AlgebraNation Dataset: Demographics-Enriched Data to Support Fair Educational Machine Learning

Wanli Xing*
University of Florida
wanli.xing@coe.ufl.edu

Chenglu Li
University of Florida
li.chenglu@ufl.edu

Walter Leite
University of Florida
walter.leite@coe.ufl.edu

ABSTRACT

One of the major challenges for studies on promoting algorithmic fairness in educational machine learning can be the limited access to demographic information due to privacy and regulations. We presented a demographics-enriched dataset called AlgebraNation, with 893,190 assessment items and 20,297,075 log entries by 12,697 Algebra I learners between 2017 and 2019. We discussed the data context, collection, and attributes of AlgebraNation, providing researchers with opportunities to adopt the dataset to investigate and implement fair EML toward building trustworthy and sustainable AI in education.

Keywords

public dataset, math learning, virtual learning environment, demographic information

1. INTRODUCTION

To provide automatic, accurate, and scalable insights to support teaching and learning, researchers have widely examined educational machine learning (EML) in K-12 [3], higher education [11], and online learning contexts [5]. There are successful investigations and applications of EML in those settings, such as predictions for early warning [2], clustering for precision teaching [13], and text analysis for individualized scaffolding [6]. However, there is a gap that educational studies have paid limited attention to algorithmic fairness (AF) of EML [1]. Although there is yet to be a universal definition of AF, a common perception holds that it is the absence or minimization of systematic discrimination against individuals or groups of specific attributes (e.g., gender, race, learning profile) in EML [1, 9].

There have been conceptual and evaluation studies in education to discuss and identify the origin, harm, and measurement of algorithmic bias [1, 10, 9], an opposite end of AF. In contrast, relatively fewer efforts have been made to enhance AF in EML proactively. As suggested by Baker and Hawn [1], one of the major challenges for studies on promoting AF in EML can be limited access to demographic information due to privacy and regulations. To support the examination and development of fair EML, we aimed to present a dataset with enriched demographic information to be published by us. The dataset consisted of more than 12,000 students' behavioral (approx. 20 million entries) and within-platform assessment (approx. 890,000 items) data in an introductory algebra course from a virtual learning environment. We discussed more details of the dataset in the Data Description section. In the following sections, we referred to this dataset as AlgebraNation and its context as Algebra Nation.

*Corresponding Author

2. RELEVANT PUBLIC DATASETS

There are other public datasets for assessing students' learning outcomes that bear similar affordances to AlgebraNation. Three of them are ASSISTments (e.g., [15]), Eedi [17], and EdNet [4] datasets. These datasets have inspired numerous studies in adopting EML to construct intelligent tutoring systems (ITS) using techniques such as knowledge tracing [18], affect detection [14], correctness prediction [16], and question quality measurement [12]. However, a major distinction between AlgebraNation and the mentioned ones is that AlgebraNation provides rich demographic information of students, including gender, race, ethnicity, grade level, and reduced-cost meal benefits. While some of the existing datasets also include students' demographics, the attributes are usually limited (e.g., ASSISTments, Eedi), providing insufficient insights to consider elements such as cultural responsiveness in EML. The enriched demographic information, along with students' assessment and behavioral data, provide researchers with desirable testbeds to examine and develop fair EML. Researchers can utilize AlgebraNation to answer EML questions related to algorithmic fairness of regression (e.g., assessment score prediction), classification (e.g., pass/fail prediction), and clustering techniques (e.g., students' learning profiles identification). Researchers can also address questions regarding the influence of fairness-aware and fairness-unaware EML on understanding and interpreting students' learning results using AlgebraNation.

