1 Datasets documentation

We introduce a new dataset named EEDI and describe two existing datasets: WORDBANK and DUOLINGO. All datasets are in a public Google Drive folder: https://drive.google.com/drive/folders/1cFez1tATsCgXOGBxpR2oHAtuWUqPqlDT?usp=sharing.

1.1 EEDI

The EEDI dataset is derived from the NeurIPS 2020 Education Challenge dataset [Wang et al., 2020], provided by the Eedi online educational platform[1]. It contains student responses to math multiple-choice questions (see Figure 1) collected between September 2018 and May 2020. The NeurIPS 2020 Education Challenge dataset provided question content in image format (e.g., Figure 1) without accompanying texts. With permission from Eedi, we have extracted the text from these question images and released this modified dataset. We excluded questions with graphs or diagrams since most current language models do not support visual inputs. The modified dataset contains 573 unique questions and 443, 433 responses to these questions from 2, 287 students, along with data on their age (mostly 11-12 years), gender, and socioeconomic status.

Our EEDI dataset can be found in the `all_data.csv` file inside the `eedi` folder. For our analysis, we randomly selected a test set of 50 questions and 150 students who answered all these questions, which can be found in the file `test_data.csv`. The dataset includes the following columns:

- **QuestionId:** Unique identifier for the question.
- **UserId:** Unique identifier for the student who answered the question.
- **AnswerId:** Unique identifier for the (QuestionId, UserId) pair.
- **IsCorrect:** Indicates whether the student's answer was correct (1 is correct, 0 is incorrect).
- **CorrectAnswer:** The correct answer option (1, 2, 3, 4 corresponding to A, B, C, D).
- **AnswerValue:** The student's chosen answer option (1, 2, 3, 4 corresponding to A, B, C, D).
- **Gender:** The student's gender (0 for unspecified, 1 for female, 2 for male, 3 for other).
- **DateOfBirth:** The student's date of birth, rounded to the first of the month.
- **age:** The student's age in years, calculated from the DateOfBirth.
- **PremiumPupil:** Indicates if the student is eligible for free school meals or pupil premium due to being financially disadvantaged.
- **DateAnswered:** The date and time when the question was answered, rounded to the nearest minute.
- **Confidence:** Percentage confidence score for the answer. 0 means a random guess, 100 means total confidence.
- **GroupId:** Identifier for the class/student group assigned the question.
- **QuizId:** Identifier for the quiz containing the question.
- **SchemeOfWorkId:** Identifier for the scheme of work under which the question was assigned.
- **problem:** The question content in text.

We prompted the gpt-4-vision-preview[2] model to extract question texts from images, and we used the prompt in Figure 2. Specifically, we asked the model to convert the question depicted in the image into a math word problem and identify whether each question contains graphs or diagrams. After gathering the model's responses for all questions, we excluded those involving graphs or diagrams. Additionally, we manually reviewed 10% of the extracted questions to verify the quality of the text extraction.

The EEDI dataset is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License. We bear all responsibility in case of violation of rights. The dataset has

[1] https://eedi.com

[2] https://platform.openai.com/docs/guides/vision, https://platform.openai.com/docs/models/gpt-4o

been uploaded to a public Google Drive folder to ensure easy accessibility. We are committed to maintaining the long-term availability of this dataset. Regular checks and updates to the folder's links and access permissions will be conducted to guarantee that it remains accessible to researchers and the public for the foreseeable future. Additionally, we will monitor the storage platform's policies and adapt our preservation strategies as necessary to ensure ongoing access. The complete NeurIPS 2020 Education Challenge dataset, including additional question metadata like subject areas, is available for download at https://eedi.com/us/projects/neurips-education-challenge (see details in Wang et al. [2020]).

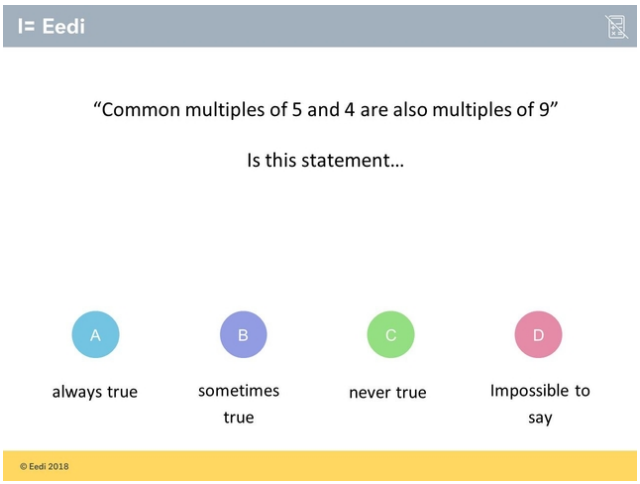

Figure 1: An example of a question from the EEDI dataset.

```
Please convert the problem in the image to a
math word problem. Keep the numbers and letters
in their original form. Put the entire problem
text (including the answer choices) in double
square brackets like this [[problem and answer
choices]]. Note that the problem and answer
choices should be in the same brackets.
Additionally, does the problem or any of the
answer choices contain graphs/diagrams? Answer
yes or no in double square brackets like this
[[yes/no]].
```

Figure 2: The prompt we used to extract question texts from images.

