

A Feasibility Study on Retrieval Practice Using Large Language Models

Marcus Leong, Jennifer Rose, Lisa Dierker, Antonio Laverghetta Jr.

Wesleyan University

Middletown, CT, USA

{mleong, jrose01, ldierker, alaverghetta}@wesleyan.edu

Abstract

Retrieval practice supports learning but requires educators to build large item banks. We compared LLM-generated and human-written retrieval practice items in an introductory psychology course to test whether LLM items match instructor-written ones in quality. LLM items exhibited overall weaker psychometric properties, suggesting that human supervision may remain necessary during item generation for retrieval practice.

1 Introduction

Retrieval practice (RP) is a well-established pedagogical strategy that helps students learn through repeated memory retrieval (Roediger and Butler, 2011). Its efficacy in supporting long-term retention, particularly when learners must apply new knowledge, has been supported by decades of research in the learning sciences (Karpicke and Roediger III, 2008; Roediger and Butler, 2011; Rowland, 2014; Butler et al., 2017; inter-alia). Typically, instructors integrate RP into their courses by administering repeated, timed, multiple-choice quizzes (MCQs) that require retrieving course-relevant information from memory (Lyle and Crawford, 2011; Roediger III et al., 2011). Yet doing so requires creating large banks of items, which remains a difficult and time-intensive bottleneck.

Large language models (LLMs) have become a prominent tool for the automatic generation of test items (AIG). LLM-generated items have demonstrated strong validity and reliability evidence across educational and psychological constructs (Lee et al., 2026; Martin Kowal et al., 2025; Russell-Lasalandra et al., 2024; Laverghetta et al., 2024; Laverghetta Jr. and Licato, 2023). Indeed, LLM-generated items may be just as psychometrically valid as their human-written counterparts, providing support for using LLMs to reduce the often significant effort required to develop new items.

Yet, to our knowledge, there has been little empirical work examining the psychometric properties of LLM-generated retrieval practice items, leaving it unclear how well LLMs can generate items for this use case.

In this paper, we investigate the validity evidence of LLM-generated retrieval practice items used in a university introductory psychology course, specifically examining whether items generated from LLMs under both zero-shot and few-shot prompting conditions can match the quality of those developed by a course instructor. Rather than the best commercial LLMs, we use models that we believe better represent what is the most accessible to instructors. We generate 30 items for one topic taught in the course, and administer them using an adaptive assessment platform to a sample of more than 700 students. We find that, while LLM-generated items are structurally similar to human-written items, they have overall weaker psychometric properties and are not fully equivalent to items developed by instructors. Our findings offer important insights into the feasibility of deploying LLM-generated items for retrieval practice and emphasize the need for continual supervision by human item writers whenever LLMs are used to generate items.

2 Related Work

2.1 Retrieval Practice in Education

The effectiveness of retrieval practice in education has been documented extensively in prior work (Roediger III and Karpicke, 2006; Roediger and Butler, 2011). This literature has demonstrated that active retrieval of learned concepts from memory enhances long-term retention (Roediger III and Karpicke, 2006), and does so more effectively than passive studying activities (Karpicke and Blunt, 2011). This effect has been demonstrated in both empirical studies (Roediger III and Karpicke, 2006;

Karpicke and Roediger III, 2008; Karpicke and Blunt, 2011) and through meta-analysis of existing work (Rowland, 2014; Adesope et al., 2017). Retrieval practice has thus emerged as a powerful teaching strategy (Dunlosky et al., 2013), whose adoption has consistently been urged (Pashler et al., 2007). However, the integration of retrieval practice into pedagogy has been stymied by the need to develop large-scale RP item banks, which the present work helps address.

2.2 Automated Item Generation

The earliest approaches for automatically generating test items were based on simple text-replacement algorithms that could swap different item stems to create content variations (Bejar et al., 2002; Wolfe, 1977). While effective, this approach suffered from methodological shortcomings, such as a limited ability to create semantically diverse item content and the need to write all stem variations by hand. Advances in generative AI spurred interest in using deep learning models to create test items (von Davier, 2018; Kurdi et al., 2020). Extensive prior work has investigated LLM-based AIG for educational and psychological assessments. Evidence from these studies indicates that LLM-generated items are often comparable to human-written items, and may even *exceed* human items in quality in some contexts (Maertens et al., 2024; Hernandez and Nie, 2023; Laverghetta et al., 2024; Laverghetta Jr. and Licato, 2023), though other studies have reported more mixed results (Law et al., 2025; Martin Kowal et al., 2025). Major testing companies have integrated LLM-generated items into high-stakes assessments, with the Duolingo English Test being a prominent example (von Davier et al., 2024; Naismith et al., 2025).

