Question Quality Assessment

20182705

Bing Gibeom

1. Summary

The goal of this project is to create a model that can make judgments about the quality of diagnostic questions similar to those of experts. With the recent development of online education, the data obtained from it is used to improve education. Assessing the quality of diagnostic questions is an important issue, and in this project, the aim is to make judgments based on the characteristics of Correctness, Difficulty, Variance of Answer, and Length of question, which explain the students' understanding of the questions. Based on the experimental and evaluation results, it was determined that the three characteristics, excluding Variance of Answer, have sufficient explanatory power, and the model that combines them with assigned weights showed the best performance. This indicates that by considering various meaningful data together, a question quality assessment model can be created that closely aligns with expert judgments.

2. Introduction

As online education becomes more accessible, many people now have the opportunity to access education. Within online spaces, people can gain knowledge through images and videos, exchange opinions with others, and receive education. This method, which is not constrained by time and space, is highly effective.

Like any other form of education, this method of education also requires improvement. When people receive education directly through online platforms, various data is collected. The processing and analysis of this data can be used to diagnose the current state of education and suggest directions for improvement.

This project aims to measure the quality of questions asked, as high-quality questions lead to better education for students. In this task, the quality of questions defined by expert panels in the domain is predicted based on data on students' responses. An example of a given question is shown in Figure 1.

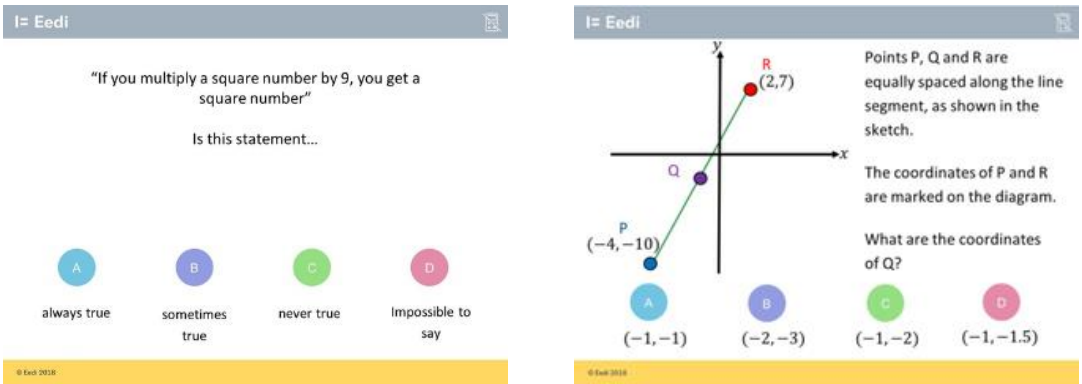

Figure 1. Examples of Question, which is a math problem consisting of a description of question and four answer choices.

In addition, the "Golden Rule" of good diagnostic questions presented by an expert is as follows.

1) They should be clear and unambiguous

2) They should test a single skill/concept

3) Students should be able to answer them in less than 10 seconds

4) You should learn something from each incorrect response without the student needing to explain

5) It is not possible to answer the question correctly whilst still holding a key misconception

By considering these rules, one can analyze the data and determine whether a given question has good quality or not.

3. Methods

3.1 Data

The data provided in this task is extensive data provided by Eedi, an online education provider used by tens of thousands of schools. It explains students' responses to multiple-choice diagnostic questions provided from September 2018 to May 2020. Each diagnostic question is a multiple-choice question with four answer options, one of which is correct. Currently, the platform focuses on math problems. The provided data includes the images of the questions, students' responses to the questions as the main training data, and various metadata. Here is a brief description of the main training data:

• QuestionId: ID of the question answered.

• UserId: ID of the student who answered the question.

• AnswerId: Unique identifier for the (QuestionId, UserId) pair, used to join with associated answer metadata.

• IsCorrect: Binary indicator for whether the student's answer was correct (1 is correct, 0 is incorrect).

• CorrectAnswer: The correct answer to the multiple-choice question (value in [1,2,3,4]).

• AnswerValue: The student's answer to the multiple-choice question (value in [1,2,3,4]).

3.2 Factors

3.2.1 Correctness

It is possible to hypothesize that questions with high accuracy rates are good questions. If a question can accurately convey the problem to students and guide them towards the correct answer, it can be considered a good question. The process of calculating the accuracy rate for each question is as follows:

1) Group the training data based on 'QuestionId' and calculate the mean value of the 'IsCorrect' attribute for each group.

2) Normalize the mean value using StandardScaler to use in PCA.

