

Do Synthetic Students Thrive? Validity and Fairness of Generated Assessment Data

Neba Nfonsang and Temple Lovelace

Assessment for Good - AERDF and Oluko Learning

{nnfonsang, tlovelace}@aerdf.org

Abstract

Synthetic assessment data can match real distributions while losing the within-group spread or correlation structure that asset-based instruments reveal in learners furthest from opportunity. Using a stratified, majority Black and Latinx sample of students, the authors evaluated whether synthetic assessment data generated through a retrieval-augmented generation (RAG) pipeline preserved the construct structure of ground-truth assessment data. Synthetic responses were evaluated using both distributional and psychometric measures, including response distributions, response agreement, confirmatory factor analysis, and model fit indices. The results of this study demonstrate that synthetic assessment data can closely reproduce the distributions of the original data. However, limitations were found. These data did not preserve underlying correlation structures and psychometric properties, highlighting the limitations of relying solely on distributional similarity when evaluating synthetic assessment data. These findings suggest that psychometric preservation should be considered alongside distributional similarity when evaluating synthetic assessment data.

without exposing real study participants. According to Kalkbrenner (2021), social scientists contribute to research by developing and validating scores on instruments as well as evaluating the psychometric properties of existing measures for use with research participants.

Synthetic data are defined as artificially generated data that are created to resemble the characteristics of real data without containing actual records from real individuals. Educational measurement research relies on access to high-quality assessment data to develop, validate, and refine instruments. While simulated datasets are not new in assessment research, particularly in educational measurement, recent advances in artificial intelligence and machine learning have renewed interest in the use of synthetic data. These approaches enable researchers to generate realistic datasets for instrument development, validation, methodological research, and privacy-preserving data sharing.

Researchers often face limitations related to sample size, privacy restrictions, incomplete responses, and the underrepresentation of certain populations, especially historically underrepresented populations. Inadequate sample sizes can hinder efforts to establish the psychometric quality and validity of assessment instruments (Anthoine et al., 2014).

Most recently, in the wake of frontier model use and generative artificial intelligence, synthetic data have continued to grow as a promising approach for addressing these challenges because synthetic data can augment limited datasets, facilitate data sharing while preserving privacy as well as support

1 Introduction

1.1 Background

Synthetic data is increasingly used to validate measures, enlarge small samples, and share data

experimentation with new assessment items, predictive modeling, methodological and machine learning research (İpek et al., 2026).

According to Jiang et al. (2024), synthetic data has potential to enable proprietary data access for statistical and research purposes and increase available data. The authors highlight the promise of synthetic data in enabling data sharing across institutions while preserving patient privacy and illustrate its utility for data augmentation and research.

Even more, Liu et al. (2025), stated that, educational researchers and practitioners can leverage synthetic data to release data without compromising student privacy, thereby promoting the development of open science and contributing to the advancement of education research. They also emphasized that high-volume, high-quality and diverse datasets are crucial for advancing research in the education field.

Finally, Vie, Rigaux & Minn (2022) argued that synthetic data generation opens new opportunities for research in education. In their paper, they present a generative model for educational data and an evaluation framework for comparing synthetic data generators. According to these authors, the comparison between real and synthetic datasets involved examining response distributions through histograms and evaluating measurement properties using Item Response Theory (IRT) estimates derived from both the real and generated data.

1.2 Problem Statement

Despite the advantages of synthetic data, it remains unclear whether synthetic data preserve the construct structure and learner variability captured by assessment instruments. Most evaluations of synthetic data focus on distributional fidelity, assessing the extent to which generated data reproduce the statistical characteristics of the original dataset. However, the psychometric properties of synthetic data may not match those of the original data for different reasons. For example, data generator functions are not neutral instruments.

Additionally, synthetic data generators, including retrieval-augmented generation (RAG) systems and large language models, are influenced by their underlying model architectures, training data, knowledge bases, and retrieval mechanisms. As a result, the synthetic data they produce partly reflect the assumptions embedded within the

generation process. This raises important concerns for asset-based assessment, particularly among learners from historically marginalized communities.

