CAHL

Computational Approaches to
Human Learning (CAHL) research lab

Evaluating ChatGPT-generated Textbook Questions using IRT

Shreya Bhandari, Yunting Liu, Zachary A. Pardos
University of California, Berkeley

Motivation for the study

Abstract. Item development is a critical component of contemporary measurement systems, playing a pivotal role in evaluating knowledge and skills. The current body of measurement literature lacks insight into an evaluation of leveraging generative AI for textbook question generation. In our study, we make use of rigorous measurement methodologies to evaluate and compare the psychometric properties of gold-standard textbook questions with those of ChatGPT4-generated questions, produced from textbook chapter summaries.

Research questions:

- What is the range of IRT item difficulty and discrimination parameters for ChatGPT generated questions?
- Do they differ significantly from the parameters fit to human-authored, gold-standard textbook questions intended for the same textbook content?

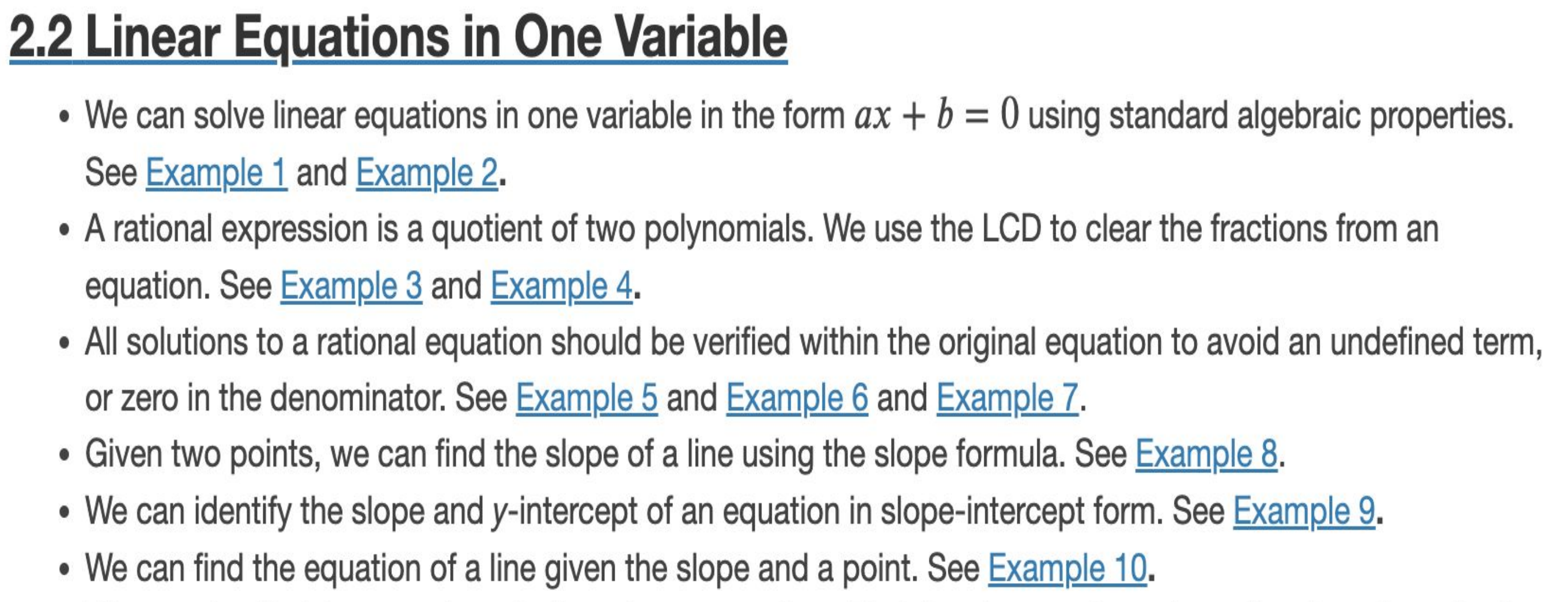
Related work:

- Past research has focused on leveraging LLMs to create math questions using a template-based approach [32], generating open-ended questions [32, 42, 13, 24], and generating multiple choice questions [4]
  (see paper for references)

