Evaluating ChatGPT-generated Textbook Questions using IRT

Shreya Bhandari
Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720
shreya.bhandari@berkeley.edu

Yunting Liu
School of Education
University of California, Berkeley
Berkeley, CA 94720
yunting99@berkeley.edu

Zachary A. Pardos
School of Education
University of California, Berkeley
Berkeley, CA 94720
pardos@berkeley.edu

Abstract

We aim to test the ability of ChatGPT to generate mathematics assessment questions, given solely a summarization of textbook content. We take a psychometric measurement methodological approach to comparing the qualities of questions, or items, generated by ChatGPT versus gold standard questions from a published textbook. We use Item Response Theory (IRT) to analyze data from 207 test respondents answer questions from OpenStax College Algebra. Using a common item linking design, we find that ChatGPT items fared as well or better than textbook items, showing a better ability to distinguish within the moderate ability group and had higher discriminating power as compared to OpenStax items (1.92 discrimination for ChatGPT vs 1.54 discrimination for OpenStax).

1 Introduction

Much research has been conducted on different processes of question generation, question difficulty control, and psychometric validation. However, the current body of measurement literature lacks insight into an evaluation of leveraging generative AI for textbook question generation. If high quality question generation is possible with AI, we will be able to develop intelligent tutoring systems capable of both unlimited question generation while effectively controlling question difficulty. The potential problems this system would solve are multifaceted. Firstly, lack of control over question difficulty poses issues such as generating questions of inappropriate difficulty that hinders student learning and requiring tedious manual question search from exam designers. With an intelligent tutoring system that is able to efficiently control difficulty, these issues would be mitigated. Secondly, a system capable of unlimited question generation reduces the learning barrier caused by lack of practice and optimizes the labor-intensive process of generating exam questions.

Item development is a critical component of contemporary assessment systems, playing a pivotal role in evaluating knowledge and skills [11]. With the help of high quality items, we are able to draw inferences about students' current knowledge state through different models [39], thus facilitating better instructions. In a formative testing environment, effective assessment can afford practitioners the opportunity to diagnose subjects in a way that leads to tailored remediation. However, this process is associated with substantial costs both in terms of finances and time, often demanding extensive

NeurIPS'23 Workshop on Generative AI for Education (GAIED).

efforts from professionals within the relevant testing domain. Consequently, transformative solutions are sought to cut down costs and enhance the production of effective assessments.

Large Language Models (LLMs), such as ChatGPT [1] are bringing rise to speculation on whether they can competently produce math questions [24, 14]. While some research has been done on other large language models, such as BART and T5, producing math questions [23, 32, 42], rigorous IRT analyses have not been conducted and little is currently known about the validity of ChatGPT generated questions and how they compare to human-generated textbook questions. Therefore, in this work, we aim to compare ChatGPT and human-generated items to answer the following research questions:

- **RQ1:** What is the range of IRT item difficulties for ChatGPT generated questions? How do they compare to human-authored questions?
- **RQ2:** What is the range of IRT item discrimination values for ChatGPT generated questions? How do they compare to human-authored questions in distinguishing between high and low ability groups?

We make all of the item content used in the study available under a creative commons license[2] and provide the source code and a standing demo of the system used to collect respondent data[3] allowing for full replication of our study environment.

2 Related work

Generative AI using ChatGPT is seeing nascent application in education. It has been used to generate help messages and other types of feedback in Algebra [27] and in a decimal learning game setting [22]. ChatGPT has also been used for evaluative purposes, evaluating the goodness of crowd-sourced multiple choice questions using a rubric fed as part of the prompt [21]. In these examples, the efficacy of the technology was promising but not without reported shortcomings. In the remainder of the related work section, we focus on work related to question generation, textbook evaluation, and LLMs in general.

