

Rethinking Pilot Data: Evaluating LLM Synthetic Data for Scale Development

Margarita Olivera-Aguilar, Samuel H. Rikoon, Paul D. Bailey, Michael B. Kruse, Kamal Middlebrook
American Institutes for Research

Abstract

This study evaluated whether LLM-synthetic data can support K-12 instrument development. We generated LLM-synthetic datasets under various prompt conditions for two surveys and one assessment. Our findings indicate that while some prompt conditions successfully reproduced the overall latent structure of the instruments, recovery of item-level parameters was generally inadequate.

1 Introduction

Developing K–12 assessments and surveys meeting rigorous standards (AERA, APA, and NCME, 2014) is resource intensive but essential for understanding what students know and can do. Rapid advances in AI have led researchers to use synthetic data (AI-produced datasets resembling real data) in educational and behavioral studies (e.g., Benedetto et al., 2024; Bisbee et al., 2024; Motoki et al., 2023). In educational assessment, Liu et al. (2025) and Säuberli et al. (2025) reported moderate to strong correlations between item parameters estimated from LLM responses and those estimated from human responses in some academic subjects, but they also showed important limitations, including restricted variability in respondent proficiency. Synthetic data could help streamline instrument development and reduce costs, but important questions remain. Can LLM synthetic data reproduce the psychometric properties observed in human responses? Do synthetic datasets introduce systematic bias across student groups? How can synthetic data be used responsibly as part of validity evidence? These questions are relevant for both self-report surveys and academic assessments. In survey research, LLMs

are often asked to generate responses that reflect people's social, emotional, or behavioral characteristics (which might differ from what is represented predominantly in LLM training data). In academic assessments, the challenge is different: LLMs must simulate how students with varying levels of knowledge and different test-taking strategies respond to content-based (i.e., without simply relying on the LLM's vast knowledge base to select the correct answer invariably). Academic assessments also provide an opportunity to examine whether LLMs can reproduce common testing patterns, such as differences in item difficulty, student misconceptions, guessing, and careless responding.

Our primary aim is to evaluate the viability of LLM-generated synthetic data as a tool for supporting early-stage psychometric analyses. Our intention is not to replace human validation samples, but rather to evaluate the use of synthetic data as a tool for refining measures before large-scale data collection. To do so, we tested multiple prompts using ChatGPT 4.0 with the following goals:

1. Determine whether synthetic datasets produce realistic item response distributions,
2. Examine whether synthetic data produces confirmatory factor analysis (CFA) model fit consistent with human data,
3. Compare synthetic and human datasets with respect to patterns of localized CFA misfit, and
4. Evaluate correspondence between item parameter estimates in human vs. synthetic data.

A second methodological goal is to develop a more systematic framework for evaluating synthetic data. Drawing on Monte Carlo simulation methods, we generate and analyze multiple synthetic datasets under each prompting condition rather than relying on a single dataset. This approach allows us to evaluate the stability, consistency, and reproducibility of findings across repeated datasets.

2 Sample

We report findings from K–12 student samples collected using two surveys and one academic subject assessments: (1) the BESSI self-management scale with five response options (Soto et al., 2022; N=478 middle school students), (2) a leadership scale (“LEADER”; N=750 intermediate/middle school students) with four response options, and (3) a science assessment (“SCIENCE”, N=750 4th grade students) with two response options (correct/incorrect). We have shielded the names of the latter two measures to comply with contractual obligations.

3 Method

3.1 Analysis of human data

We used the lavaan R package (Rosseel, 2012) to fit one-factor CFA models to the human data, treating BESSI items as continuous and LEADER and SCIENCE items as categorical. We constructed two versions for each instrument: (1) “well-fitting” versions (i.e., BESSI-9, LEADER-6, and SCIENCE-12, with 9, 6 and 12 items, respectively) and (2) “poorly fitting” versions (i.e., BESSI-13, LEADER-9, and SCIENCE-16; 13, 9 and 16 items each, with items added from other parts of each larger instrument. Unlike BESSI, whose factor structure and item properties have been established across multiple studies, LEADER and SCIENCE reflect the ambiguity typical of scale development, where the underlying structure is less certain and some items may perform differently across samples. For example, LEADER-6 was formed by items from three different subscales that nevertheless resulted in adequate fit to a unidimensional model; and although the 12-item SCIENCE form showed adequate model fit, several items had loadings at or below .20 in the human sample. This design allowed us to examine whether synthetic data

could reproduce psychometric patterns under both well-established and more uncertain measurement conditions.