Methods

Step 1) Item Generation

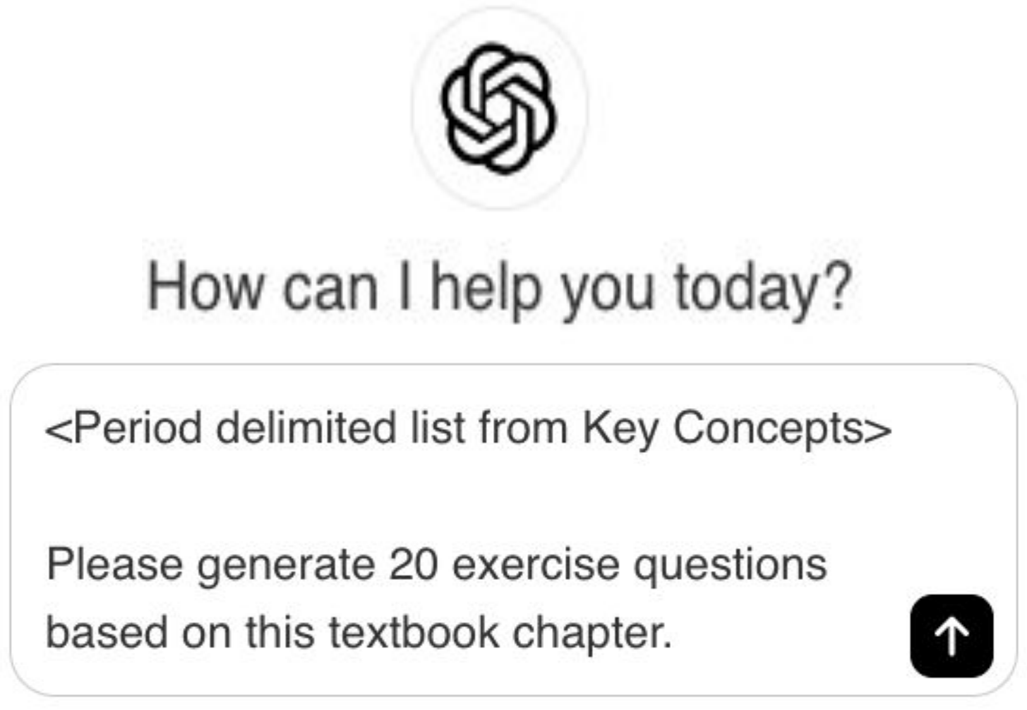
- College Algebra was selected as the subject area, 15 OpenStax questions were selected from Lesson 2.2: *Linear Equations in One Variable*. Examples of learning objectives are shown in Fig 1.
- ChatGPT prompt is shown in Fig 2. Generated questions were **manually checked** (passing rate: 90%) to make sure the question is solvable and that it leads to a single solution.



2.2 Linear Equations in One Variable

- We can solve linear equations in one variable in the form $ax + b = 0$ using standard algebraic properties. See Example 1 and Example 2.
- A rational expression is a quotient of two polynomials. We use the LCD to clear the fractions from an equation. See Example 3 and Example 4.
- All solutions to a rational equation should be verified within the original equation to avoid an undefined term, or zero in the denominator. See Example 5 and Example 6 and Example 7.
- Given two points, we can find the slope of a line using the slope formula. See Example 8.
- We can identify the slope and y-intercept of an equation in slope-intercept form. See Example 9.
- We can find the equation of a line given the slope and a point. See Example 10.

Figure 1 Textbook chapter summary from OpenStax textbook



How can I help you today?

<Period delimited list from Key Concepts>

Please generate 20 exercise questions based on this textbook chapter.

Figure 2 ChatGPT prompt

Step 2) Test design via linking:

- We utilized a measurement technique called a psychometric linking/equating strategy to map different calibration results onto a common scale and thus ensure parameters are comparable to each other across multiple test phases.
- There were **four parallel "forms"** (OpenStax test, ChatGPT test, and two link tests; Similar in length). We ranked all items by difficulty within the original two forms and form an 'easy' link form and a 'hard' link form.
- Each respondent was distributed randomly to one form

Step 3) Data collection

- We recruited respondents via **Prolific**, a popular crowd-sourcing platform, and utilized an open-source tutoring tool, **OATutor,** to deliver questions and collect responses.
- After the four-phase study was run, a total of 248 respondents had participated in the study
- All respondents who spent less than five minutes on the test and those who completed less than 70% of the questions are excluded from further analysis. After the exclusion criteria was applied, the sample size was reduced to 207 respondents

