

It Quacks Like a Response: Evaluating Synthetically-Generated Responses in Assessment Research

Cole Walsh and Rodica Ivan

Acuity Insights Inc., Toronto, ON, Canada

Abstract

This study examines the quality of synthetic open-response data generation for assessment development and psychometric analyses. We use a methodology that treats existing real response data as ground truth and we evaluate the effectiveness of generating synthetic data that has been conditioned on real test taker responses. We also evaluate different LLMs and generation strategies to identify optimal configurations for producing synthetic data that approximate human response data. We find that generating new responses using previous responses from real test takers as contextual input improves both synthetic response *fidelity* and *diversity* compared to simply prompting an LLM to generate responses of specified lengths. Furthermore, we find that increasing LLM temperature and selecting diverse responses from a pool of candidate synthetic responses can improve synthetic response *diversity*. Across configurations, however, synthetic responses resembled typical real responses without reproducing the variation among real test takers, which limits the psychometric purposes they can currently serve.

1 Background

Almost all areas of measurement science, from item parameter and reliability estimation to automated scoring development and cheating detection, depend on response data or evidence about how test takers respond to a given task. Yet empirical data are often “time consuming and costly to collect and [...] often incomplete” (Feinberg and Rubright, 2016). These constraints are especially pronounced during early-stage item development, when researchers may need evidence about item functioning before sufficient human response data are available (Bulut et al., 2024). To overcome this challenge, researchers have long relied on simulated data to augment empirical data and enable faster, cost-controlled, and more directed

research (Feinberg and Rubright, 2016; Harwell et al., 1996).

Historically, simulation studies in educational and psychological measurement have used prespecified psychometric models to simulate IRT parameters and raw/true scores to evaluate model performance, parameter recovery, equating and linking procedures, and other analytic methods under controlled conditions (Baldwin, 2011; Feinberg and Rubright, 2016; Wang and Jin, 2010). Some work moved closer to simulating the response process itself by using automated question-answering systems or virtual test-takers to produce response patterns for multiple-choice and other closed-response items (Byrd and Srivastava, 2022; Uto et al., 2024). Rather than simulating item properties directly, these studies used machine-generated responses as evidence from which item difficulty, discrimination, or related properties could be estimated.

More recent studies extended this approach by prompting large language models (LLMs) to act as test takers and generate responses to closed-response assessment items (Benedetto et al., 2024; Lu and Wang, 2024; Liu et al., 2025). By shifting the object of simulation from parameters to responses, this work opens the possibility of using synthetic response data as a flexible input for downstream psychometric analyses, including item evaluation, difficulty estimation, calibration, linking, equating, and automatic scoring system development.

LLMs have also expanded our capabilities for simulating the response generation process beyond closed-response formats by making it possible to simulate open-ended responses themselves. Recent studies have used LLMs to generate essays (Bevilacqua et al., 2025; Herbold et al., 2023), respond to short-answer tasks (Wei et al., 2025), and simulate student-teacher interactions in conversation-based tutoring and assessment systems (Burleigh et al., 2025; Sonkar et al., 2024).

This expansion is especially important because open-response formats are often used to assess complex forms of reasoning that are difficult to capture with multiple-choice or other closed-response items (Martinez, 1999). As a result, the ability to simulate open-ended response generation is particularly valuable for developing and refining assessments intended to measure higher-order skills. If synthetic responses can approximate real test taker responses, they could be used to prototype assessments before large-scale human response data are available, both improving assessment quality and reducing development times.

2 Aims

The promise of synthetic response data depends on a simple but consequential assumption: that the generated responses are good enough to stand in for real response data for the intended measurement purpose. Although LLMs can now produce large volumes of plausible assessment responses, the ability to generate text at scale does not ensure that the resulting data are useful for measurement. For synthetic responses to support assessment and item development, they must preserve the features of real response data that are relevant to the intended usage of those responses. In other words, synthetic responses must not only sound plausible; they must resemble human responses in that particular assessment context and capture the diversity therein.