Despite the substantial body of work on AIG, we know of only two studies specifically focusing on retrieval practice using LLM-generated items. Yusuf (2025) implemented RP with LLM-generated items in a measurement and evaluation course ($n = 64$ students). They found that, compared to a control group that did not use RP, students given LLM RP items had significantly higher final exam performance. An et al. (2025) studied the efficacy of LLM RP items across two college-level data science courses ($n = 120$). While they found that students who were given RP items performed better than those who were given no retrieval practice, they also cautioned that the quality of the LLM-

generated items can vary and still require instructor validation. These studies, while an important foundation for our work, suffer from a key methodological shortcoming: neither included a condition that used human-authored items for RP, and instead only compared RP with LLM-generated items to using no RP at all. This leaves a crucial question unanswered: **are LLM items more effective for retrieval practice than instructor-authored items?** We directly address this limitation by comparing the use of LLM-generated to human-written items, while also using a much larger student sample size to ensure that our findings are robust.

3 Methods

All code and the final set of generated items is available on OSF.¹

3.1 Participants

Retrieval practice items were administered to two cohorts of students enrolled in the same undergraduate introductory psychology course during the Spring 2026 term at a U.S. liberal arts university. The first cohort consisted of undergraduate students ($n = 59$, 54.2% male), while the second consisted of high school students completing the course for college credit ($n = 732$, 33.1% male). To ensure our sample size would be adequate for analysis, and given the small gap in age between the samples, we pooled the data from both cohorts together, resulting in a total $n = 791$. Both cohorts completed the class in-person, and received the same curriculum. Students completed retrieval practice quizzes each week as part of their normal course activities, and were not made aware that some items given to them were generated by LLMs. The study was approved by the local institutional review board.

3.2 Retrieval Practice Procedure

Students completed RP quizzes via an adaptive assessment platform, which delivered them multiple-choice RP items related to the course's topics.² Each RP quiz was personalized to the student, with items drawn from both new and previously introduced material. Items answered incorrectly reappeared in later sessions until the student correctly answered the item, after which they did not appear again. During each quiz, students had a maximum of 10 minutes to complete as many items

¹<https://osf.io/5yrq2/overview>

²<https://rippl.online/>

as possible, with the quiz taking place either at the beginning or end of the class period. Students were required not to consult external sources during each quiz. To detect potential cheating, the platform flagged students who navigated to another tab while completing the quiz, and we dropped any responses that had this flag set. The platform collected students’ responses to each item, timestamps, total number of items attempted, and accuracy rates. Retrieval practice began in the first week of the course. However, quizzes did not include LLM-generated items until the second half of the semester, with a total of six quiz weeks containing LLM items. Unless otherwise noted, we only analyze item responses that come from this period. Using a within-subjects design, students were given a mix of human and LLM items for each of these weeks. We chose not to divide students into human-only or LLM-only item conditions to eliminate variability caused by individual differences in item performance.

3.3 Item Generation

We used GPT-oss-20b (OpenAI et al., 2025), Qwen3-Next-80B-A3B-Instruct (Yang et al., 2025), and GPT-5-mini (Singh et al., 2026) to generate RP items. These models were chosen based on their relatively low cost and our ability to run them on our computing infrastructure, which we reasoned would make them broadly accessible to instructors. GPT-5-mini was queried via the OpenAI API.³ We used transformers (Wolf et al., 2020) to access all other models, and used an NVIDIA RTX Pro 6000 Blackwell GPU for generation. We prompted models to generate 10-20 multiple-choice RP items for a specific topic taught in the course (e.g., sensation and perception). The model was provided with the topic to generate items for, and was instructed to ensure items tested deeper conceptual understanding (rather than memorization) and that items had high discrimination. We instructed models to output items as a JSON object that included the item content, response options, and the correct answer choice. We provide the text generation hyperparameters used in Appendix A and our exact prompt in Appendix B. We used both zero-shot and few-shot item generation. In the few-shot condition, we included items authored by the course instructor as context for the LLM, using items from the same topic as what the

model was prompted to generate.