2.1 LLMs and Question Generation

Large Language Models (LLMs) are being pre-trained and fine-tuned with ever-larger text corpora, giving rise to highly advanced and domain-general models. These models typically leverage foundation model [2] architectures associated with the Transformer [35] and its novel attention mechanism. GPT [29] and BERT [8, 31] both leverage the Transformer as their base architecture. However, the main difference comes with BERT utilizing encoding components (i.e., embedding oriented) of the architecture and GPT utilizing decoding components (i.e., generating oriented). ChatGPT is an interface to a machine learning model that is based on the Generative Pre-trained Transformer (GPT) architecture and fine-tuned using Reinforcement Learning with Human Feedback [10]. The GPT model has years of development through autoregressive language modeling and has undergone several stages of evolution. The model has most recently utilized human raters to better fine-tune the model to common human prompts and desirable responses [29, 30, 3, 25]. With ChatGPT being based on such an advanced model, curiosity has arose regarding the extent of its capabilities.

The majority of past research in this field has focused on leveraging the capabilities of LLMs to create math questions either based on specifying a template-based approach [32], open-ended generation (socratic style math questions or math word problems) [32, 42, 13, 24], or multiple choice question generation [4]. The likely reason that these approaches were taken, rather than simply having the LLMs emulate the creation of textbook questions are as follows. Prior language model approaches to the prompt/response scenario treated the response as a text completion of the prompt. But users tend to interact differently with these language models. In essence, they prefer to pose a question or an instruction rather than a simple text completion. This misalignment acted as the motivation for the creation of InstructGPT (or GPT 3.5) [25], the basis for ChatGPT. With the advent of ChatGPT providing this alignment, we aim to test the capabilities of ChatGPT for textbook question generation,

[1] `https://chat.openai.com/chat`

[2] `https://github.com/CAHLR/OATutor-Question-Generation-Final`

[3] `https://cahlr.github.io/OATutor-Question-Generation-Final`

given solely the textbook content. In our study, we utilize GPT-4, the March 14, 2023 version of the model to produce questions.

2.2 Textbook evaluation

Textbook items are different from traditional summative assessment in that summative test result will not be used for other purposes such as selection criteria–a feature referred to as formative assessment. Therefore, extra requirements have been posed for textbook items in particular. Some research argues that whether the item can facilitate active learning and engagement is a vital factor [26], while other argue from a content perspective, stating that it is important for items to encompass higher level skill or show multidimensionality [5, 19]. The content perspective is also stressed in the formative assessment literature [33], they state that an important feature for current formative assessment is better coverage of the construct being measured, being termed as 'Representativeness'. In summary, an optimal set of textbook items should have below features: a) They should have good content coverage of the construct they are measuring, optimally consist of some higher level construct which may be a progression of the current one [38]. b) They should facilitate student thinking and subsequent studying, which means it should be interesting and probably innovative and should be of moderate difficulty so as not to frighten students. Because the questions found in tutoring systems are also formative assessments, questions generated to satisfy the above textbook question criteria may also be beneficial in tutoring contexts.

3 Methods

3.1 Subject selection and Item Generation

College Algebra was selected as the subject area as it is often important bridging material at many colleges. It is also a subject for which pre-authored questions were available under a CC B-Y license from a notable open textbook publisher, OpenStax[4] [1]. To decide which lesson would be utilized in the study, we looked for the first lesson in which none of the problems depended on any images or figures, since a limitation of ChatGPT and most other LLMs, as of this writing, is that they are not yet multi-modal in their inputs and outputs. The first lesson that satisfied this criteria was *Lesson 2.2: Linear Equations in One Variable*. We selected 15 OpenStax questions from that lesson to utilize in our study. This number was chosen to "right size" the test given to respondents. We utilized a random number generator to randomly select 15 questions to be included out of the pool of available items in that lesson. At the end of each chapter, there was a "Chapter Review" with "Key Concepts," summarizing each lesson with bullet points (Figure 1). We leveraged this when prompting GPT-4. In order to prompt ChatGPT to generate questions, the following prompt was given to ChatGPT:

```
<Period delimited list from Key Concepts>
Please generate 20 exercise questions based on this textbook chapter.
```