3.2 LLM synthetic data generation

Using ChatGPT 4.0, we generated synthetic datasets matching the human sample sizes for each instrument. For each respondent, the model received demographic information and survey items and was instructed to generate plausible responses. For the survey instruments, we tested 29 prompting conditions that varied the information provided to the model, including demographic characteristics, respondent personas, trait-level descriptions, response-style tendencies, item-level response probabilities, and instructions about response consistency.

For SCIENCE, we developed 33 prompts based on the set of BESSI and LEADER survey prompts but adapted specifically for academic assessments. Because responses to multiple-choice test items depend on factors such as content knowledge, item difficulty, misconceptions, and guessing, SCIENCE prompts incorporated information such as demographic profiles, expected performance levels, item difficulty, common misconceptions, test-taking effort, attention, and response strategies. Some conditions also included answer keys or expected percent-correct information. When answer keys were provided, prompts explicitly instructed ChatGPT not to simply select the correct answer, but instead to respond as a student with the specified characteristics would be likely to respond.

We initially generated 20 datasets per condition and retained those that most closely reproduced the response distributions observed in the human data. Following a Monte-Carlo simulation approach, for the selected conditions, we then generated 500 datasets per condition to reduce the likelihood that findings are driven by random variation or idiosyncratic characteristics of single or small groups of synthetic dataset(s).

3.3 LLM synthetic data analysis

Consistent with common survey and assessment development practices, our analytic procedure focused on evaluating dimensionality and identifying opportunities for scale refinement. We used the lavaan R package (Rosseel, 2012) to fit a one-factor CFA model to each synthetic dataset to assess the intended factor structure. We then fit a

second model removing items with loadings below .20. We used the same models specifications in the human and LLM synthetic data (e.g., model identification constraints, estimator) to ensure the comparability of results.

Analyses were conducted on the final models (i.e., after excluding items with factor loadings below .20). First, we evaluated the fit of a unidimensional CFA model for each dataset using the Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA). We expected acceptable fit for the well-fitting test forms and poorer fit for the extended forms in which additional items were intentionally introduced to induce model misfit. Second, we examined CFA modification indices to determine whether the added misfitting items were consistently identified as sources of local misfit. Finally, we evaluated item parameter recovery by comparing synthetic and human item parameter estimates using four complementary metrics: (a) **absolute difference (AD)**, calculated as the absolute discrepancy between the human and synthetic estimates; (b) **relative difference (RD)**, calculated as the difference between the synthetic and human estimates divided by the human estimate; (c) **mean absolute relative difference (MARD)**, calculated by first taking the absolute value of the relative bias within each replication and then averaging these values across replications; (d) **root mean squared error (RMSE)**, calculated as the square root of the mean squared discrepancy between the human and synthetic estimates across replications and used to quantify overall recovery error while giving greater weight to larger discrepancies; and (e) **Pearson correlations** between the human and synthetic parameter estimates, which assess the extent to which the relative ordering of item parameters is preserved.

4 Results

4.1 Survey results: BESSI and LEADER

Response distributions. In both surveys, only 3 of 29 prompt conditions produced synthetic datasets using all five item response options in BESSI and all four response options in LEADER. The rest of the conditions resulted in synthetic data using only one or two of the item response options. We focused analyses on the datasets within those three conditions: C18

(providing the model with examples of five previous responses, represented by randomly generated numbers), C24 (asking the model to implicitly adopt a position on the latent trait being measured while also providing randomly generated numbers as the previous five responses), and C25 (asking the model to explicitly identify the latent trait underlying the items, specify the simulated person's standing on that trait, and providing random numbers as the previous five responses).