3. DATA DESCRIPTION

3.1 Context

The dataset was collected through Algebra Nation, a virtual learning environment developed by the University of Florida Lastinger Center for Learning in collaboration with Study Edge. There are approximately 1 million active users

on Algebra Nation every year. Algebra Nation has been renamed to Math Nation to reflect its broad coverage of K-12 math, including 6th-8th grade math, geometry, and SAT preparation. However, we kept the original name of Algebra Nation for the released dataset to distinguish it from future potential releases of its new version. The new Math Nation has been completely re-designed to align with the recent Florida Benchmarks for Excellent Student Thinking (BEST) Math Standards [8]. In contrast, the current version–Algebra Nation–was designed to align with the Mathematics Florida Standards (MAFS) [7].

Algebra Nation was designed to be a supplement for teaching and learning; there is no expectation for students to complete an entire course or section. A section in Algebra Nation can be treated as a module, where a general topic (e.g., expressions, inequalities) with multiple sub-topics will be covered. Algebra Nation was designed to support teachers in the ways they teach best and to support school districts that choose different combinations of curriculum materials and where students' access to technology varies. Algebra Nation provides students with PDF workbooks, instructional videos with diverse instructors (e.g., Spanish, diverse gender and racial identities), and Q&A forums supported by paid study experts. Figure 1 illustrates the main interfaces on Algebra Nation for students' learning.

3.2 Data Collection & Pre-Processing

AlgebraNation consists of 89,982 within-course assessments ($n_{item} = 893,190$) by 12,136 unique Algebra I learners, with 20,297,075 logs recorded. The platform behavioral and assessment data were collected from 05/01/2017 to 12/31/2019 to allow the investigation of Florida end-of-course (EoC) Algebra I exams preparation and follow-up behaviors in the academic year from 2018 to 2019. All Algebra Nation users' data are stored in a MySQL database. Students are de-identified in the database and are assigned unique user IDs. Assessment and behavioral data are automatically collected when students interact with the Algebra Nation platform. The demographic information comes from the district student information system (SIS) with parents' consent, where districts integrate with Clever or Classlink. This information is not displayed in Algebra Nation accounts and would only change if the change were made in the district SIS. Demographic information is only visible in the backend database. The data collection has been reviewed and approved by IRB from the authors' institute. To prepare the dataset:

1. We identified students who took EoC in 2018-2019 from the same school district to minimize the influences on instructions caused by the pandemic in 2020 ($n_{students} = 14,251$). Since the ways and goals to use Algebra Nation vary, we selected students from the same school district, aiming to keep similar usage patterns and engagement of Algebra Nation among students.

2. We fetched from the MySQL database to retrieve these students' behavioral and assessment data. We removed students without platform interactions. There were 12,697 students who had engaged with Algebra Nation.

3. We recalculated whether students had answered an assessment item correctly. The values stored in the database can only accurately reflect the correctness of questions that do not require student inputs (e.g., multiple-choice). For example, the store value might be stored as Null or False when a student inputs "2/4" for an answer of "1/2".

Figure 2 shows students' demographic distributions on race, gender, and hispanic-ethnicity. Figure 3 shows the number of assessment items taken by students grouped by demographics.

3.3 Data Tables

There are three tables in the dataset: students, logs, and assessments. These tables can be connected using the shared "useraccount_id" column, which is the unique identifier of each student. All three tables have a "ts_created" column to suggest the creation timestamp, allowing temporal analysis of the data. The following sections provided detailed information on the attributes of each table.

3.3.1 Students

Students' metadata is stored in a separate table (see Table 1). Each row in the table represents a unique student. Students' demographics (race, gender, hispanic ethnicity, and reduced-cost lunch benefits) were coded in binary values (0 and 1).