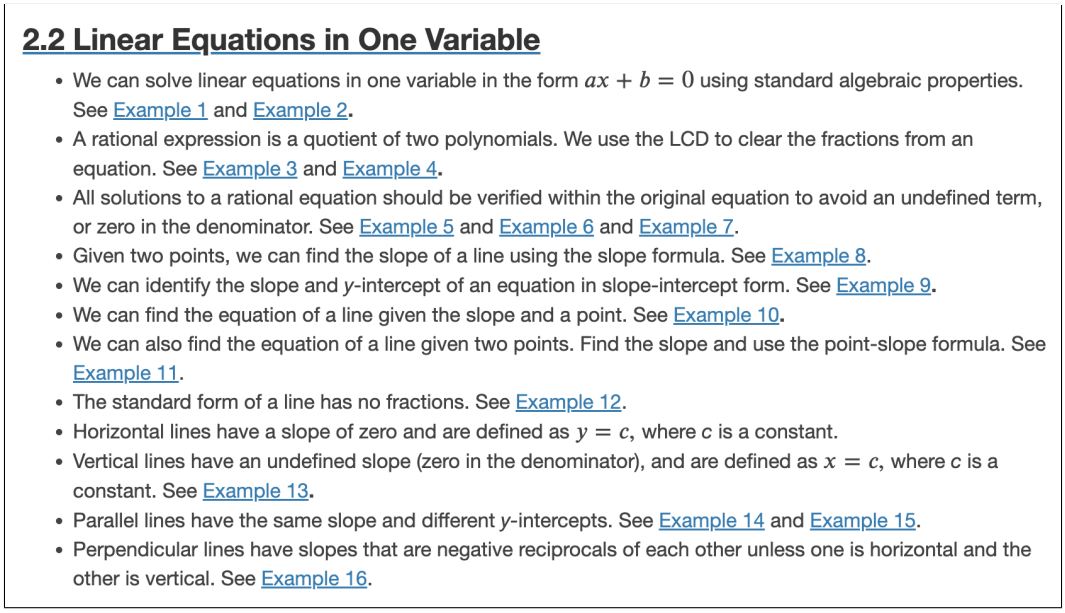
Each question was quality checked against the following 3-point criteria:

- Use of inappropriate language
- The question is solvable
- The question leads to a single solution (i.e. does not have multiple answers)

Each question generated by ChatGPT was attempted to be manually solved by the first author of the paper to ensure that the question was solvable. If a question failed one or more of the quality checks, it was eliminated from our question pool. Among the questions that passed all the checks, we utilized a random number generator to choose 15 questions to be included in our study. We did not apply a similar quality check to the OpenStax questions, as it is assumed that a separate quality control processed was used, internal to the publisher before the items were made public in their textbook.

As an example, the first question generated by ChatGPT was "Solve for x: $5x + 10 = 0$," which corresponds to the first bullet point in the OpenStax Chapter 2.2 Key Concepts (Figure 1).

[4] `https://openstax.org/details/books/college-algebra-2e`



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.
- We can also find the equation of a line given two points. Find the slope and use the point-slope formula. See Example 11.
- The standard form of a line has no fractions. See Example 12.
- Horizontal lines have a slope of zero and are defined as $y = c$, where c is a constant.
- Vertical lines have an undefined slope (zero in the denominator), and are defined as $x = c$, where c is a constant. See Example 13.
- Parallel lines have the same slope and different *y*-intercepts. See Example 14 and Example 15.
- Perpendicular lines have slopes that are negative reciprocals of each other unless one is horizontal and the other is vertical. See Example 16.

Figure 1: OpenStax Chapter 2.2 Key Concepts

3.2 Linking design

The goal of our research is to compare the quality between human generated items and ChatGPT generated items. However, we wanted to ensure that no respondent's test exceeding 15 items. To accomplish this 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 [15, 41] across multiple respondent test phases. Possible methods for linking require anchors which consist of a person who answers items from both tests or an item taken by members of two groups [34]. As the test length cannot be doubled in the current experimental environment, common item design is the most efficient and accurate way for the current study. This approach has also been shown to be effective spanning across the field of Education [6], Medical health [17], Psychology [12]. In 2013, Meyer and Zhu introduced the linking methods for assessment on online higher education platform-MOOC platform [18].