Items with low loadings. In the synthetic BESSI-13 data, one item consistently showed loading values < 0.20 across datasets (i.e., BESSI_136) (Table S1); this item corresponds to a cooperation scale and was one of the four items added to BESSI-9 to create a poor fitting model. In the synthetic LEADER data, by contrast, no single item was universally problematic (three items were dropped in $< 3\%$ of datasets).

Model fit. Overall, C24 and C25 resembled model fit patterns observed in human data, with poor model fit in BESSI-13 (Table S2) and LEADER-9 (Table S3), and better fit in BESSI-9 and LEADER-6.

Modification indices (MIs). The results (Table S4) indicate that for BESSI-13, the largest MIs were associated with residual covariances between items that were not part of the original BESSI self-management scale. For BESSI-9, the highest MI involved a pair of conceptually related items (“Think things through carefully” and “Double-check my work”). In the LEADER synthetic data, the largest MIs correspond to items taken from the same subscales.

Item parameter recovery. The results for BESSI (Table S5) indicate that while correlations with human item parameters were moderate to large, AD, RD and MARD had larger than acceptable values (i.e., $> .10$). Overall, Condition 24 was preferable for loading accuracy, while Condition 18 was preferable for intercept accuracy. Condition 24 produced the most accurate loading estimates for both BESSI forms, with the lowest AD, MARD, and RMSE. Loadings were generally underestimated, especially in Condition 25. Condition 18 produced the most accurate intercept estimates, with the lowest AD,

MARD and RMSE for both forms. Correlations showed a less consistent pattern: Condition 25 best recovered the ordering of BESSI-13 loadings, whereas Condition 18 performed best for BESSI-9.

The results for LEADER (Table S6) showed larger differences and lower (or even negative) correlations than BESSI. LEADER-9 generally showed better item parameter recovery than LEADER-6, particularly under Condition 24 and in terms of Pearson correlations. Intercepts had substantially stronger Pearson correlations than loadings. MARD varied across thresholds, with generally better recovery for t1 thresholds and considerably larger differences for t2 thresholds. Although loadings had lower RMSE, their weak Pearson correlations indicated that relatively low overall error did not correspond to recovery of the human loading pattern.

4.2 Assessment results: SCIENCE

Analysis with 20 replications. Human item percent-correct values ranged from 16.8% to 53.9%, indicating a relatively difficult assessment. In contrast, most prompt conditions produced LLM synthetic datasets with narrower score distributions and higher item percent-correct values, suggesting that the assessment was easier for the AI-generated respondents. Nineteen conditions that exhibited score ranges smaller than 20 percentage points and/or very high item percent-correct values in both forms were excluded from further consideration.

Because generating 500 dataset replications for every condition required substantial computational resources, we used a two-stage approach. First, we evaluated model fit and item parameter recovery using 20 replications for 14 candidate conditions. We then conducted the full 500 dataset generation and analyses for eight conditions (Table 1) that showed the most promising initial results (see the following sections). Results for the remaining conditions will be examined in future work.

Response distributions. Most prompt conditions produced substantially higher item percent-correct values and narrower score distributions relative to the human data (16.8%-53.9% correct), indicating higher proficiency levels

among synthetic respondents versus humans (Table S7). Among the conditions examined, Condition 22 most closely approximated the difficulty of the human assessment, whereas the remaining conditions generally yielded considerably higher percent-correct values.

Prompt	Description
C4	Includes the answer key, instructing the LLM to pick familiar-sounding or misleading answers
C5	Includes the answer key, instructing the LLM to pick the number of items to get correct before answering
C21	Includes an ability level for each individual (below average, average, above average)
C22	Includes the answer key, and an ability level for each individual (below average, average, above average)
C26	Same as 22, with values for the theta distribution (mean and standard deviation)
C28	Same as 26, encouraging chain-of-thought reasoning for the LLM to guess between the two most likely correct responses.
C30	Same as 26, with chain-of-thought reasoning around confidence in response based on item difficulty and ability level.
C32	Same as 26, with chain-of-thought reasoning around writing a paragraph explaining the LLM's answering approach.

