account for class imbalance, which is a prevalent issue in discrete-level QDE. In many datasets, mid-range difficulty levels tend to dominate, while questions at the extremes—those that are very easy or very difficult—are underrepresented [17, 18]. Standard metrics, which compute aggregate scores across all samples, inherently place greater weight on the majority classes. As a result, model performance on minority classes is often underrepresented, leading to inflated metrics that do not accurately reflect model effectiveness across the full difficulty spectrum. This is particularly problematic in educational contexts, where balanced performance across all difficulty levels is essential to ensure adequate personalized learning experiences.

As a result, the literature ignores the ordinal aspect in the modeling approaches and employs suboptimal evaluation metrics for discrete-level QDE. There is a clear need for an evaluation metric that

simultaneously accounts for ordinality and class imbalance, thereby enabling fair comparison across different modeling strategies.

3. Balanced Discrete Ranked Probability Score

The Continuous Ranked Probability Score (CRPS) is the most widely adopted scoring rule for evaluating probabilistic forecasts of real-valued variables, such as in precipitation forecasting [19]. It is defined as the integral of the squared difference (i.e., Brier score) between the cumulative distribution function (CDF) of a probabilistic forecast F and the CDF of the observed outcome, at all real-valued thresholds. The observed outcome is represented as a degenerate distribution, as its CDF is a step function. Formally, given a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$, the CRPS is computed as:

$$\text{CRPS}(F, y) = \frac{1}{N} \sum_{i=1}^N \int_{-\infty}^{\infty} (F(\hat{y}_i) - \mathbb{1}\{\hat{y}_i \geq y_i\})^2 d\hat{y}_i,$$

where $F(\hat{y}_i)$ denotes the CDF of the forecast and $\mathbb{1}\{\cdot\}$ is the step function.

The CRPS is distance-sensitive, meaning it rewards forecasts that assign higher probability mass to values near the true outcome. Specifically, when the forecast distribution concentrates probability density around the true value, the squared error between the forecast CDF and the observed step function is smaller across the integration range, resulting in a lower (better) score. In other words, even if the predicted value does not exactly coincide with the true outcome, placing substantial probability on neighboring values results in a better score than a prediction that is entirely off-target.

This property makes the CRPS particularly well-suited for ordinal prediction tasks. In such cases, the DRPS serves as a natural extension of the CRPS for discrete outcomes across K ordered categories. The DRPS has only been applied in meteorology [20] and has received little attention in the general field of ordinal regression. It is defined as:

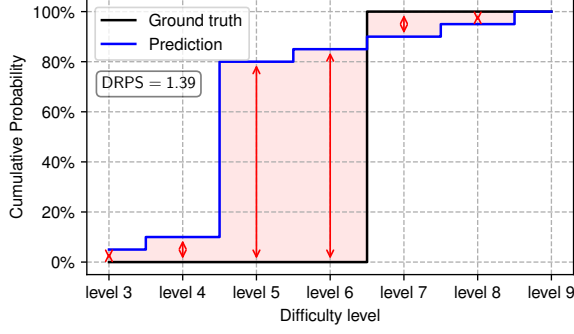
$$\text{DRPS}(F, y) = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{K-1} (F_k(\hat{y}_i) - \mathbb{1}\{k \geq y_i\})^2,$$

where $F_k(\hat{y}_i)$ denotes the predicted cumulative probability up to class k . The step function $\mathbb{1}\{\cdot\}$ moves from 0.0 to 1.0 at the position of the ground truth label.

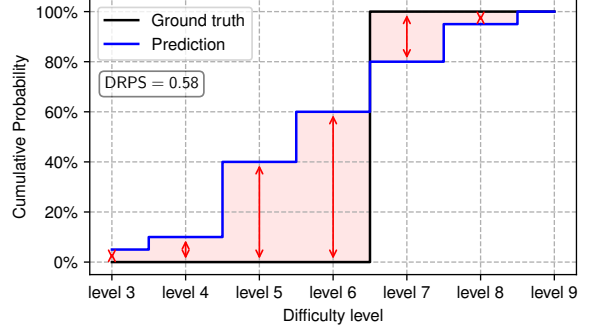
Unlike other evaluation metrics in related work, the DRPS operates on full probability distributions rather than point estimates, enabling a more nuanced assessment of ordinal predictions. Such a probability distribution over the levels is available for all classification and ordinal regression models, but not for the discretized regression model as it only outputs a predicted difficulty level. In the case of deterministic predictions, the output is treated as a degenerate distribution—analogueous to the representation of the observed outcome. Figures 1 and 2 illustrate how the DRPS is computed for single observations with probabilistic and deterministic model outputs.

Crucially, the DRPS respects the ordinal structure of the prediction task without assuming equal inter-class distances, a notable limitation of metrics such as RMSE. Additionally, when applied to deterministic predictions, the DRPS reduces to the mean absolute error, providing a direct way to compare deterministic and probabilistic predictions within a unified evaluation metric.