Even though the method does not require equivalent samples, the sampling and resampling process from a crowd-sourcing platform (we will shortly introduce in the next section) can be safely assumed equivalent, in part due to the recruitment criteria applied. This assumption ensures the estimation process is unbiased (we will elaborate more later). Specifically, we connected two types of items (human-designed textbook items, ChatGPT-generated items) using link forms. Two link forms were established to keep the test length similar across all test forms, which ensures respondents don't experience test fatigue. Another merit of using two parallel forms rather than one is that can allow us to double check the quality of the link and measurement invariance. In total, there are four forms (OpenStax form, ChatGPT form, and two link forms). Each respondent will be distributed one form randomly from the available four. The assignment of items to the two link form is conducted based on item property information gained from initial calibrations, namely, we rank all items by difficulty within the original form and form an 'easy' test and a 'hard' test . Through the assignment, the link form has consistency in the overall difficulty. The test design of our study is shown in Figure2.

3.3 Data collection

We recruited respondents via Prolific, a popular crowd-sourcing platform, and utilized an open-source tutoring platform, OATutor [28], to deliver questions and collect responses as this platform had already transcribed OpenStax College Algebra question content for data collection. OATutor was run with all correctness feedback and other tutoring facilities turned off (i.e., "test mode") for this study and the order of the question was randomized per user. To ensure the quality of answers, screening

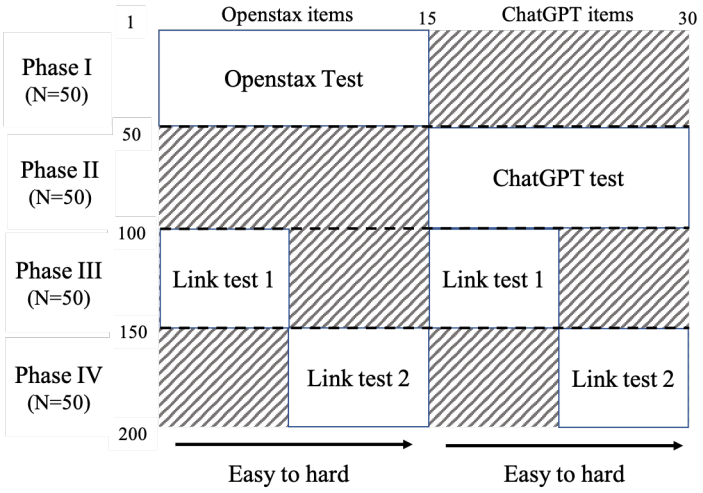


Figure 2: Our Test design using common item equating

tests and quality checks are performed. 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. Each test form comprised around 15 items focused on a specific topic in college algebra, therefore the overall test time was estimated to be under 30 minutes. We restricted all respondents to be current college students in the United States to ensure they were familiar with prerequisites of the test domain. There were four forms randomly assigned to respondents, each respondent could not take more than one form. These forms are distributed to a minimum of 50 participants each. As a result, every individual item will accumulate at least 100 valid responses, fulfilling the basic requirement for the Rasch analysis [36].

3.4 Item calibration

In the current calibration, we applied the Item Response Model (IRT) to create an analysis of item difficulty and item discrimination. The validity of new items was also examined through this process [20]. We calibrated all forms together, which is referred to as the concurrent calibration method [40], because it is more efficient in that it provides smaller standard errors and involves fewer assumptions than separate calibration procedures. As different groups in the study were assumed equivalent, the estimation with incomplete/nonrandomly missing data should not be subject to bias [7].

3.5 Evaluation Criteria

Items were generated for a formative purpose, i.e., they will be mainly used in textbooks as homework questions or reflection questions. The formative nature allows making inferences about student ability based on ongoing assessments [9], the ideal item should still be of moderate difficulty and should be able to distinguish between low proficiency and high so that a student can then self-evaluate and receive diagnostic information from those items.

To compare the quality between two sets of items, we divided the process into different stages:

1. A descriptive analysis was carried out to show differences in item parameters. Specifically, WrightMaps, a figure that can holistically represent item location(difficulty) and person location(ability) [37], allow us to compare the relative difficulty of item sets across two generation method. The discrimination parameters were also compared against each other.

2. An unpaired t-test method was applied here to check whether the item difficulty and item discrimination differ significantly for two generation methods.