This study addresses these issues by examining both the generation and evaluation of synthetic open-response data. Previous efforts to generate synthetic response data investigated a number of strategies including providing rubrics (Wei et al., 2025), prompting for specific failure modes (Lu and Wang, 2024), and conditioning on real responses (Wei et al., 2025; Zelikman et al., 2023). We focus specifically on this latter strategy here and compare the effectiveness of conditioning on real response data relative to context-free approaches. We additionally evaluate different LLMs and generation strategies to identify optimal configurations for generating synthetic data that approximate human response data for assessment use.

While studies often evaluate linguistic and structural characteristics (Wei et al., 2025), as well as psychometric properties (Lu and Wang, 2024; Sonkar et al., 2024) of synthetically generated data, there has been scant attention paid to whether syn-

thetic responses can be generated in a manner that makes them indistinguishable from real data. In this study, we focus on the *fidelity* and *diversity* of synthetic responses (Alaa et al., 2022); that is, the extent to which generated responses are semantically similar to real responses and represent the diversity of the real response space. We use a methodology that treats existing real response data as ground truth, setting a standard that can be adopted in future studies to evaluate the quality of synthetic response generation processes before their use with untested assessments and items.

3 Samples

We used data collected using a formative open-response situational judgment test (SJT) designed to measure personal and professional skills. The assessment, which is typically taken by undergraduate and graduate students in business, engineering, and health sciences fields, presents test takers with a series of hypothetical scenarios and asks how they would respond given the situation. While the assessment typically comprises 30 items and four measured dimensions (i.e., sub-scales), we focused exclusively on eight items measuring *interpersonal skills* in this study. Below is a summary of a sample scenario and associated item related to interpersonal skills:

Scenario: Kendra and Alex, two close friends co-leading a major group project, are clashing over Kendra pushing a strict schedule to meet deadlines. Alex wants more group input on task division and worries that Kendra's approach could undermine team trust and strain their friendship.

Item: How could the co-leads, Alex and Kendra, have collaborated more effectively with the group to create a clear and realistic project plan?

SJTs like this one are particularly well suited for exploring synthetic response generation as the nature of the items promotes a constrained diversity in responses. There are myriad ways to approach the example item above, for example, but the contextualized scenario constrains the response space such that responses are anchored to the specific dilemma (Lievens et al., 2008; Whetzel and McDaniel, 2009). This constraint differentiates SJTs

from more open-ended tasks like personal essays, for instance, where the space of possible responses is much larger (Alvero et al., 2021), making synthetic response generation a more feasible goal.

For this experiment we selected responses from 100 test takers for all eight items. Test takers consented to the use of their data for research purposes before starting the assessment. We selected from the entire pool of test takers by first excluding any test takers who did not respond to all eight items, then stratified by mean response length across all eight items and selected one test taker at random from each percentile. This approach ensured that a diverse representation of response profiles were included in the sampled dataset.

4 Methods

4.1 Synthetic Response Generation

As mentioned above, we hypothesized that we could improve the quality of synthetically generated data by directing LLMs to generate responses that were aligned with specific response patterns. To evaluate this, we conditioned on real test taker responses by incorporating them as part of the context provided to the LLMs. Of the eight items included in this study, we selected seven items at random and, for each test taker, created a simulated counterpart “generative test taker”, which consisted of an LLM given the real test taker’s responses to the seven items as context. Fig. 1 gives a minimum reproducible example of a prompt used for a “generative test taker”. We used these “generative test takers” to simulate responses to the remaining held-out item that was not used in their construction. Using our full complement of 100 generative test takers, we generated 100 synthetic responses to the held-out item. This process, described in greater detail in Fig. 2, created one synthetic response per generative test taker on the held-out item that was meant to reflect each test taker’s response patterns on the other items and simulate how the real test taker might respond to the held-out item.

We compared this generation procedure with a baseline approach that prompted LLMs to generate responses of specified length, but did not provide any further instruction or specification as to the characteristics that the generated responses should possess. By comparing this generation procedure with the “generative test taker” procedure described above, we were able to evaluate the effectiveness of providing real test takers’ responses as context over

and above simply prompting an LLM to generate responses of varying lengths. We additionally experimented with LLMs using different temperature settings and investigated creating multiple parallel sets of synthetic responses to increase diversity of selected responses.