In this work, we introduce the balanced DRPS to address the class imbalance in discrete-level QDE datasets, where extreme difficulty levels are typically underrepresented compared to mid-range levels [17, 18]. The popular metrics (accuracy and RMSE) and standard DRPS compute unweighted averages, which overemphasize performance on majority classes and can produce misleadingly high scores on imbalanced data. However, for educational practitioners, robust performance across the full spectrum of difficulty levels—including the rarest—is essential. To ensure fair evaluation, the balanced DRPS weights each observation inversely proportional to the prevalence of its true class, $w_i = \frac{1}{\sum_{j=1}^N \mathbb{1}\{y_j = y_i\}}$,

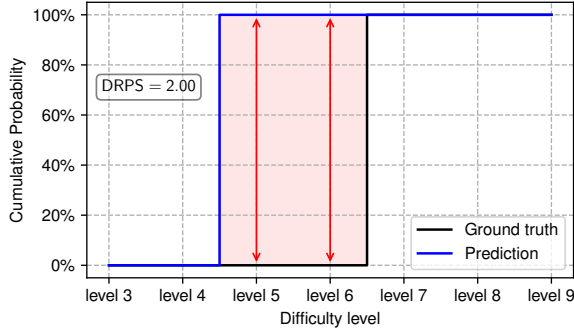


(a) Most density on “level 5”, confident

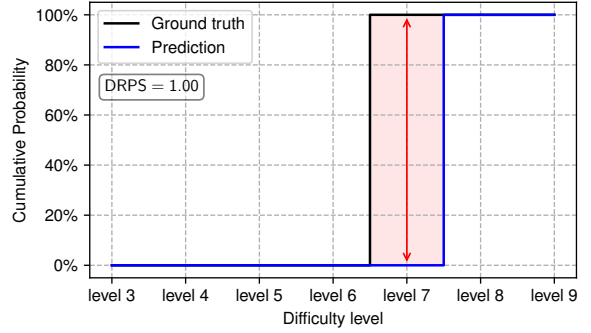


(b) Most density on “level 5”, uncertain

Figure 1: Example DRPS calculation with predicted probabilities. The ground truth for this observation is level 7 and both predictions have the most density on the incorrect level 5. (a) assigns high probability to level 5 while (b) is more uncertain and has more density on the neighboring levels, resulting in a better score.



(a) Prediction “level 5”



(b) Prediction “level 8”

Figure 2: Example DRPS calculation with a predicted label. The ground truth for this observation is level 7; (a) predicts level 5 while (b) predicts level 8.

thereby giving equal importance to all difficulty levels. Formally:

$$\text{Balanced DRPS}(F, y) = \frac{1}{N} \sum_{i=1}^N \sum_{k=1}^{K-1} w_i (F_k(\hat{y}_i) - \mathbb{1}\{k \geq y_i\})^2.$$

Thus for balanced datasets, the balanced DRPS is equivalent to the standard DRPS.

In conclusion, the balanced DRPS offers a robust and fair evaluation metric for discrete-level QDE by accounting for both ordinal structure and class imbalance. It supports both deterministic and probabilistic predictions and remains neutral to training objectives, making it well-suited for benchmarking across diverse modeling approaches.

4. Ordered Logit for NNs

OrderedLogitNN extends the classical ordered logit model to NNs, effectively bridging the gap between econometrics and deep learning. At its core, it is a latent variable model, where each observation is associated with an unobserved continuous utility value y_i^* [21] modeled as: $y_i^* = \mathbf{x}_i \beta + \epsilon_i$. The observed ordinal outcome y_i is derived from y_i^* through a censoring mechanism, whereby the continuous latent variable is mapped to one of K discrete categories based on a sequence of $K + 1$ increasing threshold values $\{\mu_{-1}, \mu_0, \mu_1, \dots, \mu_{K-1}\}$.

To identify the model parameters, several normalizations are required. First, the thresholds must be increasing $\mu_k > \mu_{k-1}$ to ensure valid (i.e., positive) probabilities. Second, the endpoints of the support

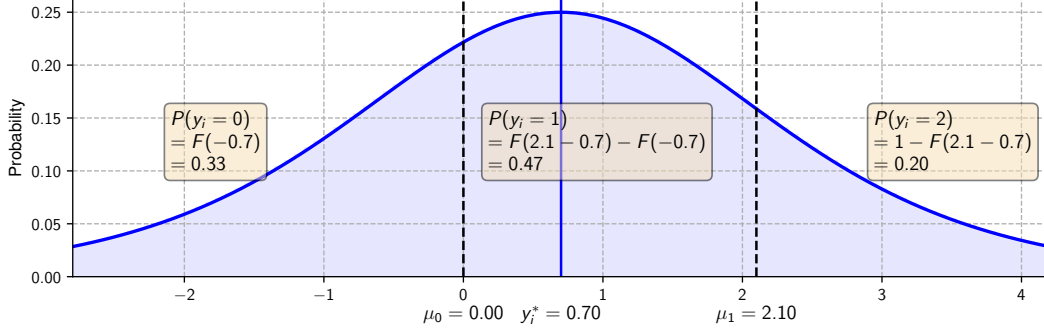


Figure 3: Ordered logit example for three ordinal categories.