3.2.2 Difficulty

We can hypothesize that questions with lower difficulty are better questions. To evaluate difficulty, we use the accuracy rate of students who attempted the question. If a high-performing student answers a question correctly, it can be interpreted as having a relatively low difficulty, but if they get it wrong, it can be seen as having a very

high difficulty. Conversely, if a low-performing student gets a question wrong, it can be interpreted as having a relatively high difficulty, but if they get it right, it can be seen as having a very low difficulty. Questions with low difficulty can be considered good questions because they can be understood and the correct answer can be deduced by students who may not study well. The process of calculating the difficulty for each question is as follows:

1) Calculate the ratio of ‘IsCorrect’ count per ‘UserId’ to the total number of questions, to obtain the accuracy rate for each student.

2) Subtract the accuracy rate of the student corresponding to the ‘UserId’ from the ‘IsCorrect’ value of each question to create a new attribute called 'difficulty'.

3) Set the learning data group based on ‘QuestionId’, and calculate the average value of 'difficulty' for each group.

4) Apply StandardScaler to the value for using in PCA.

3.2.3 Length of description

The length of a question is an important measure that determines its quality. Assuming that all the instructions are sufficient, shorter questions are easier for students to understand and take less time to respond to. Therefore, shorter question length is a condition for a good question. The process of calculating the length of each question is as follows:

1) Use tesseract-ocr to extract text from all question images.

2) Save the length of the text.

3) Normalize the value using StandardScaler to use in PCA.

3.2.4 Color number of image

The number of colors present in a problem image is likely to impact the quality of the question. Each question may either include an image or not, which affects the nature of the question. To check the quantity of these images, the number of colors present in the problem image is used. The process of determining the number of colors is as follows:

1) Extract and store the colors of the pixels present in the image.

2) Calculate the number of unique colors stored.

3) Normalize the value using StandardScaler to use in PCA.

3.2.5 Subjects

The number of subjects is expected to influence the quality of a problem. In the given problem's metadata, we can obtain information about the subject(s) of the problem. If a problem has multiple subjects, it requires knowledge of various concepts and involves solving the problem using multiple skills. Therefore, problems with a higher number of subjects are more likely to be ambiguous. The process of determining the number of subjects is as follows:

1) Extract the 'SubjectId' information from the problem's metadata.

3) Normalize the value using StandardScaler to use in PCA.

3.3 Principal Component Analysis

Principal Component Analysis (PCA) is used to reduce the dimensionality of data by linearly transforming it, making the analysis simpler. By transforming the attributes of a problem into a lower-dimensional space, it becomes easier to interpret the selected variables. PCA allows us to obtain new principal components and the components represented by each principal component. To create principal components, we derive the weights that need to be multiplied by the original components. Through this process, we can obtain principal components that effectively describe the data. By multiplying the variances of the two principal components obtained through PCA, considering the explanatory power of each component, we can obtain an overall result that combines all the problems.

3.4 Combination and Evaluation

New values are created by combining the values of each attribute, and rankings are established based on these values. The validation data provided is used to evaluate the results. This data contains evaluations by experts on which of two questions is the better one. The resulting rankings are then compared to the evaluations of each expert and their average evaluations to determine how closely they match. We evaluate the components derived before PCA using indicators based on data represented using the principal components obtained through unsupervised learning, specifically PCA. In this evaluation, we employ the forward selection method. By incrementally adding components and performing PCA, we assess the quality of the problem representation based on the evaluation results. If the evaluation results are positive, we adopt the corresponding component.

You can find the code for all the processes on the following Github:

https://github.com/kibumbing/2305_Question-Quality

4. Results

4.1 Each factor

4.1.1 Correctness

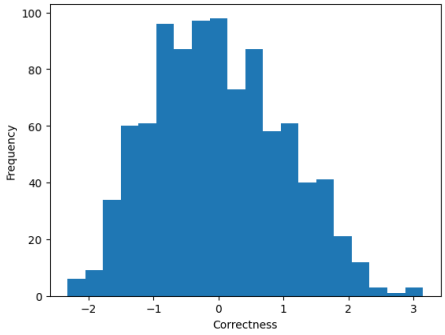

Figure 2. Histogram of Correctness

	T1 ALR	T2 CL	T3 GF	T4 MQ	T5 NS	Average
Accuracy	0.72	0.68	0.76	0.56	0.72	0.69

Table 1. Accuracy of Correctness on Validation data

Table 1 shows the results of the evaluation using only each individual attribute with validation data. T1, T3, and T5 show high agreement, but T4 is relatively lower. The overall accuracy of 0.68 is significantly higher than 0.5, indicating that these attributes can serve as good criteria for determining good questions. Additionally, according to Figure 2, Correctness is relatively normally distributed within the data.