We dropped all items that did not match the target output schema, including items missing a JSON key or that were missing answer options. To mitigate potential bias, the platform randomized the correct answer position for all items. While we generated items for all topics taught in the course, we ultimately only used items from the “treatment” topic, which covers concepts related to mental health (further details are in Appendix C). This was based on a preliminary review of the items by the course instructor, who found that items from other topics were generally either too similar to existing human items or contained flaws in their design. We therefore chose to analyze only this topic. In total, we generated 93 treatment items that passed preliminary validity checks.

3.4 Item Content Review

Following best practices (An et al., 2025; Laverghetta Jr. and Licato, 2023; Rust and Golombok, 2014), we performed a content review on the generated items to select the ones to use in the study. One of the authors (the course instructor) reviewed the items to select which ones to include. Items were flagged for the following issues: including concepts not covered in the course (“not covered”), requiring simple memorization to answer (“memorization”), or having one or more item writing flaws, such as using jargon unsuitable for the course level, or having multiple plausible interpretations (“poor item”). From this process, we identified 30 LLM-generated treatment items for inclusion in the study. These items were matched against 30 instructor-authored items covering the same topic, which were taken from prior administrations of the course. Table 1 lists how many items came from each model and prompting condition.⁴

4 Results

Model	Zero-shot	Few-shot	Total
gpt-oss-20b	4	5	9
GPT-5 mini	3	4	7
Qwen3-Next-80B	5	9	14
Human-written	–	–	30

Table 1: Summary of LLM items passing content review, broken down by model and prompt type. The total number of human-written items is included for comparison.

³<https://openai.com/api/>

⁴Due to budget constraints, we could not recruit multiple experts to perform the content review.

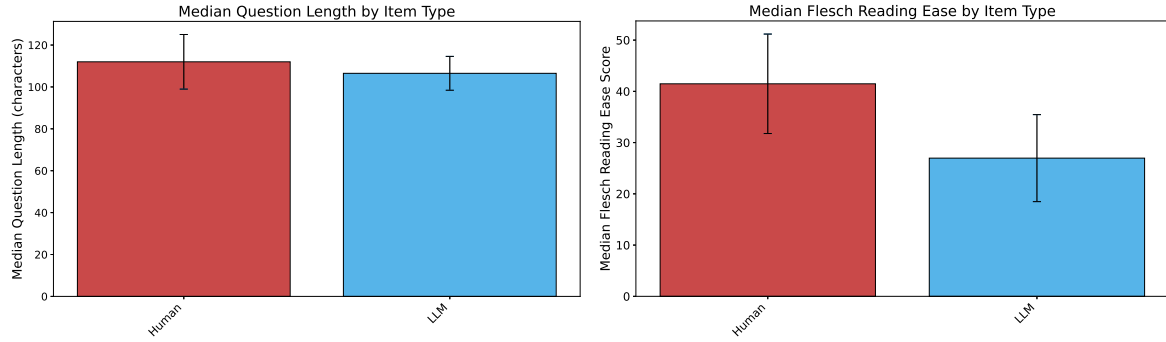


Figure 1: Left: Distribution of median length (number of characters) for human and LLM items. Right: Distribution of median Flesch Reading Ease for human and LLM items. Lower scores indicate items generated under that condition are more difficult to read. Error bars are 95% confidence intervals.

We compare human and LLM items in terms of their structure (item length and readability, Section 4.1), psychometric properties (difficulty and discrimination, Section 4.2), cognitive processing effort (response latency, Section 4.3), and criterion validity (correlation with final exam scores; Section 4.4). Our goal is to determine whether LLM-generated RP items function similarly to their human-written counterparts.

4.1 Item Length and Readability

We first examined whether human and LLM items were similar in structure. Figure 1 compares the median length — measured by number of characters — of the human-written and LLM-generated items. In general, LLM items have a similar length to human-written ones, and there was no significant difference in median length between human and LLM items (Mann-Whitney $U = 483$, $p = 0.63$). We also compared items based on their readability using Flesch Reading Ease (Flesch, 1948), with the results also in Figure 1. LLM items appear to be more difficult to read than human items. Across all conditions, LLM items had a median readability score of 26.97, compared to 41.48 for human items. This difference did not reach statistical significance ($U = 576.5$, $p = 0.06$), though this may be due to the small sample size. Collectively, our results indicate some potential structural differences between human and LLM items, though these differences were not significant.