We conducted four experiments in total, described below:

1. *Wordcount Instruction*: For each real test taker, we prompted an LLM to “respond in approximately {wordcount} words,” where {wordcount} was the mean response length for the specified real test taker across their responses.
2. *“Generative Test Takers”*: For each real test taker, we created a counterpart “generative test taker”, a prompt containing responses from the real test taker as context for generating responses to new items.
3. *Temperature*: Using the same “generative test takers” from Experiment 2, we experimented with GPT-4o mini and temperature settings of 0.1 and 1.5¹. Temperature is a hyperparameter that controls the randomness and creativity of model generations, so we hypothesized that higher temperature settings would lead to more diverse data.
4. *Parallel Datasets*: Using the same “generative test takers” from Experiment 2, we created five sets of synthetic responses, then for each test taker we selected the synthetic response that was least similar to the other synthetic responses from the same set (i.e., smallest maximum intra-set similarity; see below). This procedure prioritized selecting the most diverse data possible across multiple parallel runs.

For each experiment (other than Experiment 3), we evaluated three LLMs: GPT-4o mini and GPT-5.4 from OpenAI, and Anthropic Claude Haiku 4.5. These LLMs were chosen to represent a range of model sizes from two different providers. We manipulated temperature within a single model: GPT-5.4 does not expose the parameter, and although Haiku 4.5 does, OpenAI and Anthropic define different ranges, confounding setting with

¹We used a maximum temperature of 1.5 because we noted generation of non-english characters at temperatures above this.

Prompt Structure (Minimal Example)

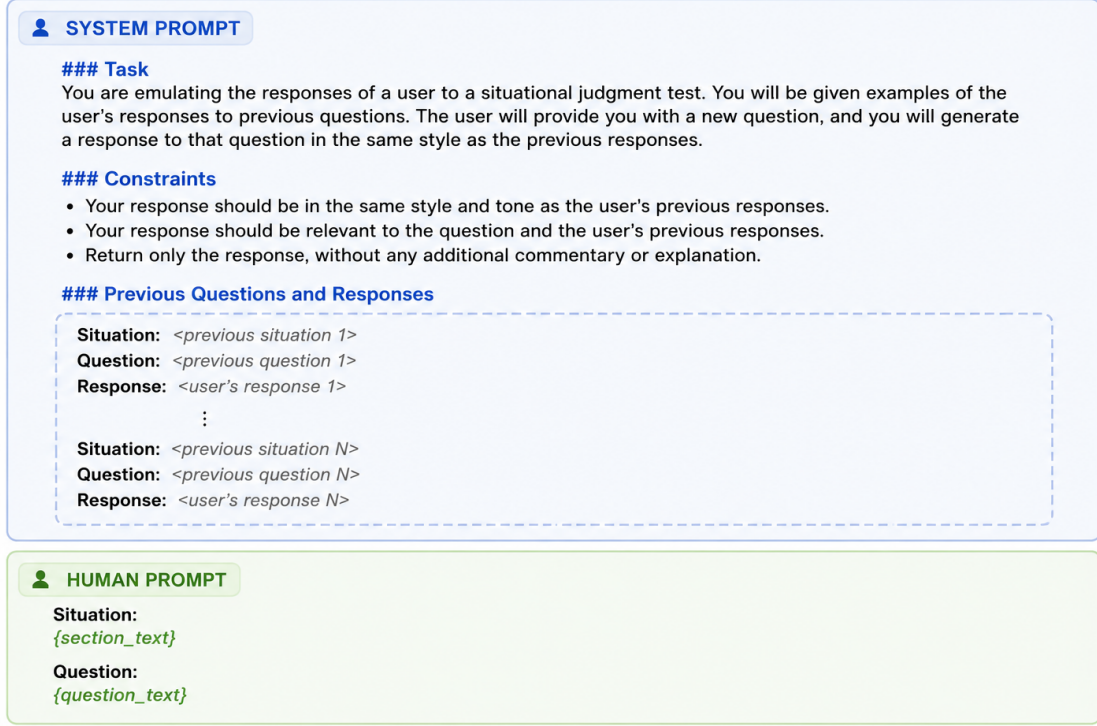


Figure 1: Minimal reproducible example of prompt used to produce synthetic responses using a test taker's responses to previous items as context when generating a response to the new item.

scale. GPT-4o mini admitted the widest range at the lowest cost.