are fixed as $\mu_{-1} = -\infty$ and $\mu_{K-1} = +\infty$, covering the entire real line. Third, the error term ϵ_i is assumed to follow a standardized logistic distribution (mean zero, variance $\frac{\pi^2}{3}$). The logistic distribution is preferred over the Gaussian (i.e., probit) for computational convenience, as the derivative has a closed form solution and is readily available as the sigmoid function. Finally, since \mathbf{x}_i includes a bias term, the threshold $\mu_0 = 0$.

The model defines the class probabilities as:

$$P(y_i = k | \mathbf{x}_i) = F(\mu_k - \mathbf{x}_i\beta) - F(\mu_{k-1} - \mathbf{x}_i\beta),$$

with F the cdf of the logistic distribution. An example for $K = 3$ is shown in Figure 3.

The NN is trained by minimizing the negative log-likelihood (NLL):

$$\text{NLL} = \sum_{i=1}^N \sum_{j=0}^{K-1} m_{ij} \log [F(\mu_k - \mathbf{x}_i\beta) - F(\mu_{k-1} - \mathbf{x}_i\beta)],$$

where $m_{ij} = 1$ if $y_i = k$ and 0 otherwise. The thresholds are reparameterized to ensure monotonically increasing values: $\mu_k = \mu_{k-1} + \exp(\delta_k) = \sum_{m=1}^k \exp(\delta_m)$.

The δ_k parameters are initialized such that the ordinal levels have equal probability mass under the logistic distribution, with the first threshold set to zero. The bias term is initialized to lie at the center of this distribution, while all other weights follow standard initialization practices in PyTorch. To facilitate convergence, the learning rates for the δ_k values and the bias term are scaled to be 100 times larger than those of the remaining network parameters.

Importantly, OrderedLogitNN is architecture-agnostic and can be integrated into any NN. Additionally, it makes no assumptions about the distances between ordinal levels, allowing it to flexibly model a wide range of ordered regression problems.

5. Existing approaches for ordinal regression with NNs

This section describes existing approaches to handle ordinal regression using NNs. Depending on the specific approach, the amount of ordinal information used and the underlying assumptions vary. This study uses three existing specialized ordinal regression methods that follow the extended binary classification framework, most widely used in the ordinal regression literature [22]. Note that the methods discussed below are not tied to any specific architecture and can be utilized with any NN.

Let $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$ be the training dataset consisting of N training examples. Here, $\mathbf{x}_i \in \mathcal{X}$ denotes the i^{th} training example and y_i the corresponding rank, where $y_i \in \mathcal{Y} = \{r_1, r_2, \dots, r_K\}$ with ordered rank $r_K \succ r_{K-1} \succ \dots \succ r_1$. The objective is to find a model that maps $\mathcal{X} \rightarrow \mathcal{Y}$. For example, the ARC dataset has $K = 7$ difficulty levels with an output space $\mathcal{Y} = \{\text{“grade 3”}, \text{“grade 4”}, \dots, \text{“grade 9”}\}$.

5.1. Discretized regression

In the regression approach, also referred to as discretized regression, the K rank indices are treated as numerical values to utilize the ordinal information (see Table 1 for references). The model f minimizes the mean squared error loss and predicts a real-valued quantity $f(\mathbf{x}_i) \in \mathbb{R}$ representing a continuous rank estimate, which is then converted to the closest rank index. For example, a regression estimate of 2.7 is converted to index 3 while estimate 5.2 is converted to index 5.

Using a discretized regression approach in an ordinal QDE problem assumes that the inter-level distances are equal. However, this condition is only rarely satisfied in practice [16]. On the ARC dataset with levels “grade 3” to “grade 9”, for example, such an approach assumes that the jump in difficulty among all grades is identical.

5.2. Classification

In the classification approach, the model’s output space is a set of K unordered labels, one for each rank (see Table 1 for references). The model is trained to minimize the cross-entropy loss and the predicted rank label is the class with the highest predicted probability. As such, the predicted rank label is $\hat{y}_i = \arg \max_{y_i \in \mathcal{Y}} p(y_i | \mathbf{x}_i)$.

This approach essentially assumes that the difficulty levels are completely independent, hence discarding the available ordinal information. For example, for a question in the ARC dataset with true level “grade 3”, predicting levels “grade 4” and “grade 5” incurs the same loss even though the difference between “grade 3” and “grade 5” is larger than the that between level “grade 3” and “grade 4”.

