Students' Learning Behavior Reflected by Self-reported Confidence Level

Richard Holmes and Shuyuan Zhang

April 5, 2022

1 Overview

In this project, we investigated a data set containing results from maths exams run on EEdi, an online student learning platform. Students could optionally state how confident they were about each of their answers. This project investigates how students' confidence level relates to the correctness of their responses. There is also a follow-up exploration on how the correct answer option in multiple-choice questions can impact students' correctness and confidence. Our findings concluded that there is a strong positive correlation between confidence and correctness. This project also discovered that when students opted to provide their confidence they were more likely to be correct. We also found that the position of the correct answer in a multiple-choice does not significantly impact students' confidence or correctness. However, exam writers typically write fewer questions where the final option is the correct one.

2 Introduction

Context and Motivation This data science study on topic of students' learning behavior is tightly linked to psychology when it comes to students' self reflection. It is interesting to ask and try to answer questions like how well students are self aware of their understanding of the problem, and how consistent or inconsistent their confidence are comparing to their actual correctness.

Previous Work

- Backtracking to 1985, a study by Dan Zakay pointed out that time pressure could not be shown to be a factor that influences confidence, but would result in different strategies being used and will affect the decision process. [5]
- In 1990, Janet A Sniezek, Paul W Paese and Fred S Switzer III noted the phenomenon of people being overconfident in general, and investigated the influence of different types of question on overconfidence. [3]
- Diving deeper into this topic, Asher Koriat and Rakefet Ackerman observed in 2010 finally a relation between confidence and the time consumed to answer the question, focusing on responses from pupils in primary schools. [2]
- In 2021, Grazziotin-Soares, R., Blue, C., Feraro, R. et al. investigated the relationship between confidence and correctness on multiple choice questions and found out that the more misconceptions made by students, the less correct they were overall. However, this study was carried out on dental students, and the dataset from EEdi was instead focused on Maths and Science questions - this left space for more research in this topic and the possibility to generalize the findings. [1]

Objectives The main question is, when controlling for other circumstances, are students' confidence levels positively correlated with their actual correctness? Diving deeper into the field of psychology, an interesting follow-up question is when students are asked to self reflect on their answers and to give their confidence levels, will that help them to spot any hidden errors and thus to increase their correctness, and further to improve their learning outcomes? The second question is, does the placing of options or the position of the actual correct answer have an impact on students' correctness rate?

3 Data

Data Provenance The dataset is downloaded from `https://www.inf.ed.ac.uk/teaching/courses/fds/data/project-2021-2022/eedi/`, and originally obtained from the NeurIPS 2020 Education Challenge. [4] As required, following the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, we will only utilise this dataset for study purposes and credit the original source. We have some custom modifications to the dataset including cleaning and transformations, which will be discussed in later sections, and we will not distribute this modified version publicly.

Data Description The dataset consists of several CSV files, including subjects with their hierarchies, tasks with corresponding question ID and answer ID as well as whether they are answered correctly or not, students with their gender, date of birth and whether they are premium pupils. For question and answer data, confidence and involved subjects are also recorded. There are 1382727 records in total, with 6148 unique users, 948 questions, and 388 subjects.

The task and answer tables were both very large containing records of students' individual responses to questions. Despite their size (over a million rows in both), Python was able to quickly process and manipulate the data set.

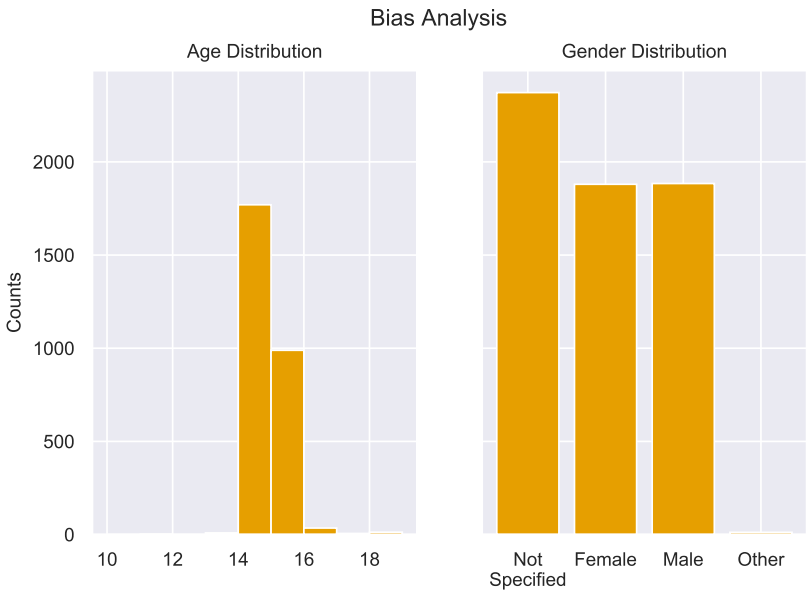


Figure 1: Bias analysis in Age distribution and Gender distribution.