4.2 Evaluation

We evaluated the quality of synthetically generated responses relative to their counterparts in our real dataset through two components (Alaa et al., 2022):

1. *Fidelity*: similarity between synthetic and real responses
2. *Diversity* of synthetic responses

We operationalized these components by examining the cosine similarity between pairs of responses. Cosine values depend on embedding-space geometry (Steck et al., 2024; Timkey and Van Schijndel, 2021), however, so we never interpret values in isolation. Every statistic is benchmarked against the real response dataset and all claims are comparative.

When measuring *fidelity*, we had ground truth data in the form of real responses from the same test takers whose responses we used to create “generative test takers”. This ground truth data allowed us to evaluate the similarity between real responses and synthetically generated responses on the held

out item. We evaluated *fidelity* in two ways: mean paired similarity (MPS) and mean maximum inter-set similarity (MMS_{inter}). For MPS, for each test taker we computed the cosine similarity between their embedded real response vector and the embedded synthetic response vector from the corresponding “generative test taker”. We then calculated the mean across all test takers. Letting $\mathcal{R}$ be the set of real response vectors, r_i , and $\mathcal{G}$ be the set of generated response vectors, g_j , we have:

$$\text{MPS} = \frac{1}{N} \sum_{i=1}^N \cos(r_i, g_i), \quad (1)$$

where i denotes the i^{th} test taker and N is the total number of vectors in the set.

For MMS_{inter}, we computed the cosine similarity of each embedded real response vector with its closest neighbor among the 100 synthetically generated responses (i.e., maximum inter-set similarity), then took the mean of these values:

$$\text{MMS}_{\text{inter}} = \frac{1}{N} \sum_{i=1}^N \max_{j \in \{1, \dots, N\}} \cos(r_i, g_j). \quad (2)$$

As opposed to MPS, which only considers the similarity between real responses from a test taker

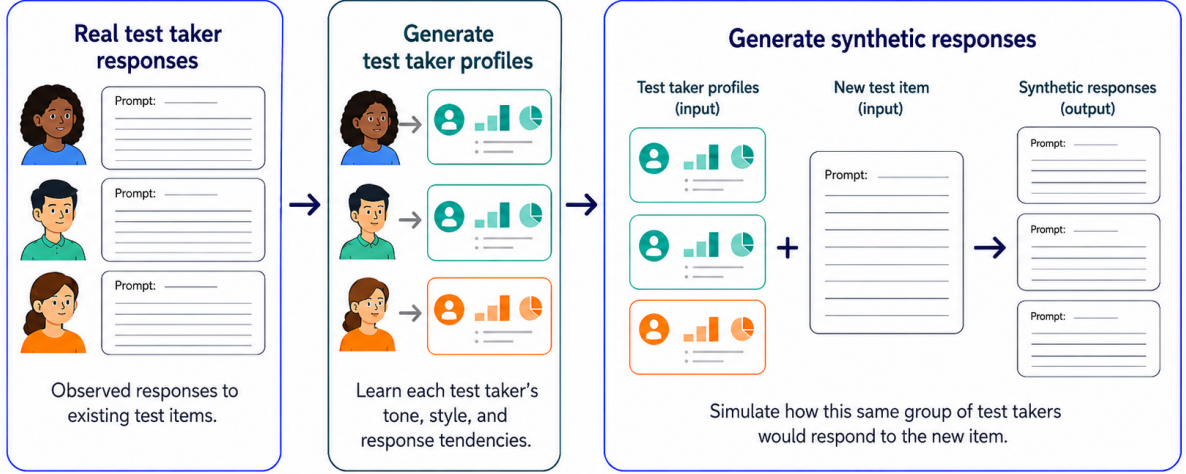


Figure 2: Synthetic response generation procedure. Responses from test takers to existing items are treated as profiles and used for in-context learning with instructions for the LLM to mimic the test taker’s tone, style, and response tendencies. For a new test item, a synthetic response is generated for each profile, simulating how the corresponding test taker might respond to the item.