4.2 Item Parameters

We compared the item parameters of the human and LLM items. Specifically, we used a Bayesian hierarchical longitudinal 3-parameter logistic item response theory model (BHL-3PL; Hans et al., 2023;

Proust-Lima et al., 2022). BHL-3PL provides a framework for modeling learning as it unfolds during repeated practice, and extends the traditional 3PL model by allowing a learner’s latent ability to change as a function of practice opportunity (Wang and Nydick, 2020). This makes a BHL-3PL model suitable for the present analysis.

We fit the model using the `cmdstanr` package (version 0.8.0; Gabry et al., 2025) in R (version 4.5.1; R Core Team, 2025). We dropped any responses flagged for being kicked out of the quiz, and dropped all students and items with 0 variance in scores. We use the data from each student’s first exposure to each item to fit the model, giving us estimates of how difficult and discriminating items were when students initially attempted them. All parameters were initialized randomly using either normal or uniform distributions; further information is provided in Appendix D. To ensure models consistently converged, we fit four BHL-IRT models with different random seeds. Each model used four parallel chains, a maximum tree depth of 12, and an `adapt_delta` of 0.9 to reduce the risk of divergent transitions. To evaluate model convergence stability, we examined the run-to-run stability of posterior predictive check (PPC) for model accuracy and posterior mean parameter estimates. We found that all models converged to a similar predicted mean score, with parameter estimates across runs correlating at above 0.99, indicating good cross-run stability. All models overestimated the true mean scores by less than 0.001, indicating good mean recovery. We then checked item-level convergence for human and LLM treatment item parameters using the $\hat{R}$ statistic, using the `bayesplot` package (version 1.15.0; Gabry et al., 2019). $\hat{R}$ measures the ratio of the average variance

draws within each BHL-3PL chain to the variance of the pooled draws across chains. If chains have not converged to a common distribution, $\hat{R}$ will be greater than one (Gelman et al., 1995). We follow the package’s recommended ratio of below 1.05. All item parameters obtained a $\hat{R} \leq 1.01$, indicating good item-level convergence. We extracted the item parameters from the first model run, for just the treatment topic, using both human and LLM items, and used this data as the basis for our parameter analysis.

We make comparisons between human and LLM item parameters in Figure 2. LLM items are both easier and less discriminating than human-written items. Mean differences in parameter values were -0.19 for discrimination and -1.01 for difficulty when comparing LLM parameters to human ones. We also calculated the posterior probability that LLM items have lower mean parameter values than human items, with this probability being 0.94 for discrimination and approximately 1 for difficulty. These lower parameter values suggest that the LLM items did not function as well for retrieval practice, given that they were both less discriminating and that easier items are less likely to require effortful retrieval.

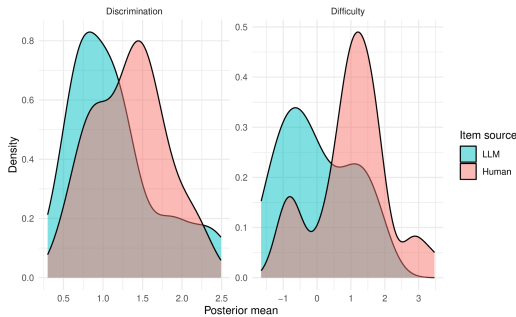


Figure 2: Comparison of difficulty and discrimination parameters for human and LLM items.

4.3 Response Latency

We compared response latencies for treatment items. As response time is often used to index cognitive processing difficulty (Pyc and Rawson, 2009), we reasoned that it would be desirable for the response time for LLM items to be similar to that of human items, to motivate their use as drop-in replacements. We fit a linear mixed effects model to predict log-transformed response time (log-RT), with this transform applied to account for heavy skew in the data. The full model predicted log-RT as a function of the item source (human or LLM)

and whether the response was correct, to account for differences between items answered correctly and incorrectly. We include the student and item IDs as random effects to control for differences in baseline response latency. Additionally, we fit a simpler model that does not include item source, and we compare the full and reduced models using a likelihood ratio test. We removed responses that were less than 1 second or greater than 500 seconds long.

Item source was found to have a very small effect on log-RT (Cohen’s $d = 0.05$, standardized effect estimate), with LLM item responses being 1.8% faster to 10.1% slower than human item responses. The full model did not significantly improve fit over the reduced model ($\chi^2(1) = 1.83, p = 0.18$). Figure 3 graphs the observed response times for correct responses (successful retrieval). The distributions appear similar to each other. In summary, our results suggest that, while LLM items may have a very modest increase in response latency compared to human items, this difference is quite small and did not reach statistical significance.