Data Processing The pair of large tables could both be indexed using the ID number of the answer so they were easy to join. The question and student metadata tables were also joined using their ID numbers which were already present in the joined table. The subject table was not joined to the main table since it did not contain any relevant information which couldn't be extracted from the question table's column containing subject IDs.

The gender data was provided in a single column with an integer value ranging from 0 to 3 representing the gender being unspecified, male, female, and other. Extra columns containing dummy variables for each of the possible genders were added to the table to allow regression analysis.

The question table's subject ID column was provided as a string like "$[3, 71, 98, 209]$" which was very difficult to work with. For convenience, it was converted into a list of integers representing subject

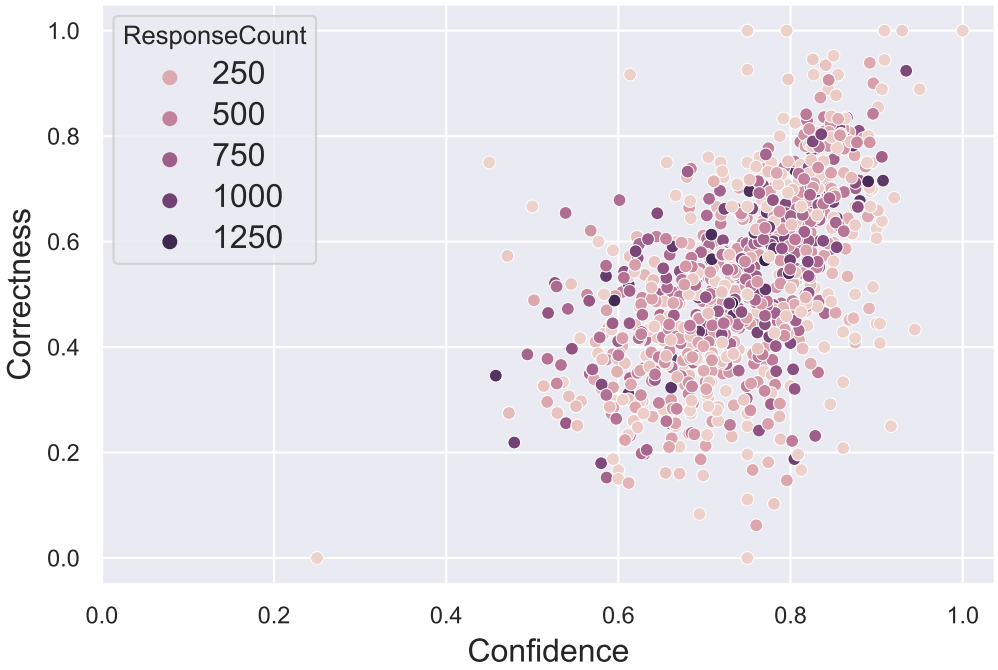


Figure 2: Distribution of mean correctness and confidence. Each point represents mean confidence and correctness for a single question, with hue representing number of responses. Note the scatters with higher response counts, and that their distribution is clearly more concentrated.

IDs then an additional column was added to the main table for each of the subject IDs, from level 0 to 5, though very few subjects went beyond level 3.

4 Exploration and analysis

Bias Analysis After simple processing of the data, the dataset is shown in Figure-1 to have the following characteristics:

- The age distribution is highly concentrated around 14 year-olds, with rare but extreme outliers. Outliers are not included in the graph for clarity.
- The gender distribution is uniform across females and males, and there is a larger sample with gender unspecified. It is then reasonable to consider the dataset to be unbiased towards genders.

4.1 Question 1: Confidence and Correctness

Null Hypothesis Students with high confidence are no more likely to be correct than students with low confidence.

Uncontrolled Observation As shown in Figure-2, there is a positive correlation between confidence and correctness per response when controlling for no other variables.