While asset-based instruments are designed to capture the diversity of learners' strengths and durable skills, synthetic generation processes may unintentionally reduce or homogenize variation within groups. For asset-based measurement of learners furthest from opportunity, the thing most at risk is within-group heterogeneity. Deficit-oriented systems compress these learners toward a single below-standard score, while the instrument and co-created construct ontology behind this work were built with learners to show the opposite, that they sit across the full range of durable-skill assets. We ask whether a synthetic generator keeps that range or flattens it, a question the field must contend with and one with direct consequences for equity.

Hence, an important question is whether synthetic data preserve the within-group heterogeneity observed among real learners. The answer has important implications for equity and fairness in assessment. Equally important is whether synthetic data preserve the construct structure reflected in learner response patterns.

1.3 Aims

This study makes three contributions. First, we present a retrieval-augmented generation (RAG) synthetic data pipeline developed for durable-skills assessment and report a finding about the role of its language-model component. Second, we propose diagnostics for evaluating whether a synthetic data generator preserves construct structure and within-group heterogeneity rather than pulling learners toward group-level averages. Third, we argue that standard algorithmic fairness approaches, which often seek to remove demographic information from models, may not be the appropriate framework for asset-based measurement.

2 Samples and Methods

2.1 Samples

The analyses used a durable-skills instrument that the authors created, administered to a sample that is majority Black and Latinx by deliberate design. The dataset consisted of 4,279 unique students who provided responses to a durable-skills assessment comprising 100 items administered to lower-grade

students and 130 items administered to higher-grade students. Across all students and items, the dataset contained approximately 339,000 item responses, yielding a large response matrix suitable for the RAG knowledge base. The ground-truth dataset used for the synthetic data evaluation consisted of 3,225 unique students who completed both the existing assessment items included in the RAG knowledge base and the newly developed items scored by the RAG system. Items were scored on 3-point, 4-point, or 5-point Likert-type scales. Assessment administration and validation activities were conducted under Institutional Review Board (IRB) approval.

2.2 Methods

Three generation approaches were used: multiclass classifiers (logistic regression, random forest, XGBoost) for generating scores of students on existing items, a retrieval or RAG system for new items, and a variational autoencoder for synthetic profiles. The retrieval finding is the methodological center of this paper. Figure 1 illustrates the architecture of the RAG pipeline. RAG is a technique used in modern AI systems to enrich prompts sent to large language models (LLMs). RAG enables LLMs to retrieve and incorporate information from specific internal knowledge bases, allowing LLMs to generate responses that are contextually relevant.

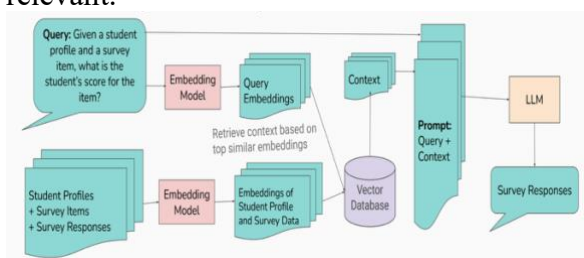


Figure 1: Retrieval-Augmented Generation (RAG) pipeline.

The RAG workflow was developed and executed within a secure, access-controlled private sandbox environment. Embedding generation, retrieval, and language model inference were performed entirely within the isolated environment, ensuring that student-level data, embeddings, retrieved context, and prompts

remained local and were not transmitted to any external model provider or third-party service.

The RAG workflow involves converting student profiles, items, and responses into documents, storing responses as metadata to avoid conditioning the context on response values, generating embeddings from documents, and storing them in a vector database. A retriever then identifies relevant context, which is combined with the query in a prompt and sent to the LLM to generate a response. We implemented hierarchical indexing to improve retrieval performance by clustering documents and retrieving context in multiple stages.

When we asked the language model to explain its response selections during retrieval, it was almost always choosing the most common response among the retrieved students, behavior consistent with a nearest-neighbor approach. We therefore replaced the LLM component of the RAG pipeline with an explicit K-nearest-neighbor (KNN) algorithm, implemented as both a classifier and regressor, which reproduced the LLM's outputs in 97.2% of cases. Replacing the LLM with a KNN classifier and regressor reduced generation time from approximately 1,000 responses per hour to 1,000 responses in under five minutes, while maintaining 97.2% agreement with the RAG-LLM outputs.