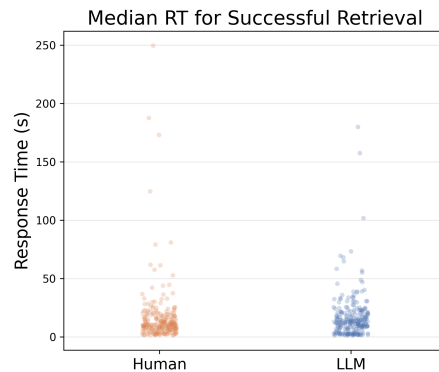


Figure 3: Per-student median response time, broken down by item type, and including only correct answers.

4.4 Criterion Validity

For our final analysis, we examined whether LLM items predicted student final exam performance as well as human items. Given that LLM items had weaker discrimination and may have also had a restricted score range by virtue of their lower difficulty, we expected that they would not be as strong a predictor of cumulative performance as human-written items. We computed each student’s overall accuracy (proportion correct) on the retrieval practice, doing so separately for human and LLM treatment items. We then computed the Pearson correlation between 1) final exam scores and LLM RP

scores, 2) final exam scores and human RP scores, and 3) LLM RP scores and human RP scores. The results are in Figure 4. Both human and LLM RP performance were significant but comparably weak predictors of final exam performance (human $r = 0.20, p = 0.003$, LLM $r = 0.15, p = 0.03$), and the difference in these correlations was not statistically significant (Steiger’s $z = 0.76, p = 0.45$). Human RP scores also correlated with LLM RP scores ($r = 0.41, p < 0.001$), providing evidence of convergent validity. In summary, LLM RP item performance predicted final exam performance, but did not do so as strongly as human RP items, though this difference did not reach statistical significance.

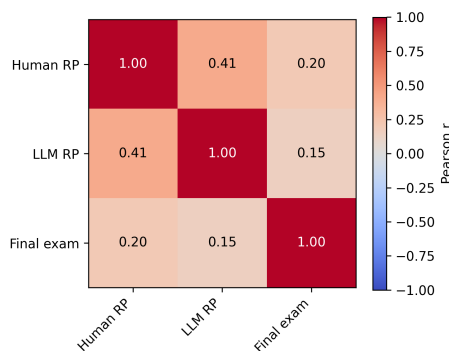


Figure 4: Pearson correlations among human RP scores, LLM RP scores, and final exam scores.

5 Conclusion

In this paper, we analyzed the functioning of LLM-generated retrieval practice items for an introductory psychology course. Unlike prior studies (Yusof, 2025; An et al., 2025), we make a direct comparison between LLM items and those developed by a human instructor for retrieval practice, thus performing a more rigorous assessment of LLM capabilities. Collectively, our results suggest that items generated by models that are more readily accessible to instructors do not yet match the quality of human-written items, though this may not be true for better performing models. LLM items were found to be structurally similar and produce similar behavioral outcomes to human ones, with no significant differences in length or response latency. However, LLM items were not found to be equivalent to human-written items in terms of their psychometric properties, with LLM items having lower discrimination and much lower difficulty.

The patterns observed between item parameters

and response latency warrant closer attention. LLM items were found to be substantially easier than human items, implying that students might answer them more quickly. Yet response latency was similar for both item types. Prior work has found that LLM items often suffer from minor syntactic issues or contain flaws in their design (Meissner et al., 2024; Olney, 2023), which may make the item content more difficult to understand, even if what is being asked is not challenging. This is also consistent with prior work on retrieval practice using LLMs, which noted that many LLM items remain unsuitable for use without modifications by instructors (Yusof, 2025). It is likely that there remained structural issues in the LLM items that were not caught during the content review and which did not affect structural characteristics like length or readability, though we cannot be certain of this without conducting a more detailed analysis.

We note that our study has several limitations that should be kept in mind. Our item generation pipeline relied on simple zero-shot and few-shot learning, which likely doesn’t surface a model’s full generative capabilities. Similarly, we relied on a limited selection of models, none of which were as capable as the best models maintained by leading providers. We chose these models because we believed their smaller size would make them more accessible to instructors, and hence that it would be more informative to examine them rather than leading commercial models. Nevertheless, future work should examine more recent and better performing generations of models to determine if our findings hold. The small item pool used, which covers only one course topic, makes it difficult to draw broad conclusions on how our findings may generalize to other domains. This also attenuated the maximum correlation we expected to observe for the criterion validity analysis, as the exam covered all course topics, not just the treatment topic. Due to how the retrieval practice was set up, the resulting dataset for the LLM items was very sparse, which limited the types of analyses we could run. Future work should conduct additional analyses, including checking for differential item functioning, to determine if LLM items have similar fairness evidence as human ones. Finally, our content review relied on evaluations from only one expert, which may have biased the item selection process.