5.3. Ordinal: OR-NN

A popular general machine learning approach to ordinal regression is to cast it as an extended binary classification problem [23], leveraging the relative order among the labels. That is, the ordinal regression task with K ranks is represented as a series of $K - 1$ simpler binary classification sub-problems. For each rank index $k \in 1, 2, \dots, K - 1$, a binary classifier is trained according to whether the rank of a sample is larger than k . As such, all $K - 1$ tasks share the same intermediate layers but are assigned distinct weight parameters in the output layer. In summary, this framework relies on three steps: (i) extending rank labels to binary vectors, (ii) training binary classifiers on the extended labels, and (iii) computing the predicted rank label from the binary classifiers.

In 2016, the authors of [24] adapted this framework to train NNs for ordinal regression; we refer to this method as OR-NN. More formally, a rank label y_i is first extended into $K - 1$ binary vectors $y_i^{(1)}, \dots, y_i^{(K-1)}$ such that the $y_i^{(k)} \in \{0, 1\}$ indicates whether y_i exceeds rank r_k , for instance, $y_i^{(k)} = \mathbb{1}\{y_i > r_k\}$. Using the extended binary labels, a single NN is trained with $K - 1$ binary classifiers in the output layer to minimize the cross-entropy loss. Based on the binary task predictions, the predicted rank label is $\hat{y}_i = r_{q_i}$. The rank index q_i is given by

$$q_i = 1 + \sum_{k=1}^{K-1} \mathbb{1}\{P(y_i > r_k) > 0.5\},$$

where $P(y_i > r_k) \in [0, 1]$ is the predicted probability of the k th binary classifiers in the output layer.

However, the authors pointed out that OR-NN can suffer from rank inconsistencies among the binary tasks such that the predictions for individual binary tasks may disagree. For example, on the RACE++ dataset, it would be contradictory if the first binary task predicts that the difficulty is not higher than middle school level while the second binary task predicts it to be more difficult than high school level. This inconsistency could lead to suboptimal results when combining the $K - 1$ predictions to obtain the estimated difficulty level. Figure 4 provides an example of a rank consistent and inconsistent prediction. In response, two methods have been proposed that overcome this drawback of rank inconsistency: CORAL and CORN.

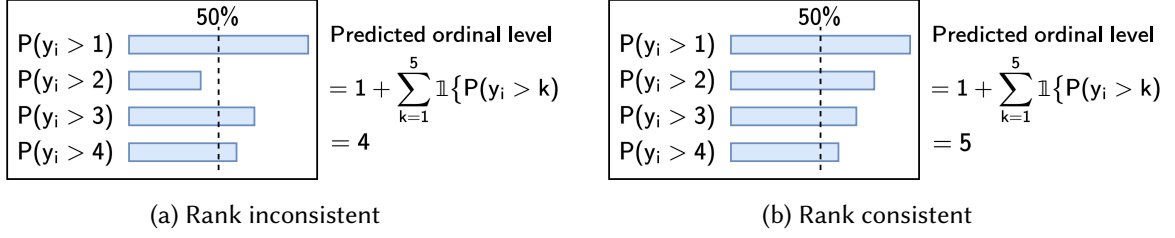


Figure 4: Example of a rank inconsistent and rank consistent prediction on a QDE task with $K = 5$ levels. Figure adapted from [22].

5.4. Ordinal: CORAL and CORN

CORAL [25] achieves rank consistency by imposing a weight-sharing constraint in the last layer. Instead of learning distinct weights between each unit in the penultimate layer and each output unit, CORAL enforces that the $K - 1$ binary tasks share the same weight parameters to the units in the penultimate layer. In addition, CORAL learns independent bias terms for each output unit as opposed to a single bias term for the output layer.

CORAL uses a cross-entropy loss over the $K - 1$ binary classifiers and the authors show theoretically that by minimizing this loss function, the learned bias terms of the output layer are non-increasing such that $b_1 \geq b_2 \geq \dots \geq b_{K-1}$. Consequently, the predicted probabilities of the $K - 1$ tasks are decreasing which ensures that the output reflects the ordinal information and is rank consistent. All other steps are identical to the extended binary classification framework.

However, while CORAL outperforms the OR-NN method in age prediction [25], the weight-sharing constraint may restrict the expressiveness and capacity of the NN.

More recently, the authors of [22] proposed CORN, which guarantees rank consistency without restricting the NN’s expressiveness and capacity. CORN achieves rank consistency by a novel training scheme which uses conditional training sets in order to obtain the unconditional rank probabilities.

More formally, CORN constructs conditional training subsets such that the output of the k th binary task $f_k(\mathbf{x}_i)$ represents the conditional probability $f_k(\mathbf{x}_i) = P(y_i > r_k | y_i > r_{k-1})$. For $k \geq 2$, the conditional subsets consist of observations where $y_i > r_{k-1}$. When $k = 1$, $f_1(\mathbf{x}_i)$ represents the initial unconditional probability $P(y_i > r_1)$ based on the complete dataset. The transformed unconditional probabilities can then be computed by applying the chain rule for probabilities: $P(y_i > r_k) = \prod_{j=1}^k f_j(\mathbf{x}_i)$.