and the synthetic response generated for their corresponding counterpart “generative test taker”, MMS_{inter} considers all possible pairs of real and synthetic responses. Thus, MMS_{inter} captures cases where the synthetically generated response most similar to a real response comes from a “generative test taker” corresponding to a different test taker than the one that provided the real response.

We evaluated synthetic response *diversity* through the mean maximum intra-set similarity:

$$MMS_{intra} = \frac{1}{N} \sum_{i=1}^N \max_{\substack{j \in \{1, \dots, N\} \\ j \neq i}} \cos(x_i, x_j), \quad (3)$$

where x_i is a vector in either $\mathcal{R}$ or $\mathcal{G}$. While we use MMS_{inter} to evaluate the similarity between real and synthetic responses, MMS_{intra} allows us to compare the similarity between responses within the same set: i.e., we compare real responses to other real responses, and synthetic responses to other synthetic responses, obtaining a measure of overlap between responses in the dataset. When all responses in a dataset are identical, $MMS_{intra} = 1$; the closer this value is to 1 the higher the degree of overlap between responses in the dataset.

When evaluating synthetic response *fidelity* we specifically compared MPS and MMS_{inter} to $MMS_{intra, real}$. If we were generating high-*fidelity*

synthetic data, then we would expect each real response to be more similar to a synthetic response than another real response from a different student (i.e., $MMS_{inter} > MMS_{intra, real}$). When evaluating synthetic response *diversity*, we were not interested in achieving low or high diversity, *per se*. Rather, we sought a diversity in the set of synthetically generated responses that mirrored that observed in the dataset of real responses, indicating that the two datasets were similarly diverse. For this, we compared the values (with standard deviations) of $MMS_{intra, generated}$ and $MMS_{intra, real}$. If we were creating appropriately diverse synthetic data then we would expect these values and standard deviations to be similar. We used OpenAI’s “text-embedding-3-small” model throughout for all embedding operations.

5 Results

Table 1 shows the MPS, $MMS_{intra, generated}$, MMS_{inter} , and mean wordcount for datasets of synthetically generated responses for each model in each experiment discussed in Sec. 4.1, as well as the $MMS_{intra, real}$ and mean wordcount for the dataset of real responses. For the real dataset, we find $MMS_{intra, real} = 0.71$ with all individual responses having similarities greater than 0.50 with at least one other response in the dataset. These

results reflect the constrained diversity discussed in Sec. 3: responses are generally similar to each other as they are all anchored by the same item context, yet we still observe meaningful diversity between responses. Below, we discuss results from each of the four experiments individually, then examine results across experiments, highlighting broader trends and observations.

5.1 Experiment 1: Wordcount Instruction

We find that LLMs accurately follow instructions to generate responses with specified wordcounts. Responses in the real dataset had a mean length of 42 words, while all three models tested in this experiment produced responses with a mean length between 42 and 44 words. The MPS and MMS_{inter} values indicate, however, that across all three models real responses are, on average, more similar with other real responses than responses from our synthetic datasets. GPT-5.4 generally produces synthetic responses that are more similar to real responses than the other models when measured by both metrics, though the effect size is small (Cohen’s $d < 0.25$).

All three models produce very homogeneous data with MMS_{intra} values of at least 0.89. GPT-5.4, in particular, while producing responses that are generally most similar to the real dataset, also produces incredibly homogeneous data. The MMS_{intra} value of 0.94 indicates that for each response in the synthetically generated dataset, there is another synthetically generated response with cosine similarity of 0.94, on average. Taking all these results together, we observe that these models produce synthetic responses that are diverse in terms of response length, but largely semantically the same.