Table 1. Prompt description for SCIENCE

Items with low loadings. The results (Table S8) show that item q27 (with a negative loading in the human sample) was dropped by over 80% of replications in every condition. Among the evaluated conditions, C5 most closely reproduced the pattern of low-loading items observed in the human data, with a high percentage of replications dropping items with loadings below .20 in the human data. However, C5 also produced loadings below .20 for several items that exceeded .20 in the human sample. In contrast, C4 was more conservative, consistently identifying two items with low loadings in the human data while rarely classifying higher-

loading items as problematic. However, C4 failed to identify several other items with low or negative loadings in the human data.

Model fit. Only C4 replicated the pattern observed in the human data, with SCIENCE_12 showing adequate fit and SCIENCE_16 failing to do so according to TLI with an average value across replications of .886 (Table S9). It should be noted that these fit measures correspond to the models after dropping the items with low loadings, explaining the adequate model fit for SCIENCE_16 in C5.

Modification indices (MIs). Overall, the average modification indices were modest, with only a few item pairs showing mean values above 10. The largest average MIs occurred primarily under Conditions 30 and 32 and involved q20, q21, and q24, where only q24 was an item added to create a poorly fitting model (Table S10). It should be noted that while q24 had a negative loading in the human data, it was rarely dropped for showing loadings below .20 in the LLM synthetic data. Other item pairs generally had lower average MIs and were infrequently flagged, suggesting that evidence of local dependence was limited and concentrated primarily among q20, q21, and q24.

Item parameter recovery. No condition exhibited AD, RD and MARD values under .10 (Table S11). Regarding loadings in SCIENCE_12, C26 produced the best overall parameter recovery, with the lowest AD (0.243), lowest MARD (0.540), and lowest RMSE (0.336), while maintaining a strong correlation with the human item loadings ($r = .630$). Although C30 achieved the highest correlation ($r = .736$), it exhibited substantially greater differences and RMSE. For SCIENCE_16, C26 and C32 showed the strongest performance. C32 minimized differences and error, whereas C26 provided a more favorable balance between recovery accuracy and association with the true parameters ($r = .461$ vs. $.359$). For loadings, C26 emerged as the most robust overall condition, offering consistently low differences, low RMSE, and comparatively strong parameter recovery.

For threshold recovery (Table S12), performance was generally weak relative to the loading estimates. For SCIENCE_12, C22 minimized AD (0.575), MARD (1.584), and RMSE (0.663). However, the average correlation with human loadings was negative ($r = -.350$). C28 provided the strongest correlation ($r = .311$) but had higher differences and RMSE. For SCIENCE_16, C22 again had the lowest AD (0.662) and RMSE (0.722), while C26 had a nearly identical MARD (1.988 versus 1.986 for C22). Nevertheless, both conditions produced negative correlations. C28 yielded the highest positive correlation ($r = .078$), although this association was very weak.

5 Conclusions

Overall, the findings indicate that developing prompts to yield usable synthetic data is not straightforward, and that AI-generated synthetic respondents are currently more effective for reproducing broad psychometric patterns than for recovering precise item parameters.

In our study, only 3 of the 29 survey-related prompts resulted in usable synthetic data. However, outcomes from these successful cases were encouraging. Under two specific prompting conditions, ChatGPT generated synthetic data useful for supporting early-stage scale development. Synthetic data were generally able to recover broad relationships among items and structural patterns, including distinctions between well-fitting and poorly fitting models. However, reproduction of specific item-level parameters was less favorable.

Performance was notably weaker for the science assessment. Although several conditions reproduced isolated features of the human data (e.g., item percent-correct distributions, identifying items with low loadings in the human data), none produced psychometrically convincing recovery across the full set of evaluation criteria. Across all conditions, threshold recovery was particularly poor, with weak and frequently negative correlations with the human threshold estimates. Thus, while some prompt conditions reproduced selected characteristics of the human data, none generated synthetic datasets that could be considered close

approximations of the human psychometric structure.

