Datasheet for 'Pandemic Schooling Models in Student Learning'*

Denise Chang

April 17, 2024

The COVID-19 pandemic transformed the educational landscape as educators move away from traditional in-person learning models towards virtual and hybrid learning models. Using data from the National Center for Education Statistics and from various district-level assessements, this paper investigates the impact of schooling modes on students' pass rate in state standardized assessments in grades 3-8 during the 2020-2021 school year. The exploration of the data across 11 states suggests that the overall student pass rates declined during the pandemic school year. The pass rates have also seen more drastic changes in schools who had a higher share of virtual. The results of this study are significant as they can be used by educational authorities and policymakers to support student-centerd models.

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The dataset was created to enable analysis of learning models. We needed a dataset that captures both learning model shares and pass rates in percentage points.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The dataset was created by the team of researchers of the original paper. The main researchers are Rebecca Jack, Clare Halloran, James Okun and Emily Oster. (Jack et al. 2023)

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - Unknown.

*Code and data are available at: https://github.com/DeniseChang9/Learning_Models.git

4. *Any other comments?*

    - None.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - Each instance represents one district in the given year from one of the 11 investigated states. As there were many years studied, a same district could appear more than once.

2. *How many instances are there in total (of each type, if appropriate)?*

    - There are 22107 instances in total, and 2328 unique districts.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - This dataset is a sample from a larger set. The larger set is the set of all district of all 51 states in the United States. This sample is representative as the 11 state in this sample are spread out evenly across the country. This dataset is also a sample of the larger set of all districts in the 11 states.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Each instance consists of a response of a survey. If a district answered the survey twice, they would have two instances in this dataset.

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - No.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Yes, there are missing information from individual instances. Certain states did not require state assessments in Spring 2021, such that the districts who were part of these states do not have data on these specific test scores.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
    - Individual instances can be related to each other by their state. Through their district ID, each instance is related to one of the 11 states. Districts from the same state operate similarly and respond to the same policies.
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
    - Unknown.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
    - Yes. Certain instances have inconsistent learning model shares. In extreme cases, the shares add up to 80%, which leaves 20% of missing data. Given that there are proportions, a full dataset would have their shares add up to wholes.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - Yes. This dataset identify sub-populations by grade. A student's grade is identified by the grade they are registered during the school year. The sub-populations are: grade 3, grade 4, grade 5, grade 6, grade 7 and grade 8

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The original raw dataset from the replication package has data on ethnic origins share, but the clean dataset for my study does not. The original dataset has shares on black students, hispanic students and students of other origins for each instance.

16. *Any other comments?*

    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data was downloaded from individual state websites. These websites collected data from schools through survey response. Whether the individual states verified the data is unknown.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Unkown.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - Sampling was based on participation rates and by available data. Only states who had information for at least 3 years in the studied timeframe (2017 to 2021) were considered. States must also have a high enough participation rate. The participation cut off is unknown.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - School authorities were involved in the collection process on a voluntary basis. No financial compensation was provided for answering the survey.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - Each state continuously collect data. The timeframe of the data associated with the instances in this paper is from 2017 to 2021.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - The data was collected via third party source.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - Unknown.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Unknown. Considering is was a voluntary survey, I assume the individuals consented.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Unknown.

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a*

description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

- Unknown.

12. *Any other comments?*

    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - For this paper, cleaning of the data was. I removed the variables I was not interested in for the scope of my analysis such as demographic factors and commuting information.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

    - Yes. The raw data can be found at https://github.com/DeniseChang9/Learning_Models/tree/d883ed

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

    - Yes. The data was cleaned with Rstudio in the statistical programming language R (R Core Team 2022).

4. *Any other comments?*

    - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

    - The dataset has been used for a previous analysis (Jack et al. 2023). This prevous analysis is the original paper of my replication.

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

    - Link to my replication: https://github.com/DeniseChang9/Learning_Models.git
    - Link to the replication package: http://doi.org/10.3886/E168843V1

3. *What (other) tasks could the dataset be used for?*

- Unknown

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

   - No/Unknown

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

   - No/Unknown

6. *Any other comments?*

   - None.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*

   - The dataset from the original paper is distributed via a replication portal for potential replicators.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

   - The initial dataset distribute on an online portal. The DOI is http://doi.org/10.3886/E168843V1

3. *When will the dataset be distributed?*

   - It has already been distributed and is available to download

4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

   - Unknown.

5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- Unknown.

6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - Unknwon.
7. *Any other comments?*
   - None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Unknown.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - The owner can be contacted by email: emily_oster@brown.edu
3. *Is there an erratum? If so, please provide a link or other access point.*
   - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
   - No/Unknown.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
   - No/Unknown.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
   - No/Unknown.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

- No/Unknown.

8. *Any other comments?*
    - None.

References

Jack, Rebecca, Clare Halloran, James Okun, and Emily Oster. 2023. *Pandemic Schooling Mode and Student Test Scores: Evidence from US School Districts*. American Economic Review: Insights. http://doi.org/10.3886/E168843V1.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.