5.2 Experiment 2: Generative Test Takers

We find that generating responses using previous test takers’ responses as context improves synthetic response *fidelity* and *diversity*. Compared to Experiment 1, for each LLM, we observe improvements in both MPS and MMS_{inter} of at least 0.05 and 0.04 points, respectively, corresponding to moderate effect sizes ($d > 0.37$) and indicating that generated data more closely resembles real data. Similarly, while we observe no change in MMS_{intra} for GPT-4o mini, this metric did decrease by 0.02 ($d = 0.42$) and 0.04 ($d = 0.88$) points for Haiku 4.5 and GPT 5.4, indicating that generated responses are more diverse.

We note, however, that, on average, generated

responses are longer than both responses in the real dataset and responses generated in Experiment 1. To investigate this discrepancy, we experimented with a variation of the prompt shown in Fig. 1 that included an additional instruction equivalent to the one used in Experiment 1, effectively combining “generative test takers” with response length instructions. The results, reported in Table 1 as Experiment 2a, indicate that combining instructions produces responses that are of more similar length to real responses and are more diverse, as we observe a decrease of at least 0.02 points ($d > 0.30$) in MMS_{intra} for all three LLMs. Conversely, generated responses are less semantically similar to real responses as we observed drops in both MPS and MMS_{inter} for both GPT-4o mini and Haiku 4.5. GPT-5.4, however, maintained similarity between real and synthetic responses while decreasing average synthetic response length and improving synthetic response *diversity*.

5.3 Experiment 3: Temperature

Adjusting the LLM temperature had negligible effect on synthetic response *fidelity* with similarly observed MPS and MMS_{inter} across experiments 2 and 3 ($d \leq 0.10$). As expected, we did, however, observe large shifts in response *diversity* as measured by MMS_{intra} : using a lower temperature produced *less* diverse responses ($d = 0.67$), while using a higher temperature produced *more* diverse responses ($d = 0.61$). These results indicate that increasing LLM temperature may be one avenue for improving synthetic response *diversity* without sacrificing response *fidelity*.

5.4 Experiment 4: Parallel Datasets

By repeating the synthetic response generation process multiple times and selecting the most distinct response for each generative test taker, we improved response *diversity* without sacrificing response *fidelity*. Across all models and configurations, MMS_{intra} decreased by at least 0.02 points ($d > 0.36$) from Experiment 2 or 3 where the corresponding model and configuration were used, while MPS and MMS_{inter} shifted by no more than 0.01 points ($d \leq 0.10$).

Across all experiments, running GPT-4o mini with a temperature of 1.5 multiple times in parallel and selecting the most distinct response per generative test taker produced the most diverse synthetic responses, though even in this case we failed to recreate the diversity observed in real re-

Configuration	Mean Maximum Similarity Intra-Set	Inter-Set	Mean Paired Similarity	Mean Wordcount
Real	0.71(0.08)	–	–	42(29)
<i>Experiment 1: Wordcount Instruction</i>				
GPT-4o mini	0.89(0.07)	0.65(0.11)	0.53(0.14)	42(28)
Haiku 4.5	0.91(0.03)	0.64(0.10)	0.53(0.12)	43(27)
GPT-5.4	0.94(0.05)	0.66(0.10)	0.56(0.13)	44(30)
<i>Experiment 2: Generative Test Takers</i>				
GPT-4o mini	0.89(0.06)	0.69(0.10)	0.58(0.13)	67(26)
Haiku 4.5	0.89(0.06)	0.70(0.10)	0.58(0.13)	82(48)
GPT-5.4	0.90(0.04)	0.70(0.10)	0.62(0.12)	60(38)
<i>Experiment 2a: Generative Test Takers + Wordcount Instruction</i>				
GPT-4o mini	0.85(0.06)	0.68(0.11)	0.55(0.14)	42(29)
Haiku 4.5	0.87(0.07)	0.68(0.10)	0.54(0.12)	43(33)
GPT-5.4	0.88(0.06)	0.70(0.10)	0.61(0.12)	46(32)
<i>Experiment 3: Temperature (GPT-4o mini)</i>				
$T = 0.1$	0.93(0.06)	0.70(0.10)	0.59(0.13)	66(29)
$T = 1.5$	0.85(0.07)	0.70(0.10)	0.57(0.13)	74(43)
<i>Experiment 4: Parallel Datasets</i>				
GPT-4o mini ($T = 1$)	0.86(0.06)	0.69(0.10)	0.58(0.13)	64(29)
GPT-4o mini ($T = 1.5$)	0.81(0.07)	0.70(0.10)	0.56(0.13)	70(44)
Haiku 4.5	0.87(0.05)	0.70(0.10)	0.58(0.13)	82(48)
GPT-5.4	0.87(0.04)	0.71(0.10)	0.61(0.12)	59(38)