Controlled Experiments Subject level 3 is the most specific level in which every question has a corresponding subject, therefore we take the mean confidence and correctness for each level 3 subject and plot them against each other in Figure-3. Linear regression is carried out and the result is shown in the

Figure 3: A closer look at the trend and the regression result in subject level 3, with residual plot provided. Since linear regression from the seaborn library is used, no specific statistics is used. This leads to further investigation including a logistic regression to ensure that confidence and correctness are correlated.

Feature	**Coefficient**
Confidence	1.751808
Unspecified Gender	0.714796
Male	-0.369473
Female	-0.431244
Premium Pupil	-0.301231

Confusion Matrix:	*Correct*	*Incorrect*
Predicated as Correct	*16408*	*9616*
Predicated as Incorrect	40342	77882

Table 1: Coefficients gained from logistic regression, with the confusion matrix gained from prediction (of an accuracy of 0.6536).

figure, alongside a plot of the residuals. From the residual plot, it is clear there is no heteroscedasticity, and that most of the points are in a cloud together, implying an accurate linear regression. The adjusted $R^2 = 0.610$ which is high for data based on human behaviour.

Logistic Regression Logistic regression for whether a student's correctness was executed on a subset of training data from all of the responses. Only responses which supplied confidence data and the student's gender and income support status were included. The regression coefficients are shown in Table-1. Confidence has a strong effect on whether a student is correct which matches the previous linear regression analysis. Surprisingly, students who did not specify their gender were expected to perform better than students who specified they were male or female, who were expected to perform similarly. Additionally, students in receipt of income support were expected to be less likely to be correct. Overall the logistic regression model had an accuracy of 65% when assessed using the test data.

Conclusion Combining the results of the above experiments shows a high correlation between confidence and correctness. Therefore there the null hypothesis is rejected.

4.2 Question 1.1: Impact of Self-Reflection on Correctness

From a psychological perspective, it is reasonable to believe that asking students to give their confidence levels will make them reflect on their responses and spend extra time on the question, thus increasing the chance of spotting some errors and further increasing the overall correctness rate.

The Null Hypothesis Students who gave their confidence levels are no more likely to be correct than students who did not give their confidence level.

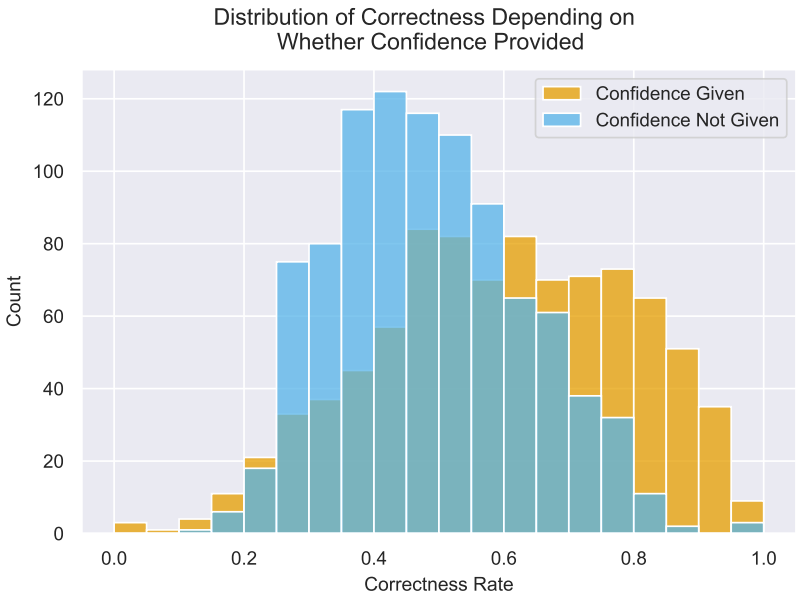


Figure 4: Comparison of distribution of correctness rate per question between pupils who did and did not provide a confidence score.

Uncontrolled Observation As shown in Figure-4, the distribution of correctness rate where a confidence level is given is more left-skewed than the distribution of that when confidence levels are not given by students. This means a higher overall correctness rate in the former group, which supports us to reject the null hypothesis.

Controlled Experiments When controlling for other aspects, for example, subjects and premium pupils, it is possible to get different distributions as results, however, this is proved to be not the case in following experiments:

- In the left part of Figure-5, we are controlling for three major subjects involved: Number, Algebra, and Geometry & Measure. The overall number of records for each of the subjects is in the hundreds thus they are considered samples large enough to prove a common trend.
- In the right part of Figure-5, the control variable is whether students are receiving income support. The trend is again visible even with the presence of outliers.

Conclusion In all cases when controlling for different variables, the students who provided their confidence level have a higher correctness rate. Therefore the null hypothesis is rejected.