4 Results

Out of the original 20 ChatGPT questions, two failed quality checks because they were deemed unsolvable and a random set of 15 from the remaining 18 were selected for the study. Our study with recruited respondents consisted of four-phases. Phase 1 consisted of 15 OpenStax items, Phase 2 had 15 ChatGPT items, Phase 3 consisted of the 8 easiest OpenStax items with the 8 easiest ChatGPT items, and Phase 4 contained the 7 hardest OpenStax items with the 7 hardest ChatGPT items. After the four-phase study was run, a total of 248 respondents were recruited via Prolific (55 for Phase 1, 60 for Phase 2, 62 for Phase 3 and 71 for Phase 4). After the exclusion criteria was applied, of a minimum of 5 minutes spent and 70% of questions attempted, the sample size was reduced to 207 respondents (47 for Phase 1, 52 for Phase 2, 56 for Phase 3 and 52 for Phase 4), which surpassed our initial target of 200 respondents. Attrition rates between the OpenStax items and ChatGPT items were similar (15% for OpenStax and 13% for ChatGPT). The accuracy for each of the phases was 54% (Phase 1), 52% (Phase 2), 62% (Phase 3), and 46% (Phase 4). The relatively high accuracy for the first three phases was expected because they contained either a mix of the "easy" and "hard" items (Phases 1 and 2) or purely contained the "easy" items (Phase 3), while Phase 4 consisted of all the "hard" questions. Out of all of the items, one item from the pool of ChatGPT items had to be removed from further analysis since no respondent was able to provide a correct response. Having a zero percent correct item makes IRT analysis infeasible, forcing the exclusion of this item.

The item difficulty parameter and respondent ability estimate were generated through the Rasch analysis, both mapping onto the logit scale. The upper portions of the WrightMaps (Figure 3) show the respondents' ability distributions. The bottom portion plots the item difficulty estimates, which is referred to as the location where the respondents have a 50% chance to get a particular item correct. The item information is maximized when the item location is the same as the respondent's location [37]. Therefore, the test is more effective when the item parameters have better coverage of the respondents' ability distributions. In Figure 3 (left), we can see that 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 generally good job in evaluating the moderate ability group. Indeed, compared to OpenStax items, the ChatGPT items did better for the moderate to low proficiency group that lies within [-2,-1], as shown in Figure 3(right). ChatGPT generated items are slightly more difficult than the Openstax versions on a logit scale(t = *-0.48, p = 0.64*), but the difference of 0.05 logit is considered almost negligible.

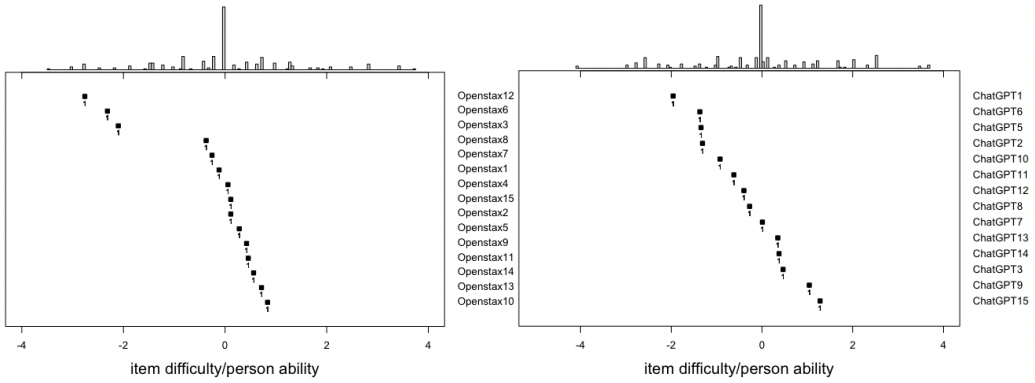


Figure 3: WrightMap for Openstax item (left) and ChatGPT item (right)

A two parameter logistic model was applied to produce the item discrimination parameters. Higher discrimination parameters better differentiate the ability levels of respondents based on the correctness of their answer [16]. A well constructed set of items should have an item discrimination distribution center around 1 but skewed to the right, which suggest moderately high (> 1) discrimination. 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. However, the difference of discrimination between the two groups of items was not statistically significantly separable(t = *-1.14, p = 0.26*), in part due to the relatively small sample size (N = 15).