Table 1: Descriptions of columns in the Students table

Column	Definition
id	id in the student table, not useful. Use useraccount_id instead
useraccount_id	User account id, which can be treated as users' unique identifiers
race	Whether a student is in the minority group (0: no, 1: yes)
hispanic_ethnicity	Whether a student is hispanic (0: no, 1: yes)
gender	Gender of a student (0: female, 1: male)
free_lunch	Whether a student is receiving free/reduced-cost lunch (0: no, 1: yes)
grade_flag	Categorical value of grade level: middle or high
grade	Current grade level of a student (e.g., 7th grade)
is_active	Whether a student's account is still active
karma_point	Rewarding points by interacting in the Q&A forum, which can be used to receive rewards
ts_created	When a student account was created
ts_modified	When a student account was modified

3.3.2 Logs

Students' behavioral logs are explained in Table 2. Students' every interaction with Algebra Nation is captured automatically in logs. Each row in the table is an action of a student. Actions can be separated into four categories: navigation

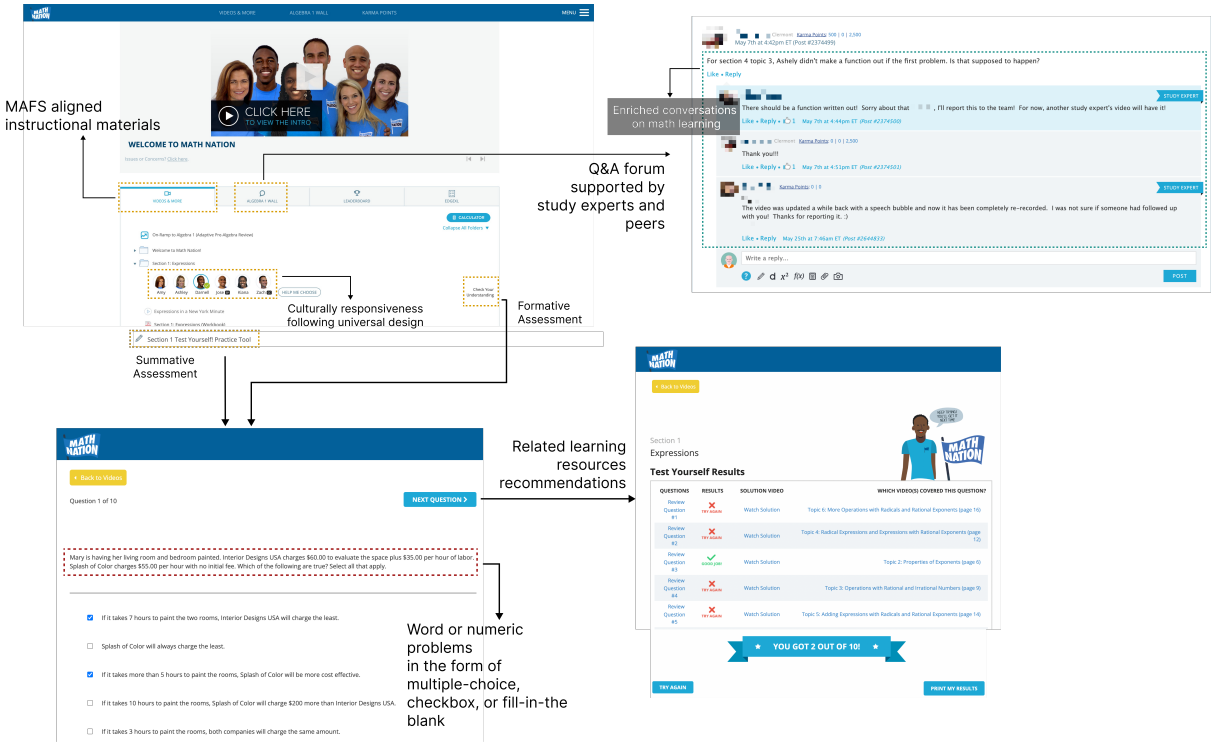