4.3 Question 2: Multiple Choice Options

All of the questions within the data-set are multiple-choice, each with 4 possible answer values numbered from 1 to 4.

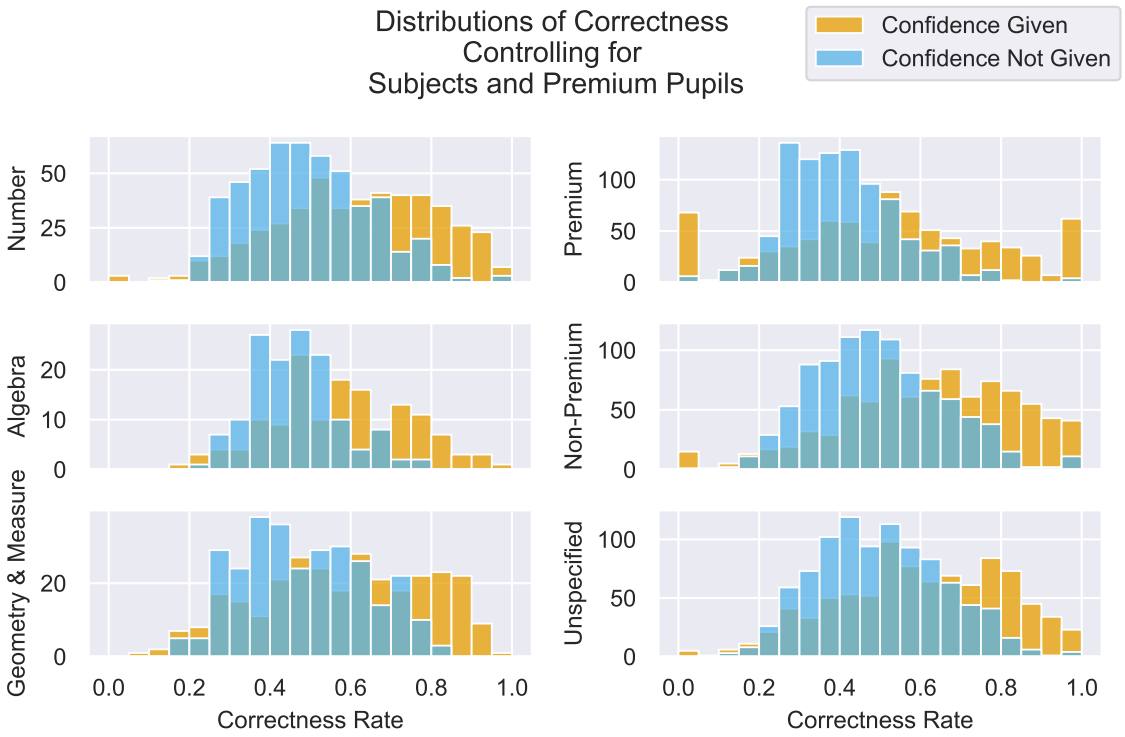


Figure 5: Different distributions of correctness controlling for different variables such as level 1 subject and whether a pupil is receiving income support. Note that trend in each case is similar to the trend in Figure-4

We investigated whether the correct answer's position here had an effect on students' correctness or confidence.

The Null Hypothesis The correct answer's value does not impact students' correctness or confidence.

Observations In terms of raw numbers of responses, there are significant variations in how often different answer values are chosen as shown in Figure-6. Answer option 4 was selected much less often than any of the other options and was only selected 20% of the time. Additionally, as shown in Figure-7a option 4 was correct in significantly fewer questions than the other options, in only 19% of questions. This shows there is some peculiarity surrounding the fourth answer option when compared with the other answer options.