5 Discussion and Conclusion

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. Out of the remaining questions, we selected 15 for our study. Based on our preliminary analysis, ChatGPT, while given the appropriate prompt, may be better at generating items with locations equally spaced within the ability distribution of respondents and even generating items with higher discrimination power.

Worth discussing is that out of the 15 selected ChatGPT questions, one had to be eliminated from analysis due to a 0% accuracy rate, as it is customary in IRT type analysis [37]. The particular item requires solving quadratic equations, which is not in the scope of the lesson and requires a higher level of ability. 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.

The study still had several limitations. 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. Another unstudied area is how the number of items will influence the result, especially the distribution of ChatGPT generated item difficulties. It is not certain whether the distribution will be more spread out or centered when more items are produced. The current analysis was also restricted in the scope of evaluating items based on their classical psychometric properties. Future work may also focus on other formative assessment concerns, such as the degree to which the generated items cover the source material with respect to concepts and skills.

Acknowledgments

We thank the UC Berkeley Peder Sather Center, the Vice Provost of Undergraduate Education's Micro Grant Program, and the Institute of Cognitive and Brain Sciences Undergraduate Research Funding Program for providing financial support for this work. This study was approved by the UC Berkeley Committee for the Protection of Human Subjects under IRB Protocol 2022-12-15943.

References

[1] Jay Abramson. *College Algebra 2e*. Houston, Texas: OpenStax, 2021.

[2] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[3] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[4] Dhawaleswar Rao CH and Sujan Kumar Saha. "Generation of Multiple-Choice Questions From Textbook Contents of School-Level Subjects". In: *IEEE Transactions on Learning Technologies* 16.1 (2023), pp. 40–52. DOI: 10.1109/TLT.2022.3224232.

[5] Nurul Hapizah Damanik and Yetti Zainil. "The analysis of reading comprehension questions in English textbook by using higher order thinking skill at grade X of SMAN 2 Padang". In: *Journal of English Language Teaching* 8.1 (2019), pp. 249–258.

[6] Matthias von Davier et al. "Evaluating item response theory linking and model fit for data from PISA 2000–2012". In: *Assessment in Education: Principles, Policy & Practice* 26.4 (2019), pp. 466–488.

[7] Christine DeMars. "Incomplete data and item parameter estimates under JMLE and MML estimation". In: *Applied measurement in education* 15.1 (2002), pp. 15–31.

[8] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[9] Kirsten DiCerbo, Valerie Shute, and Yoon Jeon Kim. "The future of assessment in technology rich environments: Psychometric considerations". In: *Learning, design, and technology: An international compendium of theory, research, practice, and policy* (2017), pp. 1–21.

[10] Shane Griffith et al. "Policy shaping: Integrating human feedback with reinforcement learning". In: *Advances in neural information processing systems* 26 (2013).

[11] Sidney H Irvine and Patrick C Kyllonen. *Item generation for test development*. Routledge, 2013.

[12] Seang-Hwane Joo, Philseok Lee, and Stephen Stark. "Evaluating anchor-item designs for concurrent calibration with the GGUM". In: *Applied psychological measurement* 41.2 (2017), pp. 83–96.

[13] Stanley Uros Keller. *Automatic Generation of Word Problems for Academic Education via Natural Language Processing (NLP)*. 2021. arXiv: 2109.13123 [cs.CL].

[14] Ghader Kurdi et al. "A systematic review of automatic question generation for educational purposes". In: *International Journal of Artificial Intelligence in Education* 30 (2020), pp. 121–204.

[15] Wim J van der Linden and Michelle D Barrett. "Linking item response model parameters". In: *Psychometrika* 81.3 (2016), pp. 650–673.

[16] Geofferey N Masters. "Item discrimination: When more is worse". In: *Journal of Educational Measurement* 25.1 (1988), pp. 15–29.

[17] Colleen A McHorney. "Use of item response theory to link 3 modules of functional status items from the asset and health dynamics among the oldest old study". In: *Archives of Physical Medicine and Rehabilitation* 83.3 (2002), pp. 383–394.

[18] J Patrick Meyer and Shi Zhu. "Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating." In: *Research & Practice in Assessment* 8 (2013), pp. 26–39.