Retrieval practice is an important tool for enhancing learning and education, yet instructors frequently do not have the time to develop the

item banks necessary to incorporate it into their courses. Although LLMs have the potential to greatly streamline the development of RP items, many questions remain about the efficacy of LLM-generated items in this context, and how they compare to items developed by human experts. By directly comparing LLM items to those developed by instructors, our study advances research at the intersection of retrieval practice and generative AI, and provides an important foundation for the future development of LLM-powered systems for retrieval practice.

Acknowledgments

This work was supported by the National Science Foundation's Improving Undergraduate STEM Education (IUSE) program Award No. 2416471.

References

- Olusola O Adesope, Dominic A Trevisan, and Narayankripa Sundararajan. Rethinking the use of tests: A meta-analysis of practice testing. *Review of educational research*, 87(3):659–701, 2017.
- Yuan An et al. Enhancing student learning with llm-generated retrieval practice questions: An empirical study in data science courses, July 2025. URL <https://doi.org/10.48550/arXiv.2507.05629>. Preprint.
- Isaac I Bejar, René R Lawless, Mary E Morley, Michael E Wagner, Randy E Bennett, and Javier Revuelta. A feasibility study of on-the-fly item generation in adaptive testing. *ETS Research Report Series*, 2002(2):i–44, 2002.
- Andrew C. Butler et al. Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied*, 23(4):433–446, 2017. doi: 10.1037/xap0000142.
- John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, 14(1):4–58, 2013.
- Rudolf Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.
- Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. Visualization in bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2):389–402, 2019.
- Jonah Gabry, Rok Češnovar, Andrew Johnson, and Steve Bronder. *cmdstanr: R Interface to 'CmdStan'*, 2025. URL <https://mc-stan.org/cmdstanr/>. R package version 0.9.0, <https://discourse.mc-stan.org>.
- Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.
- Atharva Hans, Ashish M Chaudhari, Ilias Bilionis, and Jitesh H Panchal. A bayesian hierarchical model for extracting individuals' theory-based causal knowledge. *Journal of Computing and Information Science in Engineering*, 23(3):031011, 2023.
- Ivan Hernandez and Weiwen Nie. The ai-ip: Minimizing the guesswork of personality scale item development through artificial intelligence. *Personnel Psychology*, 76(4):1011–1035, 2023.
- Jeffrey D Karpicke and Janell R Blunt. Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018):772–775, 2011.
- Jeffrey D Karpicke and Henry L Roediger III. The critical importance of retrieval for learning. *science*, 319(5865):966–968, 2008.
- Ghader Kurdi, Jared Leo, Bijan Parsia, Uli Sattler, and Salam Al-Emari. A systematic review of automatic question generation for educational purposes. *International journal of artificial intelligence in education*, 30(1):121–204, 2020.
- Antonio Laverghetta, Simone Luchini, Averie Linnell, Roni Reiter-Palmon, and Roger Beaty. The creative psychometric item generator: a framework for item generation and validation using large language models. In *CEUR Workshop Proceedings*, volume 3810, pages 59–73. CEUR-WS, 2024.
- Antonio Laverghetta Jr. and John Licato. Generating better items for cognitive assessments using large language models. In Ekaterina Kochmar et al., editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.bea-1.34. URL <https://doi.org/10.18653/v1/2023.bea-1.34>.
- Alex KK Law, Jerome So, Chun Tat Lui, Yu Fai Choi, Koon Ho Cheung, Kevin Kei-ching Hung, and Colin Alexander Graham. Ai versus human-generated multiple-choice questions for medical education: a cohort study in a high-stakes examination. *BMC Medical Education*, 25(1):208, 2025.
- Philseok Lee, Mina Son, and Zihao Jia. Ai-powered automatic item generation for psychological tests: A conceptual framework for an llm-based multi-agent aig system. *Journal of Business and Psychology*, 41(1):71–99, 2026.
- Keith B Lyle and Nicole A Crawford. Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, 38(2):94–97, 2011.