Table 1: Results from each experiment. We report baseline intra-set mean maximum similarity among responses in the real dataset for benchmarking purposes. Standard deviations are reported in parentheses for each statistic.

sponses (difference in $MMS_{\text{intra}} = 0.10$; $d = 1.33$). Though we failed to produce synthetic responses that mirrored real data in terms of *diversity* we were able to generate high *fidelity* data. All models across experiments 2-4, produced synthetic datasets with MMS_{inter} of at least 0.68, within 0.03 ($d < 0.34$) of $MMS_{\text{intra, real}}$, indicating that, on average, for any given real response there is generally a synthetic response that is as similar to it as any other real response. Further, for GPT-5.4 in particular, we see high MPS values, indicating that the synthetic response created for a generative test taker is similar to the actual response provided by the test taker.

6 Conclusions

We examined two elements of synthetic response quality, *fidelity* and *diversity*, across multiple models and generation strategies for an open-response SJT item. Treating real test taker responses to the same item as ground truth provides a directly evaluable standard that can be applied to a generation procedure before it is used on untested items.

Conditioning on real test taker responses im-

proved both *fidelity* and *diversity* relative to prompting for a specified response length, and adding a length instruction on top of conditioning improved *diversity* further. Specifying response characteristics alongside conditioning may therefore yield stronger synthetic data than conditioning alone. Raising LLM temperature and generating multiple parallel sets each improved *diversity* while preserving *fidelity*, whereas lowering temperature produced more homogeneous responses. Because recent commercial releases have deprecated the temperature parameter, the practical reach of this finding may be limited.

7 Limitations and Future Work

Across all experiments we produced high-*fidelity* synthetic responses that resembled typical real responses, but real test takers varied from one another far more than our synthetic responses did. This determines what synthetic data can support. Uses requiring typical, on-construct responses, such as screening a new item or exercising a scoring pipeline before pilot data exist, are plausibly within reach. Uses that depend on the full range of real

responses, such as rubric and rater-training development, augmenting sparse score points, or item parameter estimation, are not, because those uses depend on precisely the variation we failed to reproduce. Combining strategies narrowed this gap without closing it.

These conclusions also rest on a single, untested operationalization: responses to an SJT item share a scenario and much of their vocabulary, compressing them into a narrow semantic space, and we do not know whether the differences we report are meaningful to human readers. The measures capture whether responses say similar things, not whether they would be scored the same way or reflect the same reasoning. The evidence is likewise narrow (i.e., one held-out item from one SJT) and because SJT scenarios constrain the response space (Sec. 3), *fidelity* should degrade and the *diversity* gap widen on less anchored formats; replication across items and assessments would establish where the approach holds.

Future work should also validate the measures themselves against human similarity judgments and psychometric criteria by scoring real and synthetic responses with a common rubric and comparing score distributions, item statistics, and reliability. Overcoming the sycophancy that made poor responses difficult to generate (Do et al., 2026) may require fine-tuning on real test taker responses, or open-weight models, which retain temperature control and, with less alignment tuning, may better reproduce the weakest responses.

