Data sheet:

1. Motivation:

- For what purpose was the dataset created:

The dataset was created for a randomized controlled trial implemented in 22 elementary schools across 3 states (Texas, Illinois and Arizona) in U.S. in 2016-2017 school year. This experiment was designed to test the effect of Descubriendo la Lectura (DLL), which is an one-to-one literacy training program taught in Spanish language for first-grade Spanish-speaking students whose Instrumento de Observacion (IdO, a type of standard Spanish literacy assessment used to test the literacy skills of students who participated in DLL program) test scores were within the lowest 25% of their own school, on the literacy skills of these students. Students involved in this experiment were randomly assigned to a control group that received regular education and a treatment group that received DLL training. Therefore, the dataset was created to record the pre-test and post-test scores of Logramos (another type of Spanish literacy assessment) for students participated during the first semester of 2016-2017 school year and also compare the differences in post-test scores of Logramos between these 2 groups to explore whether DLL program is effective.

- Who created the dataset and on behalf of which entity:

The authors of the original paper designed this experiment and created this dataset. They are Trisha H. Borman, Geoffrey D. Borman, Scott Houghton, So Jung Park, Bo Zhu, Alejandra Martin, and Sidney Wilkinson-Flicker.
Trisha H. Borman, Scott Houghton, So Jung Park, Bo Zhu, Alejandra Martin, and Sidney Wilkinson-Flicker on behalf of American Institutes for Research, Geoffrey D. Borman on behalf of University of Wisconsin–Madison.

- Who funded the creation of the dataset:

The Institute of Education Sciences, U.S. Department of Education funded this research and creation of the dataset through Grant #R305A160060 to American Institutes for Research.

2. Composition:

- What do the instances that comprise the dataset represent? Are there multiple types of instances?

The instances that comprise the dataset represent the first-grade Spanish-speaking students that participated in this randomized controlled trial in the first semester (ie. Fall semester) of 2016-2017 school year. The only type of instances is people.

- How many instances are there in total:

There are 152 instances in total.

- 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?

The dataset is a sample of instances from a larger set and it is not randomly sampled. The sample is 152 first-grade Spanish-speaking students whose IdO test scores are within the lowest 25% of their schools among 22 qualified schools in Texas, Illinois and Arizona in U.S. during Fall semester of 2016-2017 school year. The population (ie. larger set) is all entry-level elementary school students (ie. young children) whose native language is Spanish and have weak literacy skills in U.S.. The sample is not a representative of the larger set, because this experiment does not randomly sample from the qualified schools and students using some random sampling methods like Simple Random Sampling or Stratified Random Sampling, and this may lead to the selection bias. But meanwhile, this sample includes students across 3 states, which reduces the bias due to geographical difference.

- What data does each instance consist of ("Raw" data or features):

Each instance consists of the pre-test and post-test scores of Logramos (which has 3 subtests: Vocabulary, Reading and Language, and a combined overall total score), as well as the assignment of each student (control or treatment). All the data is in raw format and some missing values exist.

- Is there a label or target associated with each instance:

There is no label associated with each instance because this is not a machine learning task, but we aim to build a statistical model (ie. linear regression) to explore the effect of DLL on students' literacy skills and predict their post-test scores based on the assignment and pre-test scores.

- Is any information missing from individual instances:

The values for pre-test and post-test scores of some students are missing, probably because of the measurement errors, or some students did not finish the pre-test, or some exited the experiment early and they did not finish the post-test. Also, the demographic information for schools and students is missing in the dataset because it is unavailable.

- Are relationships between individual instances made explicit:

Each individual instance has a random assignment (control/treatment group), and the students' post-test scores are compared between these 2 groups.

- Are there recommended data splits (e.g., training, development/validation, testing):

There is no recommended data split since this task is to do statistical analysis on an experiment, rather than a machine learning task that needs to split the data into training and test data.