We assessed the quality of the synthetic data using distributional and psychometric metrics to determine whether the generated data preserved within-group variability and construct structure.

3 Results

Since the RAG pipeline was used to generate student responses, we evaluated the quality of the synthetic data generated by the RAG pipeline by comparing synthetic and ground-truth response distributions, assessing the agreement between generated and observed responses, and examining whether the synthetic data preserved the psychometric properties and measurement characteristics of the ground truth dataset.

3.1 Evaluation of Distributions

The RAG-generated responses exhibited a distribution that closely matched the distribution of the actual responses as shown in Figure 2 (Grades 3-5).

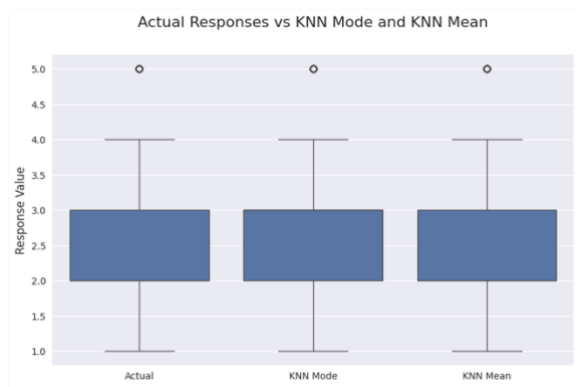


Figure 2: Distribution of actual and RAG-generated responses.

3.2 Evaluation of Accuracy

Agreement between generated and observed responses was assessed using exact match accuracy and accuracy with predictions falling within ± 1 response category as shown in Table 1.

Grade Band	Method	Exact Match	Accuracy Within ± 1
Lower Grades	KNN Mode	37.1%	83.3%
Lower Grades	KNN Mean	38.4%	90.6%
Higher Grades	KNN Mode	37.0%	82.6%
Higher Grades	KNN Mean	38.7%	90.5%

Table 1: Exact match accuracy and accuracy within ± 1 .

KNN Mean consistently outperformed KNN Mode across both grade bands, achieving approximately 38% exact-match accuracy and 90% accuracy within one response category.

3.3 Evaluation of Psychometric Properties

Although the RAG approach performed reasonably well based on accuracy within ± 1 , the generated responses often failed to preserve the correlation structure among items. The confirmatory factor analysis (CFA) revealed substantial differences between the RAG-generated and ground-truth data. As shown in Figure 3 (Grades 3-5), the RAG-generated constructs exhibited weak and, in some cases, negative factor loadings. Consistent with these results, RMSEA values indicated good fit for the ground-truth constructs (0.034-0.047) but poor

fit for the RAG-generated constructs (0.138-0.173), suggesting that the synthetic data failed to preserve the original construct structure.

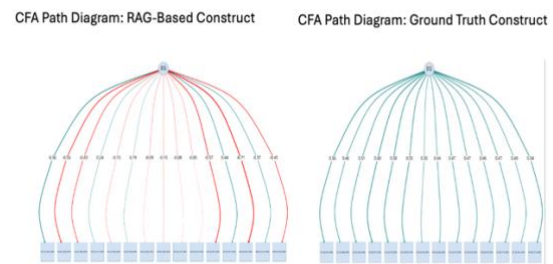


Figure 3: Standardized factor loading of RAG-based and ground truth constructs.

3.4 Variational Autoencoder Synthetic Data Evaluation

Variational Autoencoders (VAEs) generate synthetic data by reconstructing the underlying patterns in the original data. The psychometric properties of the VAE-generated assessment responses more closely resembled those of the ground-truth data than did the RAG-generated responses, as shown in Figure 4. Consistent with these results, RMSEA values indicated good fit for the ground-truth constructs (0.034-0.047), whereas the six VAE-generated constructs yielded RMSEA values ranging from 0.087 to 0.117. Although the VAE-generated data did not fully reproduce the fit observed in the ground-truth models, they demonstrated substantially better fit than the RAG-generated constructs (0.138-0.173), indicating stronger preservation of the original construct structure.