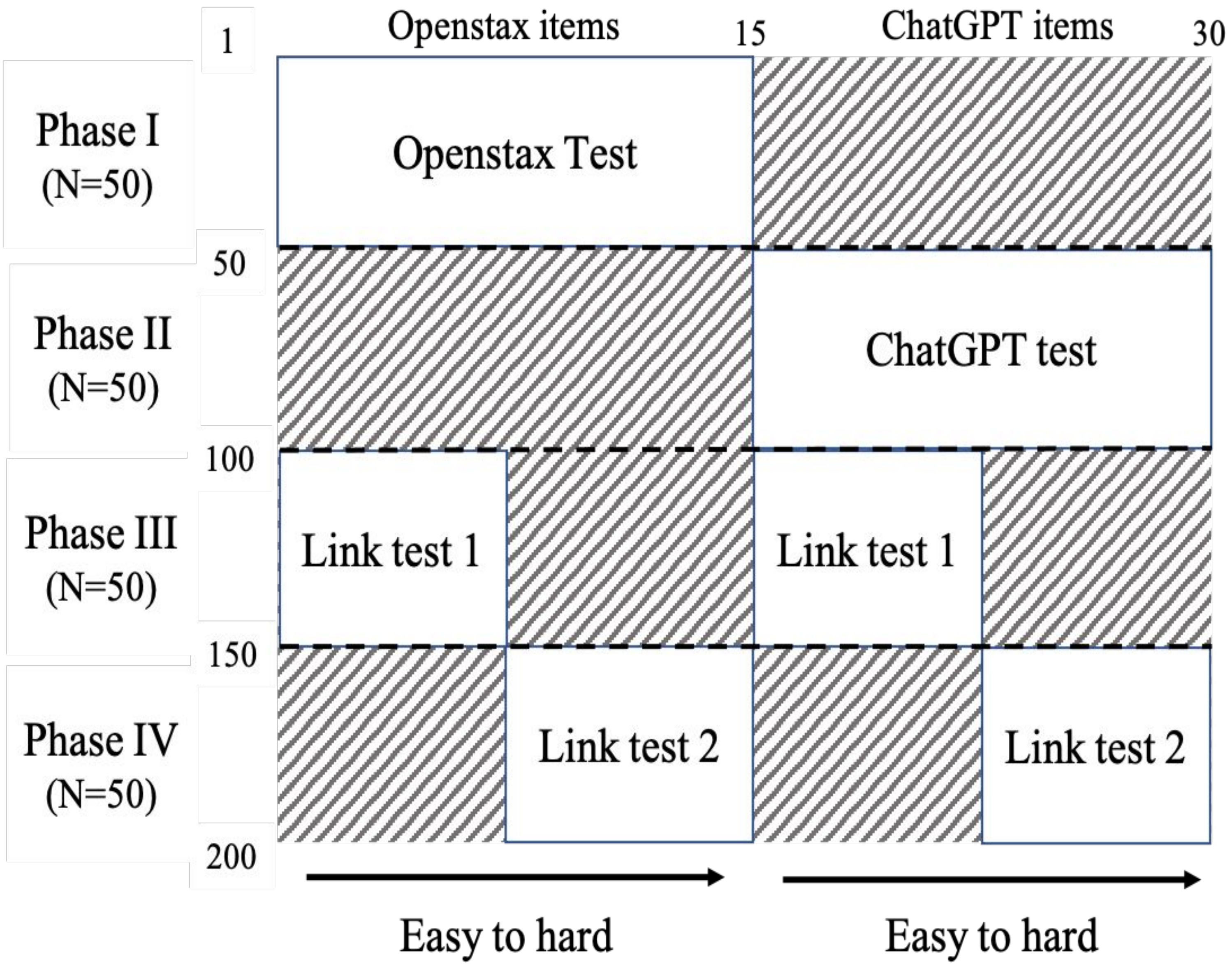
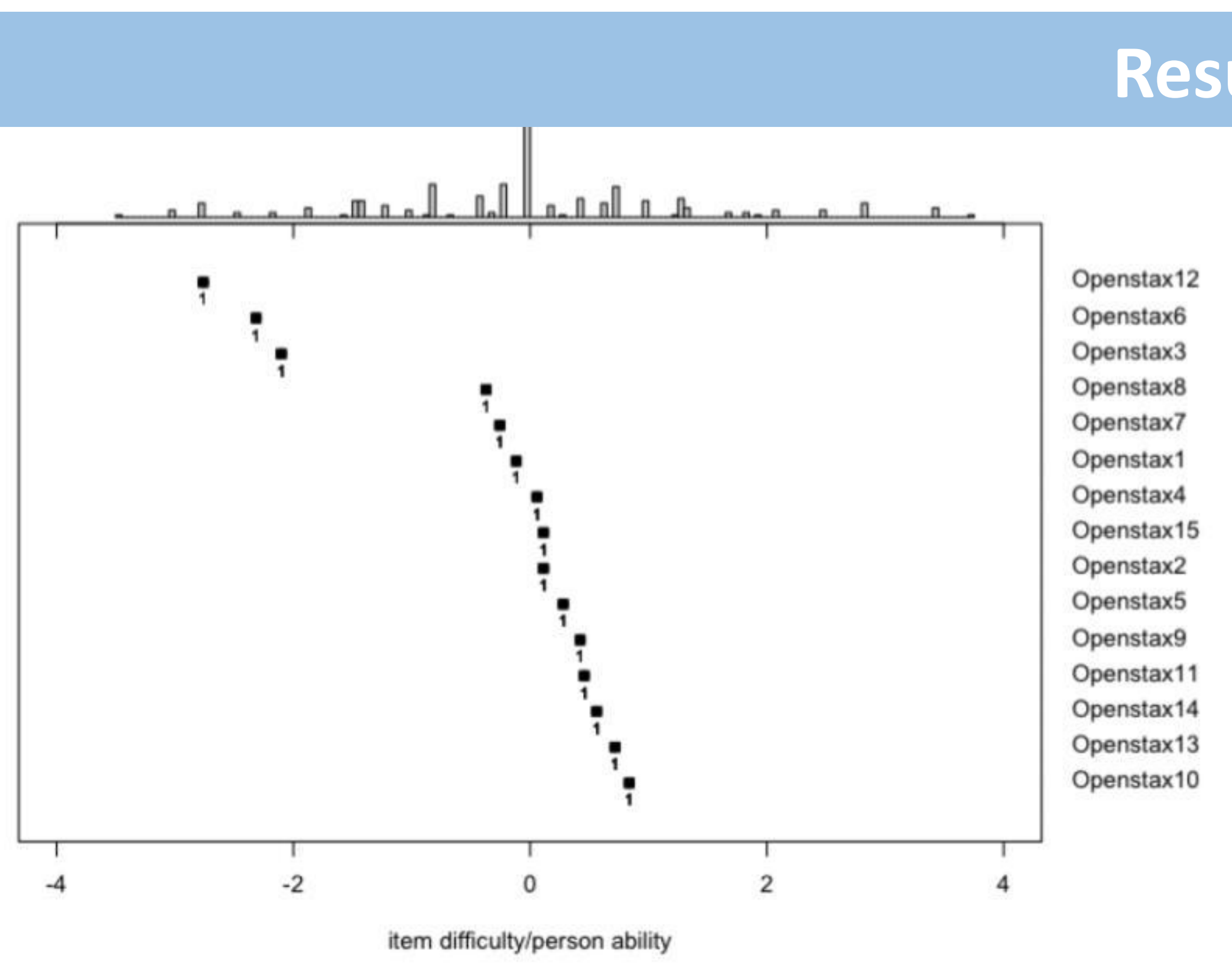