- Are there any errors, sources of noise, or redundancies in the dataset:

There are few missing values in the dataset and there are some extreme values for the response variable (post-test scores) and explanatory variables (pre-test scores). It is appropriate to remove the null values since only few of them exist, but this will still lower the accuracy of results. We need to explore the extreme values further to decide whether they exist due to errors (remove it or replace with mean values) or they are outliers (that will affect the analysis results). Also, the column 'Group' and 'T_assignment' are replicated columns that both show the assignment result (0 means control/delayed group, 1 means treatment/immediate group), so the 'Group' column can be removed.

- Is the dataset self-contained, or does it link to or otherwise rely on external resources:

The dataset can be downloaded from the OPENICPSR website (a repository where people can share the research data) freely following this link: https://www.openicpsr.org/openicpsr/project/118041/version/V1/view, as long as the users have registered a free account.

- Does the dataset contain data that might be considered confidential:

The dataset does not contain the confidential information, because each instance is represented using a unique id number and there is no demographic information included.

- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

The dataset contains the test scores of each student, which might be offensive for some people, but each instance is represented using an id number instead of the student's name directly.

- Does the dataset relate to people:

The dataset relates to people and each instance represents a student participated in this experiment.

- Does the dataset identify any subpopulations? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

The dataset identifies 2 subpopulations (control group and treatment group) based on the random assignment. The summary statistics and scatter plots of test scores are created to see the distributions, the pre-test scores are distributed similarly for these 2 groups regarding all of the 3 subtests and the overall total score, but the post-test scores for the treatment group tend to be higher than the control group regarding all of the 3 subtests and the overall total score.

- Is it possible to identify individuals, either directly or indirectly (i.e., in combination with other data) from the dataset:

It is possible to identify individuals directly by the unique id number.

- Does the dataset contain data that might be considered sensitive in any way:

The dataset contains the test scores of each student, which may be considered as sensitive data because it is related to the academic performance of students. But these students were told that they were participating an experiment and they agreed to participate in it.

3. **Collection process**:

- How was the data associated with each instance acquired? Was the data directly observable, reported by subjects, or indirectly inferred/derived from other data?

The data associated with each instance was acquired using a randomized controlled trial. The experimenters recorded each student's school id, group, pre-test and post-test scores directly during the experiment, so the data was directly observable.

- What mechanisms or procedures were used to collect the data? How were these mechanisms or procedures validated?

Since the dataset contains the basic information (like school id and group) and test scores of each student, and the sample size is relatively small (152 students), the data was recorded by experimenters manually. There may exist measurement error, but how this data collection procedure was validated was unknown.

- If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The sample was formed by recruiting the qualified schools and students in the 3 U.S. states, then 22 qualified schools and 152 students enrolled in this experiment, so the sampling strategy was deterministic and there was no random sampling strategy like Simple Random Sampling or Stratified Random Sampling to ensure the representativeness of sample and better generalization of analysis results.

- Who was involved in the data collection process and how were they compensated?

The qualified students (ie. first-grade Spanish-speaking students whose IdO scores were within the lowest 25% in their schools) in 22 qualified schools in 3 U.S. states (Texas, Illinois and Arizona) were involved in the data collection process. Whether they were compensated or not was unknown.

- Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances?

The data was collected during the Fall semester in 2016-2017 school year. This timeframe matches the creation timeframe of the data associated with the instances, because this data was collected during the implementation of the experiment to record its results.

- Were any ethical review processes conducted:

Whether there were ethical review processes conducting was unknown.

- Does the dataset relate to people:

The dataset relates to people and each instance represents a student participated in the experiment.

- Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources:

I obtained the dataset by downloading it from the website OPENICPSR where people can share the research data. The download link is: https://www.openicpsr.org/openicpsr/project/118041/version/V1/view.

- Were the individuals in question notified about the data collection:

Unknown.

- Did the individuals in question consent to the collection and use of their data:

Unknown.

- If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses:

Unknown.

- Has an analysis of the potential impact of the dataset and its use on data subjects been conducted:

Unknown.