4.1.2 Difficulty

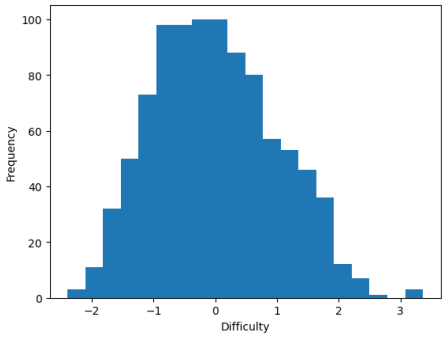

Figure 3. Histogram of Difficulty

	T1_ALR	T2_CL	T3_GF	T4_MQ	T5_NS	Average
Accuracy	0.8	0.68	0.76	0.56	0.8	0.72

Table 2. Accuracy of Difficulty on Validation data

Table 2 shows the results when evaluating based on the attribute of Difficulty with validation data. As it started with the same 'IsCorrect' base, the overall results are similar to the results of Correctness. However, T1 and T5 show improved accuracy compared to Correctness. This can be interpreted as the additional attribute of student grades being reflected. In addition, according to Figure 3, Difficulty is distributed relatively close to a normal distribution within the data and has a distribution similar to Correctness.

4.1.3 Length of description

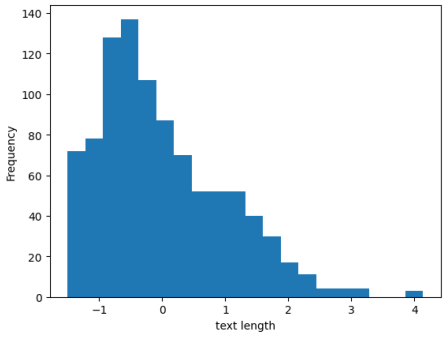

Figure 4. Histogram of Description length

	T1_ALR	T2_CL	T3_GF	T4_MQ	T5_NS	Average
Accuracy	0.56	0.52	0.6	0.64	0.56	0.58

Table 3. Accuracy of Description length on Validation data

Table 3 shows the results of evaluating based solely on the length of the questions. The results are not particularly outstanding, but compared to the other attributes, T4 shows a relatively high level of agreement. Therefore, it is suitable to be considered as a minor factor. According to Figure 4, the distribution has a right-skewed shape, and it can be observed that longer questions are relatively fewer in number.

4.1.4 Color number

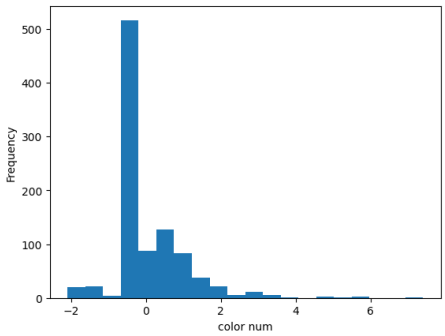

Figure 5. Histogram of Color number

	T1 ALR	T2 CL	T3 GF	T4 MQ	T5 NS	Average
Accuracy	0.48	0.52	0.44	0.56	0.64	0.52

Table 4. Accuracy of Color number on Validation data

Table 4 shows the results of evaluating based solely on the color number. The results good accuracy for T5. However, they exhibit shortcomings for T1 and T3. The range of results is relatively broad, as shown in Figure 5. It exhibits results that are skewed towards a specific number of colors.

4.1.5 Subjects

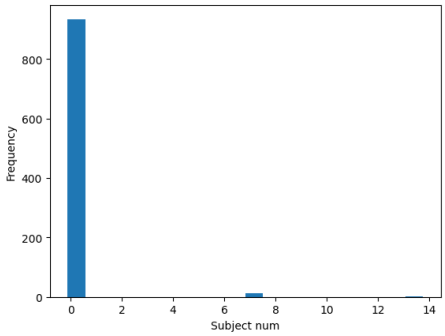

Figure 6. Histogram of Subjects

	T1 ALR	T2 CL	T3 GF	T4 MQ	T5 NS	Average
Accuracy	0.60	0.48	0.56	0.60	0.52	0.55

Table 5. Accuracy of Subjects on Validation data

Table 4 shows the results of evaluating based solely on the subjects. No prominent results were obtained from any source. According to Figure 6, occasional variations are observed, but overall, the results exhibit significant skewness.

4.2 Combination of factors

During the PCA analysis, I checked the usage of each component. I added the components in the order of Difficulty, Correctness, Length of description, Subjects, and Color number based on their high accuracy when each component was used individually.

The results obtained from PCA and forward selection can be seen in Table 6. Accuracy was highest when

Difficulty, Correctness, and Length of description were used together. There was a slight decrease in accuracy when Subjects were added, and a significant drop in accuracy when Color number was added. The addition of Length of description showed a slight increase in accuracy.