Since $\forall j, 0 \leq f_j(\mathbf{x}_i) \leq 1$, we have $P(y_i > r_1) \geq P(y_i > r_2) \geq \dots \geq P(y_i > r_{K-1})$, which guarantees rank consistency among the $K - 1$ binary tasks. During model training, CORN minimizes the cross-entropy loss over the binary tasks. All other steps are identical to the extended binary classification framework.

6. Experiments

6.1. Data

We evaluate our models on two multiple-choice question (MCQ) datasets: RACE++ and ARC. These datasets vary in domain and granularity of difficulty levels.

RACE++ [17, 26] is a dataset of reading comprehension MCQs. Questions are labeled with one of three difficulty levels—1 (middle school), 2 (high school), and 3 (university level)—which we treat as ground truth labels for QDE. The label distribution is imbalanced: 25% of the questions are labeled as level 1, 62% as level 2, and 13% as level 3. The dataset is partitioned into training, validation, and test sets containing 100,568; 5599; and 5642 questions, respectively.

ARC [18] is a dataset of science MCQs across grades 3 through 9. Question difficulty is indicated by the target grade level (i.e., 7 levels), which we use as ground truth. The training, validation, and test

Table 2

Results on RACE++ (3 levels)

Output type	Bal. DRPS ↓	Bal. DRPS ↓ (degenerate)	RMSE ↓	Accuracy ↑
Random	0.893 ± 0.009	0.893 ± 0.009	1.024 ± 0.005	0.334 ± 0.003
Majority	0.667 ± 0.000	0.667 ± 0.000	0.616 ± 0.000	0.620 ± 0.000
Regression	0.167 ± 0.003	0.167 ± 0.003	0.391 ± 0.004	0.853 ± 0.003
Classification	0.133 ± 0.003	0.170 ± 0.004	0.402 ± 0.006	0.847 ± 0.004
OR-NN	0.131 ± 0.004	0.168 ± 0.004	0.400 ± 0.008	0.847 ± 0.006
CORAL	0.201 ± 0.003	0.185 ± 0.004	0.476 ± 0.018	0.782 ± 0.017
CORN	0.127 ± 0.003	0.164 ± 0.001	0.397 ± 0.004	0.851 ± 0.003
OrderedLogitNN	0.130 ± 0.001	0.162 ± 0.002	0.384 ± 0.005	0.861 ± 0.003

splits contain 3358, 862, and 3530 questions, respectively. The distribution is highly imbalanced: level 8 appears approximately 1400 times, level 5 about 700 times, and level 6 only 100 times. To decrease this imbalance, we follow [14] and downsample the two most frequent levels to 500 examples each, resulting in a partially balanced training set of 2293 questions.

6.2. Model Architecture

We focus on end-to-end Transformer-based NNs, as they have been shown to outperform traditional NLP approaches that rely on separate feature engineering and modeling stages [14]. We fine-tune the Transformer BERT (“bert-base-uncased”) on the task of QDE, stacking an output layer on top of the pre-trained language model. During fine-tuning, both the weights of the output head and the pre-trained model are updated. We follow the input encoding of [14] and concatenate the question and the text of all the possible answer choices in a single sentence, divided by separator tokens.

Additionally, we investigate the performance of two baselines which serve as a lower bound on performance: (i) Random and (ii) Majority. The Random baseline randomly predicts a difficulty level, while the Majority baseline consistently predicts the majority level in the training set.

The experiments are implemented in PyTorch [27] using the HuggingFace [28] package, and results are averaged over five independent runs with random seeds.

7. Results and Discussion

7.1. Balanced DRPS

Tables 2 and 3 present the results for the RACE++ and ARC datasets, which contain 3 and 7 difficulty levels, respectively. Balanced DRPS with probabilistic inputs is the main evaluation metric, as these predictions express uncertainty and are particularly valuable in downstream decision-making. In addition, we report the balanced DRPS computed using degenerate distributions—i.e., where all probability mass is placed entirely on the predicted level. This setup removes any representation of uncertainty from the predictions, thereby altering the scores for both classification and ordinal regression methods. Notably, the baselines and the discretized regression model remain unaffected in this setting, as they do not express uncertainty. Recall that for both metrics, lower is better. Furthermore, we include the commonly used but flawed metrics RMSE and accuracy.

On RACE++ with 3 levels (Table 2), the classification model performs comparably to the ordinal methods OR-NN, CORN, and OrderedLogitNN in terms of balanced DRPS. The discretized regression model, by contrast, shows slightly inferior performance. Interestingly, the CORAL model underperforms significantly, likely due to its weight-sharing constraint limiting the NN’s capacity. Nonetheless, all models substantially outperform the baseline approaches. We hypothesize that the small differences in performance across models are due to the limited number of ordinal levels.