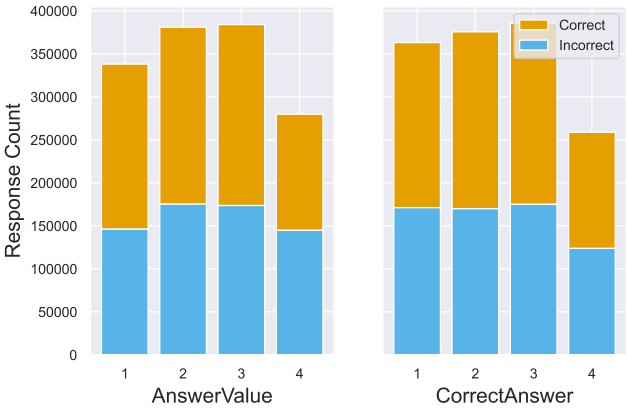


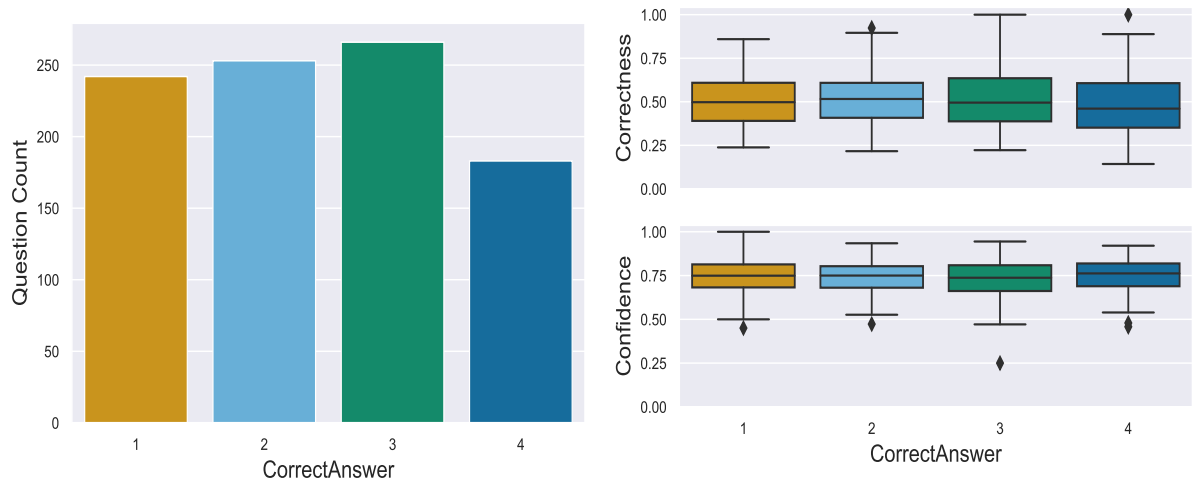
Figure 6: Frequency of Students' Selected Answer Values

Confidence and Correctness Distribution The mean confidence and correctness were calculated per question and then grouped according to the correct answer as shown in Figure-7b. The distribution of confidence does not significantly vary between correct answer values. The distribution of correctness does vary slightly between correct answer values, but not by much. The proportional difference in mean

correctness between option 4 and option 1 is less than 5%. This shows that on a per-question basis the correct answer option does not significantly impact students' confidence or correctness.

KNN Clustering Since the correct answer can be used as a classifier, clustering analysis using K-Nearest-Neighbours was run. For each question its mean confidence and correctness were used, as well as their gender and income support rates. The mean accuracy of the 10 classifiers for each value of the K hyper-parameter is shown in Figure-8. However, the accuracy is not significantly above the $\frac{1}{4} = 0.25$ accuracy expected from guessing between the 4 options of the classifier. An additional evaluation data split would have been created if a value of K had been selected to create a final classifier, but the classifier was deemed to inaccurate for all values of K.

Conclusion While there are fewer questions and a lower overall correctness rate when students respond to questions with option 4, this trend is insignificant when analysed on a per-question basis. The fact that much fewer questions' correct answers are option 4 is the main peculiarity. This could be explained by the answer option 4 being treated differently by exam writers and, for example, used as "all of the above" or "none of the above", and therefore not being as common as other options. The null hypothesis was not rejected.



(a) Number of Questions with Correct Answer Value

(b) Distribution of Correct Answers' Correctness and Confidence per Question

Figure 7: Multiple-Choice Correct Answer Summary

5 Discussion and conclusions

Summary of findings In conclusion to above data analysis, we have the following:

- Confidence and Correctness is positively correlated.
- Students that are asked to give their confidence levels tend to have higher correctness rate.
- Confidence is unaffected by the actual correct answer on a per question basis.
- Correctness is slightly impacted on a per question basis but it may be within margin of error.

Evaluation of own work: strengths and limitations We looked into many different aspects of the dataset that might have had an impact on students' correctness and confidence. We identified several variables which had a significant impact and some that did not. However, our conclusion for question 2 could have been improved if we had access to the actual questions so we could see why there were much fewer questions with option 4 as the correct answer.

We wanted to fit a lot of visualisations into the report, but since both team members were inexperienced with LaTeX formatting a lot of time was wasted attempting to fit them into the page limit.