Several explanations may account for the relatively weak correspondence between the human and LLM synthetic datasets observed in the present study. First, the results suggest that generating realistic synthetic data may be more challenging for knowledge assessments than for survey scales. The strongest evidence of correspondence between human and synthetic data has been observed for BESSI, a behavioral inventory, whereas the SCIENCE assessment showed substantially poorer recovery of item characteristics, model fit, and response distributions. Similarly, although LEADER did not show the same level of correspondence observed for BESSI, its results were generally more favorable in terms of overall structural patterns than those obtained for SCIENCE. This suggests that AI-generated data may be better able to reproduce responses to attitudinal, behavioral, or self-report survey items than responses to items requiring specific knowledge or competencies. One possible explanation is that survey responses are more directly tied to the semantic and conceptual information available in the training data of large language models, whereas assessment responses require accurate modeling of variations in knowledge, misconceptions, and ability levels across individuals.

At the same time, differences in instrument maturity may also contribute to the observed findings. BESSI is the only instrument examined in this project that has extensive prior psychometric evidence supporting its structure and item functioning. As a result, the item parameters obtained from human respondents may be relatively stable and more likely to generalize across samples. In contrast, LEADER and SCIENCE are newer instruments for which item parameters have been studied less extensively. Consequently, some item characteristics observed in the human sample may be sample dependent rather than stable properties of the instrument. Under this interpretation, at least some of the discrepancies between the human and synthetic datasets may reflect instability in the human parameter estimates rather than failures of the LLM

synthetic data. It is possible that the synthetic data are capturing patterns that would emerge in other human samples, even if those patterns do not align perfectly with the specific sample used in the present analysis.

These considerations suggest an important direction for future research. Rather than using newer instruments as the primary benchmark for evaluating prompting strategies, future studies should first examine a broader set of instruments with well-established psychometric characteristics and evidence of parameter stability across multiple samples. Such instruments would provide a stronger basis for identifying the prompting conditions that most effectively reproduce human data. Once effective prompting strategies have been identified using mature instruments, they can then be applied and evaluated on newer instruments such as LEADER and SCIENCE.

Finally, the present study evaluated a particularly stringent use case in which all responses were generated exclusively by the language model. Future work may benefit from hybrid approaches in which AI-generated data augment, rather than replace, human data. For example, a language model could be informed by a smaller human sample and then generate additional respondents consistent with the observed response patterns. Such approaches may prove more effective for reducing data collection requirements while maintaining psychometric properties that more closely resemble those observed in human samples.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. <https://www.apa.org/science/programs/testing/standards>
- Benedetto, L., Aradelli, G., Donvito, A., Lucchetti, A., Cappelli, A., & Buttery, P. (2024, November). Using LLMs to simulate students' responses to exam questions. In A-O, Yaser, M. Bansal, & Y-N Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 11351–11368). Association for Computational Linguistics.

- Bisbee, J., Clinton, J. D., Dorff, C., Kenkel, B., & Larson, J. M. (2024). Synthetic replacements for human survey data? The perils of large language models. *Political Analysis*, 32(4), 401–416.
<https://doi.org/10.1017/pan.2024.5>
- Motoki, F. Y. S., Monteiro, J., Malagueño, R., & Rodrigues, V. (2023). *Large language models in survey research: Generating synthetic data and unlocking new possibilities*.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4595896
- Liu, Y., Bhandari, S., & Pardos, Z. (2025). Leveraging LLM respondents for item evaluation. A psychometric analysis. *British Journal of Educational Technology*, 56, 1028–1052 <https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13570>
- Säuberli, A., Frassinelli, D., & Plank, B. (2025). Do LLMs give psychometrically plausible responses in educational assessments? <https://aclanthology.org/2025.bea-1.21.pdf>
- Soto, C. J., Napolitano, C. M., Sewell, M. N., Yoon, H. J., & Roberts, B. W. (2022). An integrative framework for conceptualizing and assessing social, emotional, and behavioral skills: The BESSI. *Journal of Personality and Social Psychology*, 123(1), 192–222.
<https://doi.org/10.1037/pspp0000401>
- Rosseel, Y. (2012). *lavaan: An R package for structural equation modeling*. *Journal of Statistical Software*, 48(2), 1–36.
<https://doi.org/10.18637/jss.v048.i02>

A Supplementary Material

All tables are included in the Supplementary materials:

<https://docs.google.com/document/d/1YihzSTljcgox9XF0IUdjDI5236KEARth/edit?usp=sharing&oid=110493509676671251703&rtpof=true&sd=true>