Table 3
Results on ARC (7 levels)

Output type	Bal. DRPS ↓	Bal. DRPS ↓ (degenerate)	RMSE ↓	Accuracy ↑
Random	2.296 ± 0.017	2.296 ± 0.017	2.730 ± 0.006	0.144 ± 0.002
Majority	2.286 ± 0.000	2.286 ± 0.000	2.195 ± 0.000	0.409 ± 0.000
Regression	1.030 ± 0.015	1.030 ± 0.015	1.412 ± 0.013	0.356 ± 0.003
Classification	0.720 ± 0.003	1.038 ± 0.005	1.522 ± 0.013	0.421 ± 0.005
OR-NN	0.722 ± 0.005	1.005 ± 0.003	1.435 ± 0.009	0.400 ± 0.002
CORAL	0.963 ± 0.003	1.440 ± 0.020	2.083 ± 0.033	0.156 ± 0.005
CORN	0.725 ± 0.007	1.046 ± 0.012	1.428 ± 0.004	0.389 ± 0.008
OrderedLogitNN	0.674 ± 0.004	0.980 ± 0.007	1.468 ± 0.007	0.393 ± 0.006

The ARC dataset with 7 levels (Table 3) presents a more challenging setting, as it includes a greater number of difficulty levels and exhibits more pronounced class imbalance. Here, the OrderedLogitNN model considerably outperforms all other methods. For the remaining methods, the insights are consistent with those observed on RACE++.

When restricting the predictions to degenerate distributions, scores generally deteriorate (see the third column in Tables 2 and 3) because balanced DRPS is designed to reward well-calibrated probabilistic predictions while penalizing overconfident, incorrect ones. This shift brings the regression model’s performance in line with the classification model, OR-NN, and CORN. OrderedLogitNN again performs on par for RACE++ and performs substantially better on ARC.

When considering RMSE and accuracy—metrics that are poorly suited for ordinal prediction tasks (see Section 2.2)—we observe that models closely related to these objectives unsurprisingly achieve the best performance. On the RACE++ dataset, the results are close and OrderedLogitNN performs comparably to or even slightly better than the regression and classification models. In contrast, on ARC, the regression approach achieves the lowest RMSE, while classification achieves the highest accuracy. These results are expected, as the regression method is directly optimized with RMSE loss while the classification method entirely omits the ordinal information, just like the accuracy metric.

7.2. Confusion matrix

To further investigate the behavior of the models, Figure 5 presents confusion matrices for the ARC dataset, the more complex task in this study. These matrices are normalized by the true class (i.e., row-wise) and are based on discrete predicted levels—rather than full probability distributions—thus aligning with the evaluation setting of the balanced DRPS with degenerate predictions.

The results indicate the inherent difficulty of the task, as demonstrated by the substantial deviations from the diagonal (highlighted in black). All models hardly predict level 6 as these observations are least present in the training set. Notably, the CORAL model exhibits highly atypical behavior, predicting exclusively at the extreme levels (3 and 9). This pattern suggests that the model fails to converge to a meaningful solution, consistent with its poor performance on the balanced DRPS metric.

Among the remaining models, all except OrderedLogitNN show concentrated errors in specific off-diagonal cells, i.e., predicting level 4 instead of level 3, or level 8 instead of level 9. These errors are associated with the outermost classes (levels 3 and 9), which are often neglected entirely by the models. In contrast, OrderedLogitNN stands out as the only method that successfully captures both extremes of the ordinal scale, avoiding these consistent misclassifications.

8. Conclusion

This study approaches discrete-level QDE through the lens of ordinal regression, reflecting the inherent ordering of difficulty levels from easiest to hardest. Prior work in this area has ignored this ordinal