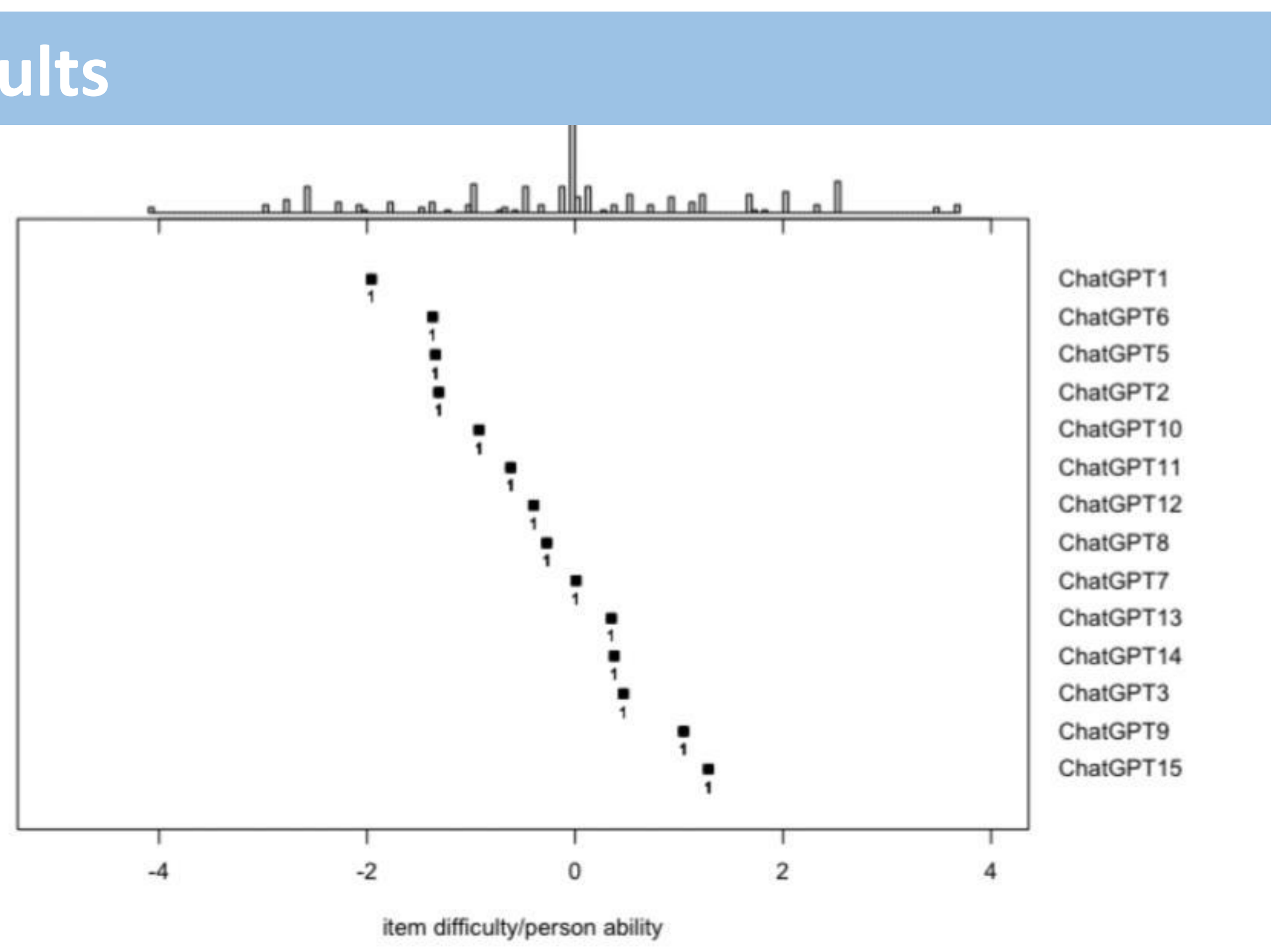