We managed to cover a wide variety of aspects of the dataset provided to us, but they all fed into the simplistic conclusion that students who are more confident are more likely to be correct. We could have chosen a more specific goal when beginning the project to arrive at a more interesting conclusion.

Comparison with any other related work Compared to the work done by Grazziotin-Soares, R., Blue, C., Feraro, R. et al. mentioned in the previous section, this study is advantageous in terms of the size and variety of samples, however, with a similar limitation of highly focused fields of multiple-choice questions - previous one is extremely limited to a preclinical endodontics course, while this one covers more topics but is still restricted to Maths and Science subjects. [1]

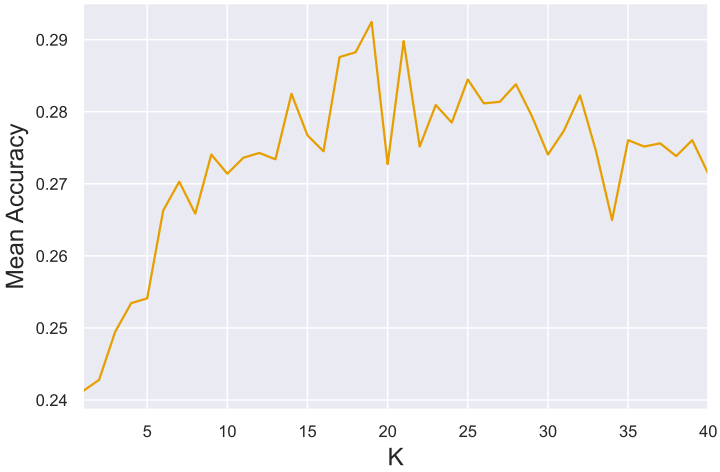


Figure 8: Accuracy of KNN Model Predicting Level 1 Subject: mean accuracy of 10 random models. Expected accuracy of random guessing is 0.25.

The current study also has more samples compared to the one done by Asher Koriat and Rakefet Ackerman, however, it lacks response time, limiting the possibility to investigate wider aspects that might influence the relationship between confidence and correctness. It is possible to obtain a rough response time by comparing the timestamps from the same user ID, however, there is no guarantee that the student is spending all of the time inbetween on the question. Considering the amount of potential errors and how complicated the whole calculation would be, this attempt is not carried out. [2]

A similar trend of a higher confidence than correctness rate is observed as mentioned in the study by Janet A Sniezek et al., however, this project's focus was more on the distribution of these two statistics, and possible conclusions about overconfidence are limited by the narrow range of subjects included in the given dataset.

Improvements and extensions Further investigation about the relationship between confidence and correctness in subjects other than Maths is possible. However, considering the complexity of other possible subjects, they would likely result in responses other than multiple-choice answers.

This study can be also improved if more data is given or found, especially on the response time of each task or questions from other subjects, as mentioned in the above sections. Alternatively, consulting experts from the field of Education or Psychology might help increase the soundness of this study by verifying the assumptions made throughout the study, for example in Section 4.2.

References

[1] Renata Grazziotin-Soares et al. "The interrelationship between confidence and correctness in a multiple-choice assessment: Pointing out misconceptions and assuring valuable questions". In: *BDJ Open* 7.1 (2021). DOI: 10.1038/s41405-021-00067-4.

[2] Asher Koriat and Rakefet Ackerman. "Choice latency as a cue for children's subjective confidence in the correctness of their answers". In: *Developmental Science* 13.3 (Aug. 2009), pp. 441–453. DOI: 10.1111/j.1467-7687.2009.00907.x. URL: https://doi.org/10.1111/j.1467-7687.2009.00907.x.

[3] Janet A Sniezek, Paul W Paese, and Fred S Switzer. "The effect of choosing on confidence in choice". In: *Organizational Behavior and Human Decision Processes* 46.2 (1990), pp. 264–282. ISSN: 0749-5978. DOI: https://doi.org/10.1016/0749-5978(90)90032-5. URL: https://www.sciencedirect.com/science/article/pii/0749597890900325.

[4] Zichao Wang et al. "Diagnostic questions: The neurips 2020 education challenge". In: *arXiv preprint arXiv:2007.12061* (2020).

[5] Dan Zakay. "Post-decisional confidence and conflict experienced in a choice process". In: *Acta Psychologica* 58.1 (1985), pp. 75–80. ISSN: 0001-6918. DOI: https://doi.org/10.1016/0001-6918(85)90035-6. URL: https://www.sciencedirect.com/science/article/pii/0001691885900356.