- Rakoén Maertens, Friedrich M Götz, Hudson F Golino, Jon Roozenbeek, Claudia R Schneider, Yara Kyrychenko, John R Kerr, Stefan Stieger, William P McClanahan, Karly Drabot, et al. The misinformation susceptibility test (mist): A psychometrically validated measure of news veracity discernment. *Behavior Research Methods*, 56(3):1863–1899, 2024.
- Jaclyn Martin Kowal, Kenzie Hurley Bryant, Dan Segall, and Tracy Kantrowitz. Harnessing generative ai for assessment item development: Comparing ai-generated and human-authored items. *International Journal of Selection and Assessment*, 33(3):e70021, 2025.
- Roy Meissner, Alexander Pögel, Katja Ihsberner, Martin Grützmüller, Silvana Tornack, Andreas Thor, Norbert Pengel, Heinz-Werner Wollersheim, and Wolfram Hardt. Llm-generated competence-based e-assessment items for higher education mathematics: methodology and evaluation. In *Frontiers in Education*, volume 9, page 1427502. Frontiers Media SA, 2024.
- Ben Naismith, Ramsey Cardwell, Geoffrey T. LaFlair, Steven Nydick, and Masha Kostromitina. Duolingo English test: Technical manual. Duolingo research report, Duolingo, July 2025. URL https://duolingo-papers.s3.us-east-1.amazonaws.com/other/technical_manual/DET_technical_manual_2025_07.pdf.
- Andrew M Olney. Generating multiple choice questions from a textbook: Llms match human performance on most metrics. In *CEUR workshop proceedings*. CEUR-WS. org, 2023.
- OpenAI et al. Gpt-oss-120b & gpt-oss-20b model card, August 2025. URL <https://doi.org/10.48550/arXiv.2508.10925>. Preprint.
- Harold Pashler, Patrice M Bain, Brian A Bottge, Arthur Graesser, Kenneth Koedinger, Mark McDaniel, and Janet Metcalfe. Organizing instruction and study to improve student learning. ies practice guide. ncer 2007-2004. *National Center for Education Research*, 2007.
- Cécile Proust-Lima, Viviane Philipps, Bastien Perrot, Myriam Blanchin, and Véronique Sébille. Modeling repeated self-reported outcome data: a continuous-time longitudinal item response theory model. *Methods*, 204:386–395, 2022.
- Mary A Pyc and Katherine A Rawson. Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60(4):437–447, 2009.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2025. URL <https://www.R-project.org/>.
- Henry L. Roediger and Andrew C. Butler. The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1):20–27, 2011. doi: 10.1016/j.tics.2010.09.003. URL <https://doi.org/10.1016/j.tics.2010.09.003>.
- Henry L Roediger III and Jeffrey D Karpicke. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3):249–255, 2006.
- Henry L Roediger III, Pooja K Agarwal, Mark A McDaniel, and Kathleen B McDermott. Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of experimental psychology: applied*, 17(4):382, 2011.
- Christopher A Rowland. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological bulletin*, 140(6):1432, 2014.
- Lara L Russell-Lasalandra, Alexander P Christensen, and Hudson Golino. Generative psychometrics via ai-genie: Automatic item generation and validation via network-integrated evaluation. *PsyArXiv Preprints*, 2024.
- John Rust and Susan Golombok. *Modern psychometrics: The science of psychological assessment*. Routledge, 2014.
- Aaditya Singh et al. Openai gpt-5 system card, 2026. URL <https://doi.org/10.48550/arXiv.2601.03267>. Preprint.
- Alina A von Davier, Andrew Runge, Yena Park, Yigal Attali, Jacqueline Church, and Geoff LaFlair. The item factory. *Machine learning, natural language processing, and psychometrics*, pages 1–26, 2024.
- Matthias von Davier. Automated item generation with recurrent neural networks. *psychometrika*, 83(4):847–857, 2018.
- Chun Wang and Steven W Nydick. On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics*, 45(3):339–368, 2020.
- Thomas Wolf et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://doi.org/10.18653/v1/2020.emnlp-demos.6>.
- John H Wolfe. Reading retention as a function of method for generating interspersed questions. 1977.
- An Yang et al. Qwen3 technical report, May 2025. URL <https://doi.org/10.48550/arXiv.2505.09388>. Preprint.

Ibnatul Jalilah Yusof. Chatgpt-assisted retrieval practice and exam scores: Does it work? *Journal of Information Technology Education: Research*, 24: 008, 2025. doi: 10.28945/5474. URL <https://doi.org/10.28945/5474>.