Figure 3 Linking design

Results





Item difficulty : The item difficulty and respondent ability estimate were generated through Rasch analysis, both mapping onto the logit scale. The Openstax items have three items within the range of [-4,-2], which suggests they can better assess the low ability group. ChatGPT is more equally spaced between [-2,2], which means it can do a good job in evaluating the moderate ability group.

Item discrimination : A two parameter logistic model was applied to produce the item discrimination parameters.The average discrimination of ChatGPT questions (1.92) is higher as compared to OpenStax items (1.54), which shows ChatGPT questions indeed do better in discriminating respondents.

Discussion and Limitations

Study implications: The results of our study showed that ChatGPT generated questions have a comparable power to evaluate students' ability when compared with gold standard, human authored textbook questions in College Algebra (no statistically significant differences)

Quality check is needed: One ChatGPT generated item had to be eliminated from analysis due to a 0% accuracy rate. The phenomenon has pointed out that not all ChatGPT items have appropriate qualities, stressing the importance of manual checks from subject matter experts after items are automatically generated.

Limited generalizability of the study result: Our current item generation was based on only a single lesson in the Openstax College algebra textbook. Due to this, we do not know if the results hold for other domains or levels of mathematics.

NeurIPS'23 Workshop: Generative AI for Education (GAIED)