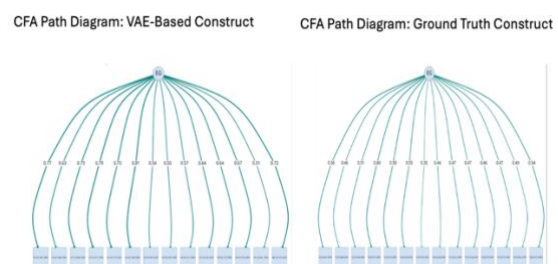


Figure 4: Standardized factor loading of VAE-based and ground truth constructs.

The distributions of the VAE-generated student profiles closely resembled those of the ground truth data. The VAE synthetic data demonstrated high similarity to the original dataset, achieving a 92.9% column shapes score, indicating strong preservation of individual variable distributions, and an 86.2% column pair trends score, indicating

that relationships between variables were largely maintained. These metrics yielded an overall quality score of 89.6%, calculated as the average of the column shapes and column pair trends scores, indicating strong preservation of both variable-level distributions and inter-variable relationships. Unlike the RAG-generated responses, the VAE-generated data preserved both the statistical distributions and psychometric structure of the original dataset, making them suitable for privacy-preserving data sharing and data augmentation applications.

4 Conclusions

We offer three contributions. First, we present a working synthetic-assessment pipeline and a finding about when a language model is doing no more than nearest-neighbor selection.

Second, we treat the generator as a context-laden tool rather than a neutral instrument, and we argue that the standard fairness move of removing demographic information may not be an appropriate frame for asset-based measurement, because it treats a demographic-neutral default as fair and erases the context the instrument exists to honor.

Third, we introduce a set of evaluation diagnostics that extend beyond distributional similarity. Our results suggest that preserving statistical distributions alone is insufficient. Although the RAG synthetic data closely resembled the ground-truth data in their overall distributions, they failed to preserve important correlation structures and psychometric properties.

These findings highlight the need to evaluate synthetic data not only on distributional similarity but also on their ability to preserve meaningful measurement characteristics, construct structure, and learner variability. Synthetic assessment data should therefore be evaluated not only on their ability to reproduce observed distributions, but also on their ability to preserve measurement structures that support valid inferences about learners.

This study is limited to six durable-skills instruments. Although the instruments span multiple durable-skills constructs, the RAG-related findings reported here may not generalize to other assessment domains, populations, foundation models, or synthetic data generation approaches. Moreover, the present results do not establish that the generated data are safe for public release,

unrestricted sharing, or downstream data augmentation.

Future research should examine whether the tendency of RAG-generated assessment data to preserve response distributions while failing to preserve construct structure is observed across other educational assessments and synthetic data generation approaches.

References

- Anthoine, E., Moret, L., Regnault, A., Sébille, V., & Hardouin, J.-B. (2014). Sample size used to validate a scale: A review of publications on newly-developed patient reported outcomes measures. *Health and Quality of Life Outcomes*, 12, 176. <https://doi.org/10.1186/s12955-014-0176-2>
- İpek, Z. H., Poyraz, E., & Yiğit, G. (2026). The role of synthetic data in educational research: A systematic review. *International Journal of Interactive Mobile Technologies*, 20(16). <https://doi.org/10.3991/ijim.v20i16.62224>
- Jiang, Y., García-Durán, A., Losada, I. B., Girard, P., & Terranova, N. (2024). Generative models for synthetic data generation: Application to pharmacokinetic/pharmacodynamic data. *Journal of Pharmacokinetics and Pharmacodynamics*, 51(6), 877–885. <https://doi.org/10.1007/s10928-024-09935-6>
- Kalkbrenner, M. T. (2021). A practical guide to instrument development and score validation in the social sciences: The MEASURE approach. *Practical Assessment, Research, and Evaluation*, 26(1), Article 1. <https://doi.org/10.7275/svg4-e671>
- Liu, Q., Shakya, R., Jovanovic, J., Khalil, M. I., & De La Hoz Ruiz, J. (2025). Ensuring privacy through synthetic data generation in education. *British Journal of Educational Technology*, 56(3), 1053–1073. <https://doi.org/10.1111/bjet.13576>
- Vie, J.-J., Rigaux, T., & Minn, S. (2022). Privacy-preserving synthetic educational data generation. *arXiv*. <https://arxiv.org/abs/2207.03202>