1.2 WORDBANK

The WORDBANK dataset is from the WordBank database[3] [Frank et al., 2017] and is licensed under a Creative Commons Attribution 4.0 International License. We focus on the English (American) subset, which includes responses from $5,520$ children aged between 16 and 30 months. Each child responded to 680 vocabulary items. We only consider items that are words. The responses, reported by parents, are binary and indicate whether the child can produce each word. The dataset also contains demographic details for each child such as age, gender, ethnicity, and the education level of the mother.

The WORDBANK dataset is stored in the `all_data.csv` file within the `wordbank` folder. We randomly selected a test set of 50 words and 150 children for our analysis (see `test_data.csv`). The dataset includes the following columns:

[3] github.com/langcog/wordbankr

- **QuestionId:** Unique identifier for the vocabulary item.
- **UserId:** Unique identifier for the child.
- **AnswerId:** Unique identifier for the (QuestionId, UserId) pair.
- **IsCorrect:** Indicates whether the child can produce the vocabulary item (1 is yes, 0 is no).
- **sex:** The child's gender.
- **age:** The child's age in months.
- **ethnicity:** The child's ethnicity.
- **mom_ed:** The education level of the child's mother.
- **DateAnswered:** This is the same as the QuestionId; specific timing data for when the word was produced is not available.
- **problem:** Definition of the vocabulary item.

1.3 DUOLINGO

The DUOLINGO dataset is from the 2018 Duolingo Shared Task on Second Language Acquisition Modeling (SLAM)[4] [Settles et al., 2018] and is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License. This dataset contains anonymized data from users of the educational application Duolingo[5]. We focus on the subset of English speakers learning Spanish through lesson sessions. Each user's data consists of a series of binary responses to vocabulary words, with each word presented multiple times. Following the approach described in Wu et al. [2020], we adapted this dataset for Item Response Theory modeling by averaging responses to each vocabulary item, rounding the average score to a binary outcome (0 or 1). For instance, if a user was shown the word "hola" 10 times and correctly translated the word 5 times, the average score would be 0.5 and rounded to 1. After processing, the dataset includes $2,783$ vocabulary words and $573,321$ responses from $2,640$ users, with missing data due to user dropout. The dataset also includes additional user information such as country and device type.

The DUOLINGO dataset is stored in the `all_data.csv` file within the `duolingo` folder. We randomly selected a test set of 50 words and 500 users who responded to these words for our analysis (see `test_data.csv`). The dataset includes the following columns:

- **QuestionId:** Unique identifier for the vocabulary word.
- **UserId:** Unique identifier for the user.
- **AnswerId:** Unique identifier for the (QuestionId, UserId) pair.
- **IsCorrect:** Indicates whether the user correctly translated or spelled the vocabulary word (1 is yes, 0 is no).
- **countries:** The user's countries.
- **client:** The user's device platform.
- **DateAnswered:** The number of days since the user started learning the language on Duolingo.
- **problem:** Definition of the vocabulary word.

2 Code

Our code is available here: https://github.com/joyheyueya/psychometric-alignment. Detailed instructions for reproducing our main experimental results can be found in the `README.md` file.

[4] sharedtask.duolingo.com/2018.html
[5] duolingo.com

References

Michael C Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A Marchman. Wordbank: An open repository for developmental vocabulary data. *Journal of child language*, 44(3):677–694, 2017.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. Second language acquisition modeling. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 56–65, 2018.

Zichao Wang, Angus Lamb, Evgeny Saveliev, Pashmina Cameron, Yordan Zaykov, José Miguel Hernández-Lobato, Richard E Turner, Richard G Baraniuk, Craig Barton, Simon Peyton Jones, et al. Instructions and guide for diagnostic questions: The neurips 2020 education challenge. *arXiv preprint arXiv:2007.12061*, 2020.

Mike Wu, Richard L Davis, Benjamin W Domingue, Chris Piech, and Noah Goodman. Variational item response theory: Fast, accurate, and expressive. *arXiv preprint arXiv:2002.00276*, 2020.