Figure 1: Students' learning interface on Algebra Nation

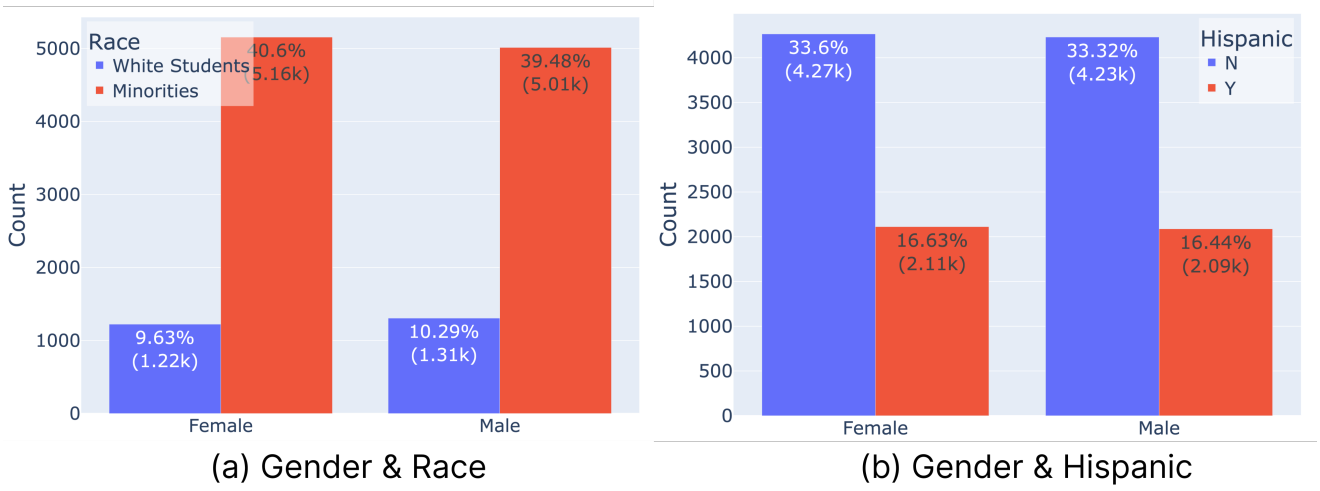

Figure 2: Students' demographic distributions

(e.g., page-loading), assessment (e.g., start, answer, or finish an assessment), video (e.g., play, pause, seek, progress), and discussion (e.g., create posts).

3.3.3 Assessments

Table 3 describes columns in the Assessments table. Each table row represents an item of an assessment. If an assessment has ten items, there will be ten rows to make up the complete assessment. Questions of assessments were generated from a bank of questions and randomly assigned, but not necessarily in increasing order of difficulty. The questions were random for the first time. When students retook an assessment, they saw the same questions drawn from the question bank. Students might not be required to take assessments as some teachers used Algebra Nation as materials to supplement their curriculum. Therefore, students may have logs without completing assessments.

3.4 Availability

The dataset will be released through a Kaggle competition on constructing fair EML with AlgebraNation. We are currently in the final stage of publishing the data with Kaggle.