	Difficult+Correctness	Difficult+Correctness +Length	Difficult+Correctness +Length+Subjects	Difficult+Correctness +Length+ColorNum
Average Accuracy	0.720	0.744	0.728	0.640

Table 6. Average Accuracy of Combinations on Validation data

The final model's performance on the validation and test dataset is shown in Table 7. Overall, the results are less satisfactory than those on the validation data only, but on average, the model shows a 68% agreement rate with experts. Figure 7 provides a scatter plot of PCA results, allowing us to examine the relationship between components. From this graph, we can observe that Difficulty and Correctness have distinct characteristics compared to Length of description.

Figure 7. Scatter plot of PCA

	T1 ALR	T2 CL	T3 GF	T4 MQ	T5 NS	Average
Accuracy	0.72	0.68	0.80	0.64	0.72	0.712

Table 7. Accuracy of Best Mode on Total(Validation+Test) data

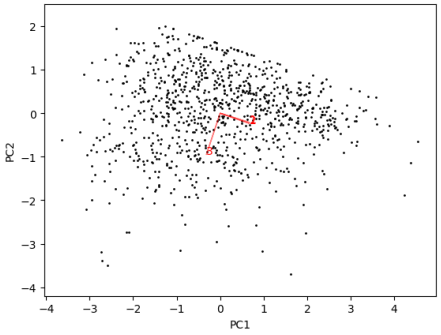

Figure 7. Scatter plot of PCA

5. Discussion

This experiment starts from the hypothesis that four characteristics, namely Correctness, Difficulty, Length of question Color number of image, and Number of subjects, can explain the quality of diagnostic questions. It is believed that students' ability to accurately understand and quickly answer questions is a determining factor in question quality, and this can be measured through the four aforementioned indicators. By examining whether these indicators have sufficient explanatory power and combining them, it is possible to make predictions about the quality of diagnostic questions that are similar to the judgments of experts.

In a single-feature experiment, Correctness and Difficulty yielded predictions that were similar to those of five experts with a probability of around 70%. Among the two characteristics, Difficulty showed a slightly higher improvement in prediction accuracy compared to Correctness. This can be attributed to the inclusion of students' grades as an additional factor in Difficulty, while both characteristics were primarily based on the ‘IsCorrect’ value of the data. This indicates that questions that are easy to understand and lead to accurate answers have good quality.

The Length of question did not show overall strong predictive performance. However, it exhibited good predictions for specific experts compared to other characteristics. Thus, it has some value as a metric for explaining experts' judgments.

Subject and Color number did not yield favorable results. Both components showed significant variability in accuracy among evaluators. The quality of the question could not be consistently predicted solely based on the

concepts covered and the number of images. This indicates that these factors can make the question either easy or difficult, depending on the context. It is likely that these components can be interpreted in conflicting ways. Although these components showed high accuracy for specific evaluators, they do not contribute to consistent judgments.

In the final stage, the three characteristics with good explanatory power, Correctness, Difficulty, and Length of question, were normalized and combined to create a model. The weights for each characteristic were adjusted based on the PCA. Forward selection method was used to add one feature at a time, and the results showed that using all three characteristics achieved approximately 74.4% accuracy in predicting the judgments of experts. On the total(validation+test) data, it yielded a result of 71.2%, slightly lower than the validation data only, but still showed good performance.

6. Conclusion

The goal of this project is to create a model that can make judgments about the quality of diagnostic questions similar to those of experts. According to the hypothesis, Correctness, Difficulty, Length of question Color number of image, and Number of subjects, can be used to assess how easily a problem is understood and answered by students, making them indicative of high-quality questions. Through the experiment, Correctness and Difficulty showed overall high similarity to expert judgments, while Length of question exhibited partial similarity, and Subjects and Color number showed low similarity, indicating an incorrect hypothesis. In the end, by calculating each values that provided by PCA, the model determined rankings that were highly similar to the experts.

However, this project has a limitation. The number of features is too small. Three characteristics are insufficient as indicators for judging question quality. Specifically, these indicators are all focused on whether the problem is easily understood by students, without considering other aspects. Therefore, even when using the features together, there is not a significant change in accuracy. As mentioned in the introduction, if more diverse and additional features are considered, it can improve the quality determination.

Due to these limitations, the accuracy of this model remains around 70%. The proportions for each expert are also similar, making it difficult to claim that it effectively explains specific experts. Therefore, future projects should focus on finding other relevant features to assign weights effectively to address this limitation.

7. References

1) Instructions and Guide for Diagnostic Questions: The NeurIPS 2020 Education Challenge
https://arxiv.org/pdf/2007.12061.pdf

2) What makes a good Diagnostic Question?
https://medium.com/eedi/what-makes-a-good-diagnostic-question-b760a65e0320

3) Assignment Github
https://github.com/ssuai/question_quality_assessment