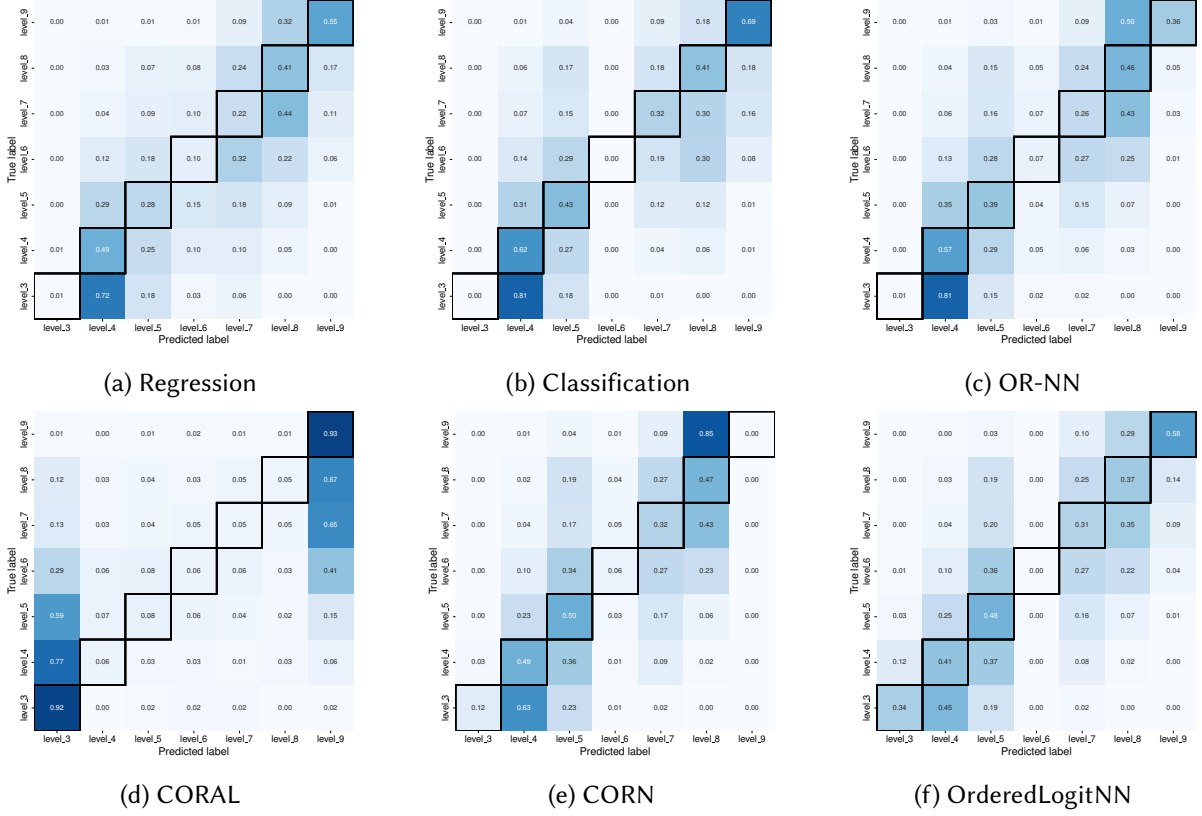


Figure 5: Confusion matrices on the ARC dataset for all models, normalized over the true labels (rows).

structure, both in the choice of modeling paradigms and in the design of evaluation metrics.

To address these gaps, we benchmark three types of model outputs—discretized regression, classification, and ordinal regression—using the balanced DRPS, a novel metric that captures both ordinality and class imbalance. Moreover, we propose a novel ordinal regression model OrderedLogitNN, extending the ordered logit model from econometrics to NNs. We fine-tune the Transformer model BERT on the RACE++ and ARC datasets.

Experimental results indicate that OrderedLogitNN considerably outperforms existing methods on more complex tasks while performing comparably on simpler ones, both on probabilistic and degenerate predictions. Probabilistic predictions express uncertainty and are particularly valuable in downstream human-in-the-loop tasks such as selective prediction or active learning [29, 15]. These findings underscore OrderedLogitNN as a highly attractive approach for discrete-level QDE. The learning dynamics of OrderedLogitNN deserve further investigation and future work can assess the impact of adapting the weight initializations and learning rate multiplier.

Our work has important practical implications for QDE, as more reliable difficulty estimation enables scalable personalized learning paths in educational platforms. Furthermore, it has broader relevance for the automated evaluation of assessment content with ordinal labels, such as essay correction. The balanced DRPS provides a principled foundation for evaluating such systems in production and for future research. Finally, OrderedLogitNN’s robustness makes it well-suited for integration in educational applications where strong performance over the entire label range is critical.

Acknowledgments

We thank the anonymous reviewers for providing valuable feedback. This study was supported by the Research Foundation Flanders (FWO) (grant number 1S97022N).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o in order to check grammar and spelling. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