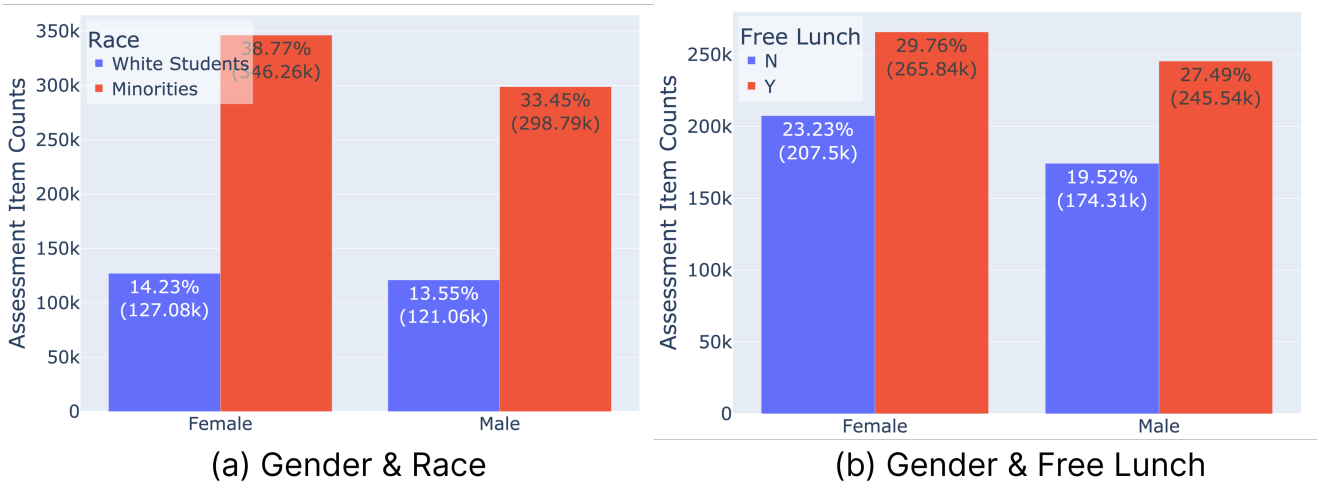

Figure 3: Assessment item distributions by demographics

Table 2: Descriptions of columns in the Logs table

Column	Definition
id	Unique identifier of logs.
useraccount_id	User account id, which can be treated as users' unique identifiers
subject_id	What context of learning subject where a log was recorded
session_log_id	User's login session id
action_name	What a log is about in terms of specific actions.
field_val	Extra info value recorded in the log
field_name	Extra info recorded in the log. Sometimes this causes duplicate log ids. Because there might be several pieces of extra info recorded in one log as primitive types (e.g., string, number), and they will be recorded in separate rows with the same log id.
video_id	Video id of a log. This can be none depending on the action type
time_position	Video's time position in milliseconds. This is only available in video-related actions
video_action	Video-specific action
ts_created	When a log was created

The Kaggle competition will link to this paper to inform participants of essential data information. Readers can check this URL for data availability updates, where the dataset will be available to download once released.

4. CONTRIBUTIONS & CONCLUSIONS

EML has attracted broad attention in the educational community to automate teaching and learning support across settings. Recently, the issue of FATE in EML has received increasing investigations toward building trustworthy and sustainable AI applications in education. However, a major challenge in addressing FATE in EML is the lack of demographic information in large-scale datasets for STEM education. This paper contributes by presenting a large dataset of over 12,000 Algebra I learners' behaviors and learning outcomes with rich demographic information. To the best of our knowledge, AlgebraNation is the first of its kind to provide various demographic information on a large scale. We expect to see more researchers conduct various tasks utilizing AlgebraNation to examine FATE issues and potentially develop solutions to address or mitigate such issues.

5. ACKNOWLEDGMENTS

The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant R305C160004 to the University of Florida, the University of Florida AI Catalyst Grant -P0195022, and the University of Florida Informatics Institute Seed Funding. The opinions expressed are those of the authors and do not represent the views of the University of Florida, Institute of Education Sciences, or those of the US Department of Education.

6. REFERENCES

[1] R. S. Baker and A. Hawn. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, pages 1–41, 2021.

[2] D. Bañeres, M. E. Rodríguez, A. E. Guerrero-Roldán, and A. Karadeniz. An early warning system to detect at-risk students in online higher education. *Applied Sciences*, 10(13):4427, 2020.

[3] A. L. Beerwinkle. The use of learning analytics and the potential risk of harm for k-12 students