A Hyperparameters Used for Models

Table 2 lists the hyperparameters used for generation. Hyperparameters were manually tuned to generate items that did not violate the instructions (e.g., using an incorrect number of options) and which, based on manual inspection, appeared relevant to the course content. Any hyperparameters not listed below were left at their defaults.

Model	Hyperparameters
gpt-oss-20b	{'max_new_tokens': 4096, 'temperature': 0.7, 'top_p': 0.9, 'top_k': 50, 'repetition_penalty': 1.1, 'do_sample': True}
Qwen3-Next-80B-A3B-Instruct	{'max_new_tokens': 4096, 'do_sample': False}
GPT-5-mini	{'max_output_tokens': 4096}

Table 2: Hyperparameters used for text generation.

B Prompt Used for Generation

We list our exact prompt for generating items below. The zero-shot variant of the prompt simply omits the section with example items.

You are an expert educational assessment designer specializing in creating high-quality multiple choice questions for college students for an introductory psychology college class. Your questions should test understanding of concepts rather than simple memorization, where the test is a form of retrieval practice.

Make sure the items generated have item discrimination, meaning each question helps differentiate between students who truly understand the topic and those who do not.

The topic for generation is **{lesson}**.

Generate **10-20** multiple choice questions. Each question should have either **2** or **4** response options, **2** if True/False question, **4** otherwise, with **exactly one** correct answer. Make sure incorrect options (distractors) are plausible but clearly wrong to someone who understands the concept.

Make sure that not all the questions have the same correct response options and that there are indeed **4** response options provided if the item generated is supposed to have **4** response options. Do not only produce True/False questions, there must be **4** response options items.

Your output must be **valid JSON**, and **must not contain any text, explanation, commentary, or notes outside the JSON object**.

Output only this JSON structure and nothing else:

```
{
  "questions": [
    {
      "question": "Clear,
well-written question text
here",
      "responses": {
        "A": "First option",
        "B": "Second option",
        "C": "Third option",
        "D": "Fourth option"
      },
      "correct": "A",
      "lesson": "",
      "unit": "",
      "topic": "",
      "difficulty": "medium",
      "subject": ""
    }
  ]
}
```

Field definitions:
- "question": The main question text (clear, unambiguous, college-level)

- "responses": Object with all possible options
- "correct": The correct answer
- "lesson": Leave empty for now ("")
- "unit": Leave empty for now ("")
- "topic": Leave empty for now ("")
- "difficulty": Use "easy", "medium", or "hard"
- "subject": Leave empty for now ("")

Here are some examples for different difficulty levels:
{examples}

IMPORTANT:

- Do **not** include any explanations, reasoning, or text before or after the JSON.
- Do **not** include code fences or markdown.
- Output **only** a single valid JSON object as shown above.
- Make sure to output 10-20 multiple choice questions, even if there are no examples of all different difficulty levels. Make sure to still output questions, do **not** rely on examples.
- Validate that the JSON output is in valid JSON format after generation, and make sure to implement fixes to the JSON format if not.
- Make sure that not all the items have the same correct response option, and that not all the items generated are True/False questions, there must be 4 response option items.

If your answer includes any non-JSON text, delete it before outputting. Output exactly one valid JSON object and nothing else.

C Definition of the Treatment Topic

We used LLM-generated items from the "treatment" topic for analysis. This topic provides a broad survey of mental health treatment across five areas: (1) the history of mental health care, from early asylums through deinstitutionalization; (2) major psy-

chotherapy approaches, including psychodynamic, humanistic/client-centered, behavioral, cognitive, and cognitive-behavioral therapy (CBT), mindfulness, and group formats; (3) methods for evaluating treatment efficacy, including randomized controlled trials, placebo controls, manualized therapy, and meta-analysis; (4) biomedical treatments, including antipsychotic medications, antidepressants (MAO inhibitors, tricyclics, SSRIs), mood stabilizers, anxiolytics, and brain stimulation techniques such as ECT and rTMS; and (5) barriers to treatment access, including insurance, cultural and ethnic disparities, and stigma.

D BHL-IRT Parameter Initializations

Table 3 lists the distributions each BHL-IRT parameter was sampled from. Starting distributions were selected based on plausible values to position the parameters near the center of the parameter space.

Parameter	Distribution	Minimum	Maximum
theta0	normal	0	0.5
learning rate	uniform	0.01	0.08
a (discrimination)	uniform	0.5	1.5
b (difficulty)	normal	0	1
c (guessing)	uniform	0.05	0.25

Table 3: BHL-IRT model parameter initializations.