- [1] W. J. Van der Linden, C. A. Glas, Computerized adaptive testing: Theory and practice, Springer, 2000. doi:10.1007/0-306-47531-6.
- [2] Q. Wang, J. Liu, B. Wang, L. Guo, A regularized competition model for question difficulty estimation in community question answering services, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1115–1126. doi:10.3115/v1/D14-1118.
- [3] Y. Attali, L. Saldivia, C. Jackson, F. Schuppan, W. Wanamaker, Estimating item difficulty with comparative judgments, ETS Research Report Series 2014 (2014) 1–8. doi:10.1002/ets2.12042.
- [4] S. Lane, M. R. Raymond, T. M. Haladyna, et al., Handbook of test development, volume 2, Routledge New York, NY, 2016.
- [5] L. Benedetto, P. Cremonesi, A. Caines, P. Buttery, A. Cappelli, A. Giussani, R. Turrin, A survey on recent approaches to question difficulty estimation from text, ACM Computing Surveys 55 (2023) 1–37. doi:10.1145/3556538.
- [6] R. K. Hambleton, R. W. Jones, Comparison of classical test theory and item response theory and their applications to test development, Educational measurement: issues and practice 12 (1993) 38–47.
- [7] R. K. Hambleton, Fundamentals of item response theory, Sage, 1991.
- [8] F.-Y. Hsu, H.-M. Lee, T.-H. Chang, Y.-T. Sung, Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques, Information Processing & Management 54 (2018) 969–984.
- [9] H. Yang, E. Suyong, Feature analysis on english word difficulty by gaussian mixture model, in: 2018 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2018, pp. 191–194.
- [10] J. Fang, W. Zhao, D. Jia, Exercise difficulty prediction in online education systems, in: 2019 International Conference on Data Mining Workshops (ICDMW), IEEE, 2019, pp. 311–317.
- [11] L.-H. Lin, T.-H. Chang, F.-Y. Hsu, Automated prediction of item difficulty in reading comprehension using long short-term memory, in: 2019 International Conference on Asian Language Processing (IALP), IEEE, 2019, pp. 132–135.
- [12] Y. Zhou, C. Tao, Multi-task bert for problem difficulty prediction, in: 2020 international conference on communications, information system and computer engineering (cisce), IEEE, 2020, pp. 213–216. doi:10.1109/CISCE50729.2020.00048.
- [13] E. Loginova, L. Benedetto, D. Benoit, P. Cremonesi, Towards the application of calibrated transformers to the unsupervised estimation of question difficulty from text, in: RANLP 2021, INCOMA, 2021, pp. 846–855.
- [14] L. Benedetto, A quantitative study of nlp approaches to question difficulty estimation, in: International Conference on Artificial Intelligence in Education, Springer, 2023, pp. 428–434. doi:10.1007/978-3-031-36336-8_67.
- [15] A. Thuy, E. Loginova, D. F. Benoit, Active learning to guide labeling efforts for question difficulty estimation, arXiv preprint arXiv:2409.09258 (2024). doi:10.48550/arXiv.2409.09258.
- [16] R. Coe, J. Searle, P. Barmby, K. Jones, S. Higgins, Relative difficulty of examinations in different subjects, Durham: CEM centre (2008).
- [17] Y. Liang, J. Li, J. Yin, A new multi-choice reading comprehension dataset for curriculum learning, in: Asian Conference on Machine Learning, PMLR, 2019, pp. 742–757.

- [18] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, O. Tafjord, Think you have solved question answering? try arc, the ai2 reasoning challenge, arXiv preprint arXiv:1803.05457 (2018). doi:10.48550/arXiv.1803.05457.
- [19] T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation, *Journal of the American statistical Association* 102 (2007) 359–378.
- [20] A. P. Weigel, M. A. Liniger, C. Appenzeller, The discrete brier and ranked probability skill scores, *Monthly Weather Review* 135 (2007) 118–124. doi:10.1175/MWR3280.1.
- [21] W. H. Greene, D. A. Hensher, *Modeling ordered choices: A primer*, Cambridge University Press, 2010.
- [22] X. Shi, W. Cao, S. Raschka, Deep neural networks for rank-consistent ordinal regression based on conditional probabilities, *Pattern Analysis and Applications* 26 (2023) 941–955. doi:10.1007/s10044-023-01181-9.
- [23] L. Li, H.-T. Lin, Ordinal regression by extended binary classification, *Advances in neural information processing systems* 19 (2006).
- [24] Z. Niu, M. Zhou, L. Wang, X. Gao, G. Hua, Ordinal regression with multiple output cnn for age estimation, in: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4920–4928.
- [25] W. Cao, V. Mirjalili, S. Raschka, Rank consistent ordinal regression for neural networks with application to age estimation, *Pattern Recognition Letters* 140 (2020) 325–331. doi:10.1016/j.patrec.2020.11.008.
- [26] G. Lai, Q. Xie, H. Liu, Y. Yang, E. Hovy, Race: Large-scale reading comprehension dataset from examinations, arXiv preprint arXiv:1704.04683 (2017). doi:10.48550/arXiv.1704.04683.
- [27] J. Ansel, E. Yang, H. He, N. Gimelshein, A. Jain, M. Voznesensky, B. Bao, P. Bell, D. Berard, E. Burovski, G. Chauhan, A. Chourdia, W. Constable, A. Desmaison, Z. DeVito, E. Ellison, W. Feng, J. Gong, M. Gschwind, B. Hirsh, S. Huang, K. Kalambarkar, L. Kirsch, M. Lazos, M. Lezcano, Y. Liang, J. Liang, Y. Lu, C. Luk, B. Maher, Y. Pan, C. Puhersch, M. Reso, M. Saroufim, M. Y. Siraichi, H. Suk, M. Suo, P. Tillet, E. Wang, X. Wang, W. Wen, S. Zhang, X. Zhao, K. Zhou, R. Zou, A. Mathews, G. Chanan, P. Wu, S. Chintala, *PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation*, in: *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*, ACM, 2024. URL: <https://pytorch.org/assets/pytorch2-2.pdf>. doi:10.1145/3620665.3640366.
- [28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45. doi:10.18653/v1/2020.emnlp-demos.6.
- [29] A. Thuy, D. F. Benoit, Explainability through uncertainty: Trustworthy decision-making with neural networks, *European Journal of Operational Research* (2023). doi:10.1016/j.ejor.2023.09.009.