4. **Preprocessing/cleaning/labeling:**

- Was any preprocessing/cleaning/labeling of the data done:

Since most data in the dataset is categorical/discrete/continuous data and a few missing values exist in the dataset, there was no preprocessing/cleaning/labeling of the data done. But I cleaned the data by removing the missing values and replacing the outliers with the average.

5. **Uses:**

- Has the dataset been used for any tasks already:

The dataset was created for a randomized controlled trial, so it has been used by the original authors who performed this experiment and wrote this paper to analyze the potential benefits of DLL program.

- 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.

The paper that uses the dataset can be accessed using the open access page on SAGE: https://us.sagepub.com/en-us/nam/open-access-at-sage, and the paper is available here: https://journals.sagepub.com/doi/pdf/10.1177/2332858419870488.

- What (other) tasks could the dataset be used for?

The dataset and the experiment were designed to explore the effect of DLL program, and I cannot come up with other tasks that could use this dataset so far.

- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? Is there anything a future user could do to mitigate these undesirable harms?

The assignment of students to the control or treatment group was randomly decided using a random number generator to avoid the self-selection bias. Also the pre-test and post-test for 2 groups were performed at the same time to control some other factors same except the intervention, because the students who took the test later may learn more at school and had the potential to have stronger literacy skills. Besides, even though the dataset only records the Fall semester data, in the experiment, the control group students in the dataset would receive DLL training in the Spring semester, so this avoids some ethical harms to the control group.

• Are there tasks for which the dataset should not be used?
Unknown.

6. **Distribution**:

- Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?

The dataset has been shared on the OPENCISPR website that is a public access research data sharing repository.

- How will the dataset will be distributed? Does the dataset have a digital object identifier (DOI)?

The dataset has been uploaded in the csv file directly on the website. The dataset has a DOI: 118041.

- When will the dataset be distributed?

The dataset was distributed on 2020-03-04.

- 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.

The dataset is distributed under an Other License, the link is available here: https://www.openicpsr.org/openicpsr/project/118041/version/V1/download/terms?path=/openicpsr/118041/fcr:versions/V1/DLL_lgrm.csv&type=file.

- 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.

The OPENICPSR website has imposed restrictions on the use of the dataset about the privacy of research subjects: 1. Do not use these datasets for investigation of specific research subject, except when identification is authorized in writing by ICPSR. 2. Do not use the identity of any research subject discovered inadvertently, and advise ICPSR of any such discovery. The link of these restrictions is: https://www.openicpsr.org/openicpsr/project/118041/version/V1/download/terms?path=/openicpsr/118041/fcr:versions/V1/DLL_lgrm.csv&type=file.

- 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.

There are regulatory restrictions applying to the dataset: 1. Users must agree to reference the recommended bibliographic citation in any publication that employs resources provided by ICPSR. 2. Users must acknowledge that the original collector of the data, ICPSR, and the relevant funding agency bear no responsibility for use of the data or for interpretations or inferences based upon such uses. 3. If ICPSR determines that you have violated the terms of this agreement, ICPSR will act according to our policy on terms of use violations and impose sanctions. The link of these restrictions is: https://www.openicpsr.org/openicpsr/project/118041/version/V1/download/terms?path=/openicpsr/118041/fcr:versions/V1/DLL_lgrm.csv&type=file.

7. **Maintenance**:

- Who is supporting/hosting/maintaining the dataset?

The original authors who performed the experiment and collected the data are supporting/hosting/maintaining the dataset.

- How can the owner/curator/manager of the dataset be contacted:

One of the authors Bo Zhu can be contacted by email: bzhu@air.org.

- Is there an erratum? If so, please provide a link or other access point.

Unknown.

- Will the dataset be updated? If so, please describe how often, by whom, and how updates will be communicated to users?

Unknown. The dataset was downloaded from the website.

- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances? If so, please describe these limits and explain how they will be enforced.

Unknown.

- 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 users.

Unknown.

- 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 other users? If so, please provide a description.

Unknown.