The procedure’s value lies in its application to new, untested items: once verified on items with existing responses, a standard set of “generative test takers” can simulate how the same population would respond where no data exist, mirroring the train-evaluate logic of machine learning workflows.

Acknowledgments

We would like to acknowledge Gill Sitarenios and Josh Moskowitz for thoughtful discussions on the ideas presented in this manuscript.

References

Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela Van Der Schaar. 2022. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International conference on machine learning*, pages 290–306. PMLR.

AJ Alvero, Sonia Giebel, Ben Gebre-Medhin, Anthony Lising Antonio, Mitchell L Stevens, and Benjamin W Domingue. 2021. Essay content and style are strongly related to household income and sat scores: Evidence from 60,000 undergraduate applications. *Science advances*, 7(42):eabi9031.

Peter Baldwin. 2011. A strategy for developing a common metric in item response theory when parameter posterior distributions are known. *Journal of Educational Measurement*, 48(1):1–11.

Luca Benedetto, Giovanni Aradelli, Antonia Donvito, Alberto Lucchetti, Andrea Cappelli, and Paula Buttery. 2024. Using llms to simulate students’ responses to exam questions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11351–11368.

Marielena Bevilacqua, Kezia Oketch, Ruiyang Qin, Will Stamey, Xinyuan Zhang, Yi Gan, Kai Yang, and Ahmed Abbasi. 2025. When automated assessment meets automated content generation: Examining text quality in the era of gpts. *ACM Transactions on Information Systems*, 43(2):1–36.

Okan Bulut, Guher Gorgun, and Bin Tan. 2024. Item difficulty and response time prediction with large language models: An empirical analysis of usmle items. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 522–527.

Tyler Burleigh, Jing Chen, and Kristen Dicerbo. 2025. Pre-pilot optimization of conversation-based assessment items using synthetic response data. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 61–68.

Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130.

Heejin Do, Shashank Sonkar, and Mrinmaya Sachan. 2026. Simulating students or sycophantic problem solving? on misconception faithfulness of llm simulators. *arXiv preprint arXiv:2605.12748*.

Richard A Feinberg and Jonathan D Rubright. 2016. Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35(2):36–49.

Michael Harwell, Clement A Stone, Tse-Chi Hsu, and Levent Kirisci. 1996. Monte carlo studies in item response theory. *Applied psychological measurement*, 20(2):101–125.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific reports*, 13(1):18617.

- Filip Lievens, Helga Peeters, and Eveline Schollaert. 2008. Situational judgment tests: A review of recent research. *Personnel Review*, 37(4):426–441.
- Yunting Liu, Shreya Bhandari, and Zachary A Pardos. 2025. Leveraging llm respondents for item evaluation: A psychometric analysis. *British Journal of Educational Technology*, 56(3):1028–1052.
- Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM conference on learning@ Scale*, pages 16–27.
- Michael E Martinez. 1999. Cognition and the question of test item format. *Educational Psychologist*, 34(4):207–218.
- Shashank Sonkar, Xinghe Chen, Myco Le, Naiming Liu, Debshila Basu Mallick, and Richard Baraniuk. 2024. Code soliloquies for accurate calculations in large language models. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 828–835.
- Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, pages 887–890.
- William Timkey and Marten Van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546.
- Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2024. Question difficulty prediction based on virtual test-takers and item response theory. In *EcalLAC@ AIED*.
- Wen-Chung Wang and Kuan-Yu Jin. 2010. Multilevel, two-parameter, and random-weights generalizations of a model with internal restrictions on item difficulty. *Applied Psychological Measurement*, 34(1):46–65.
- Yuchen Wei, Dennis Pearl, Matthew Beckman, and Rebecca J Passonneau. 2025. Concept-based rubrics improve llm formative assessment and data synthesis. *arXiv preprint arXiv:2504.03877*.
- Deborah L Whetzel and Michael A McDaniel. 2009. Situational judgment tests: An overview of current research. *Human Resource Management Review*, 19(3):188–202.
- Eric Zelikman, Wanjing Ma, Jasmine Tran, Diyi Yang, Jason Yeatman, and Nick Haber. 2023. Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2205.