Table 3: Descriptions of columns in the Assessments table

Columns	Definition
assessment_id	The whole assessment's id. This can be duplicate depending on the number of items in one assessment.
useraccount_id	User account id, which can be treated as users' unique identifiers
session_log_id	User's login session id
number_of_questions	How many questions/items are in an assessment
is_finished	Whether an assessment is finished and submitted
type	Type of assessment, mainly affecting the number of items in an assessment (e.g., 3 or 10)
item_id	A question instance id, this should be unique
question_id	A question blueprint id. This can be duplicate. For example, students might take the same assessment, which contains the same question. The question_id (namely, content) will be the same, however, item_id (instance) will be different.
user_given_answer	What id/text has been given to answer a question.
user_gave_correct_answer	Whether an item/question has been answered correctly.
question_time_taken	How much time a student used to finish a question in seconds
question_text	The question text
question_choices	What are the available choices in this question
correct_answer_json_array	For question type other than multiple-choice only. Correct answers in json format
solution_explanation_video_url	Where is the solution video.
subject_matter_video_name	What video a question is covered in
ts_created	When an assessment was created
ts_modified	When an assessment was modified (e.g., opened -> closed -> reopened and answered questions)
question_createdAt	When a user answered a question
question_modifiedAt	When a user changed answers of a question

participating in digital learning environments. *Educational Technology Research and Development*, 69(1):327–330, 2021.

[4] Y. Choi, Y. Lee, D. Shin, J. Cho, S. Park, S. Lee, J. Baek, C. Bae, B. Kim, and J. Heo. Ednet: A large-scale hierarchical dataset in education. In *International Conference on Artificial Intelligence in Education*, pages 69–73. Springer, 2020.

[5] S. Crossley, D. S. McNamara, R. Baker, Y. Wang, L. Paquette, T. Barnes, and Y. Bergner. Language to completion: Success in an educational data mining massive open online class. *International Educational Data Mining Society*, 2015.

[6] D. Dessì, G. Fenu, M. Marras, and D. R. Recupero. Bridging learning analytics and cognitive computing for big data classification in micro-learning video collections. *Computers in Human Behavior*, 92:468–477, 2019.

[7] FL-DoE. Mafs: Mathematics standards.

[8] FL-DoE. Standards review.

[9] J. Gardner, C. Brooks, and R. Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics & knowledge*, pages 225–234, 2019.

[10] R. F. Kizilcec and H. Lee. Algorithmic fairness in education. *arXiv preprint arXiv:2007.05443*, 2020.

[11] O. Moscoso-Zea, P. Saa, and S. Luján-Mora. Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining. *Australasian Journal of Engineering Education*, 24(1):4–13, 2019.

[12] L. Ni, Q. Bao, X. Li, Q. Qi, P. Denny, J. Warren, M. Witbrock, and J. Liu. Deepqr: Neural-based quality ratings for learnersourced multiple-choice questions. *arXiv preprint arXiv:2111.10058*, 2021.

[13] Y. Park, J. H. Yu, and I.-H. Jo. Clustering blended learning courses by online behavior data: A case study in a korean higher education institute. *The Internet and Higher Education*, 29:1–11, 2016.

[14] M. O. Z. San Pedro, R. S. d Baker, S. M. Gowda, and N. T. Heffernan. Towards an understanding of affect and knowledge from student interaction with an intelligent tutoring system. In *International Conference on Artificial Intelligence in Education*, pages 41–50. Springer, 2013.

[15] D. Selent, T. Patikorn, and N. Heffernan. Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 181–184, 2016.

[16] D. Shin, Y. Shim, H. Yu, S. Lee, B. Kim, and Y. Choi. Saint+: Integrating temporal features for ednet correctness prediction. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, pages 490–496, 2021.

[17] Z. Wang, A. Lamb, E. Saveliev, P. Cameron,

J. Zaykov, J. M. Hernandez-Lobato, R. E. Turner, R. G. Baraniuk, C. Barton, S. P. Jones, et al. Results and insights from diagnostic questions: The neurips 2020 education challenge. In *NeurIPS 2020 Competition and Demonstration Track*, pages 191–205. PMLR, 2021.

[18] Y. Yang, J. Shen, Y. Qu, Y. Liu, K. Wang, Y. Zhu, W. Zhang, and Y. Yu. Gikt: a graph-based interaction model for knowledge tracing. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 299–315. Springer, 2020.