[19] Rishabh Kumar Mishra. "Mapping the knowledge topography: a critical appraisal of geography textbook questions". In: *International Research in Geographical and Environmental Education* 24.2 (2015), pp. 118–130.

[20] Robert J Mislevy, Linda S Steinberg, and Russell G Almond. "Focus article: On the structure of educational assessments". In: *Measurement: Interdisciplinary research and perspectives* 1.1 (2003), pp. 3–62.

[21] Steven Moore et al. "Assessing the Quality of Multiple-Choice Questions Using GPT-4 and Rule-Based Methods". In: *European Conference on Technology Enhanced Learning*. Springer. 2023, pp. 229–245.

[22] Huy A Nguyen et al. "Evaluating ChatGPT's Decimal Skills and Feedback Generation in a Digital Learning Game". In: *European Conference on Technology Enhanced Learning*. Springer. 2023, pp. 278–293.

[23] Huy A Nguyen et al. "Towards generalized methods for automatic question generation in educational domains". In: *European conference on technology enhanced learning*. Springer. 2022, pp. 272–284.

[24] Sinan Onal and Derya Kulavuz-Onal. "A Cross-Disciplinary Examination of the Instructional Uses of ChatGPT in Higher Education". In: *Journal of Educational Technology Systems* 0.0 (0), p. 00472395231196532. DOI: 10.1177/00472395231196532. eprint: https://doi.org/10.1177/00472395231196532. URL: https://doi.org/10.1177/00472395231196532.

[25] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *arXiv preprint arXiv:2203.02155* (2022).

[26] Michelle Overman et al. "Textbook questions in context-based and traditional chemistry curricula analysed from a content perspective and a learning activities perspective". In: *International Journal of Science Education* 35.17 (2013), pp. 2954–2978.

[27] Zachary A Pardos and Shreya Bhandari. "Learning gain differences between ChatGPT and human tutor generated algebra hints". In: *arXiv preprint arXiv:2302.06871* (2023).

[28] Zachary A. Pardos et al. "OATutor: An Open-Source Adaptive Tutoring System and Curated Content Library for Learning Sciences Research". In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. CHI '23. Hamburg, Germany: Association for Computing Machinery, 2023. ISBN: 9781450394215. DOI: 10.1145/3544548.3581574. URL: https://doi.org/10.1145/3544548.3581574.

[29] Alec Radford et al. *Improving language understanding by generative pre-training*. 2018.

[30] Alec Radford et al. *Language models are unsupervised multitask learners*. 2019.

[31] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[32] Kumar Shridhar et al. *Automatic Generation of Socratic Subquestions for Teaching Math Word Problems*. 2022. arXiv: 2211.12835 [cs.CL].

[33] Stephen G Sireci and April L Zenisky. "Innovative item formats in computer-based testing: In pursuit of improved construct representation". In: *Handbook of test development*. Routledge, 2011, pp. 343–362.

[34] C David Vale. "Linking item parameters onto a common scale". In: *Applied Psychological Measurement* 10.4 (1986), pp. 333–344.

[35] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[36] Wen-Chung Wang and Cheng-Te Chen. "Item parameter recovery, standard error estimates, and fit statistics of the WINSTEPS program for the family of Rasch models". In: *Educational and Psychological Measurement* 65.3 (2005), pp. 376–404.

[37] Mark Wilson. *Constructing measures: An item response modeling approach*. Taylor & Francis, 2023.

[38] Mark Wilson. "Measuring progressions: Assessment structures underlying a learning progression". In: *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 46.6 (2009), pp. 716–730.

[39] Mark Wilson. "Using the concept of a measurement system to characterize measurement models used in psychometrics". In: *Measurement* 46.9 (2013), pp. 3766–3774.

[40] Marilyn S Wingersky and Frederic M Lord. "An investigation of methods for reducing sampling error in certain IRT procedures". In: *Applied Psychological Measurement* 8.3 (1984), pp. 347–364.

[41] Benjamin D Wright and Mark H Stone. *Best test design*. Mesa press, 1979.

[42] Zihao Zhou et al. *Learning by Analogy: Diverse Questions Generation in Math Word Problem*. 2023. arXiv: 2306.09064 [cs.CL].