Datasheet for 'Nudge Data Repository dataset'*

A Randomized Study to Evaluate the Effect of a Nudge via Weekly Emails on Students' Attitudes Toward Statistics

Shuangyuan Yang

December 3, 2024

This datasheet provides a detailed description of a dataset compiled to evaluate the effect of weekly email nudges on attitudes towards statistics. The dataset includes over 1,000 records refined from a larger pool of student responses, especially focus on participants who experienced email nudges or different instructional formats. It includes variables such as demographic information, academic performance metrics, changes in attitudes such as affect, cognitive competence, interest, and difficulty, and behavioral engagement, each selected for their relevance to understanding predictors of educational outcomes.

Overview

Extracted from the questions outlined in Gebru et al. (2021), this datasheet was analyzed and processed using `R` (R Core Team 2023). The dataset, sourced from the Nudge Data Repository (Taback and Gibbs 2022a), originates from a randomized experiment conducted on students enrolled in the "Practice of Statistics I" (STA220H1F) course at the University of Toronto (Taback and Gibbs 2023). The dataset is accessible via its digital object identifier: (Taback and Gibbs 2022b).

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 investigate the effects of personalized email nudges on students' attitudes toward statistics. It focuses on whether tailored, engaging emails highlighting real-world applications of statistics could improve students' confidence, interest, and motivation compared to generic course emails. The dataset fills a gap in understanding how non-mandatory, scalable digital interventions influence educational engagement and behavioral outcomes.

*Code and data are available at: https://github.com/Sophiaaa-Y/student-effort-analysis.

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 Nathan Taback and Alison L. Gibbs from the Department of Statistical Sciences at the University of Toronto.

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.*

    - The funding source is not explicitly mentioned in the study, but the project was approved by the University of Toronto research ethics board.

4. *Any other comments?*

    - This dataset provides valuable insights for educators exploring the role of digital nudges in shaping student engagement and attitudes.

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 student who participated in the randomized experiment. Key features for each instance include:

    1. Pre- and post-survey data: Measuring attitudes toward statistics, such as interest, confidence, and perceived difficulty.
    2. Email intervention data: Details of the email received (plain or interesting) and whether the email was opened.
    3. Behavioral metrics: Engagement levels, such as participation in assignments, attendance, and study habits.
    4. Academic performance: Metrics like quiz scores, final exam grades, and cumulative GPA (CGPA).
    5. Demographics: Variables such as gender, academic year, and program type.

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

    - The final dataset includes a subset of the original 1,430 observations, refined through pre-processing and cleaning. Instances with incomplete responses or missing data were excluded, with each remaining record corresponding to a unique student who participated in the email intervention study.

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).

- The dataset is a sample of students who completed the study and interacted with the emails. Students were randomly assigned to groups, ensuring representativeness within the study's context.

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 structured data capturing student-level information, including demographic details (e.g., program type, section enrollment), academic performance metrics (e.g., quiz scores, final exam results), and behavioral variables (e.g., participation in flipped or online sections). Additionally, it includes survey-based measures of attitudes toward statistics, such as affect, interest, and perceived difficulty, both before and after the intervention, as well as details of email interventions received. This data is derived from course records, randomized experiment logs, and self-reported survey responses.

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

   - Yes. Each instance includes labels capturing the target variable, "change in effort," representing variations in student engagement. Supplementary labels include changes in affect, cognitive competence, difficulty, and interest, reflecting shifts in attitudes and behaviors post-intervention.

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. When pooling pre- and post-surveys, 23 missing data patterns were identified. The most common pattern (19% of students) involved completing all pre-SATS-36 items but none of the post-SATS-36 items, while another pattern (2% of students) involved completing all pre-SATS-36 components but only partial post-SATS-36 items. The remaining 21 patterns each accounted for less than 2% of students. These occur may due to non-responses or incomplete participation by students during data collection.

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.*

   - Yes. Relationships are made explicit through section enrollment and shared instructional methods. Students in the same section are exposed to identical teaching strategies (in person or online) and receive the same type of email intervention (plain or interesting). This allows for comparisons both within sections to assess

individual variation and across sections to evaluate the effectiveness of different instructional approaches and interventions. Additionally, relationships are indirectly established through shared demographic or performance metrics, such as participation in similar assignments or exams.

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.*

    - Yes. In 2016, the course had 1,611 enrolled students, of which 1,430 (89%) consented to participate in the study. Among these participants, 703 students (49%) were randomly allocated to the "nudge" group, while 727 students (51%) were assigned to the "no nudge" group.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Yes. Errors and noise include missing data due to incomplete survey responses, potential response biases in self-reported survey items, and redundancy in demographic data for students with identical responses within the same section.

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. Although it references study materials such as surveys and email templates, these are internal resources that accompany the dataset. This ensures reproducibility and long-term usability.

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, the dataset does not include confidential information. All data has been anonymized to protect student identities.

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, the dataset is focused on educational outcomes and does not contain offensive or distressing content.

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.*

    - Sub-populations are identified based on gender, academic year, and instructional section. These attributes are included to analyze the differential effects of the interventions.

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, the dataset is fully anonymized, and there is no risk of identifying individuals either directly or indirectly.

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.*

    - No, the dataset does not contain sensitive information such as race, religion, or financial data.

16. *Any other comments?*

    - This dataset offers valuable insights into the effectiveness of digital nudges in educational contexts, providing a resource for research on behavioral interventions in academia.

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 primarily collected through a combination of direct observation and self-reported survey responses. Academic metrics such as GPA, exam scores were obtained from institutional records, while attitudinal data such as interest, effort, affect was collected through validated pre- and post-study surveys. Email interaction data like open rates was automatically logged by the email system. Survey responses were validated by cross-referencing with institutional records to ensure consistency.

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?*

    - Data collection involved electronic surveys for capturing attitudes, an automated email tracking system to log email interactions like open rates, and administrative systems for academic performance data. Validation included manual reviews of survey completion rates and automated checks to ensure email log data was accurately recorded.

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

    - The dataset represents a random sample of students who voluntarily participated in the study. Participants were randomly assigned to either the treatment (nudge) or control (no nudge) groups. This randomization ensured unbiased comparisons and balanced representation between groups.

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)?*

    - Faculty members, graduate students, and research assistants from the Department of Statistical Sciences at the University of Toronto were responsible for collecting and managing the data. Participants (students) voluntarily contributed to the study without financial compensation. But student would receive an incentive mark of 1% to their final grade for completing pre- and post-SATS-36.

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.*

    - Data collection occurred over the duration of a single academic semester (fall 2016). This timeframe aligned with the delivery of the email interventions and the administration of pre- and post-surveys, ensuring data relevance to the study objectives.

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.*

    - Yes, the study was reviewed and approved by the University of Toronto's research ethics board. The review ensured compliance with ethical standards for research involving human participants, including informed consent, data anonymization, and protection of participant confidentiality. The study met all institutional and ethical requirements. The detailed information can be found in the study (Taback and Gibbs 2023).

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

    - Data was collected directly from participants through surveys and email logs. Academic performance data was obtained from institutional records with the necessary permissions.

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.*

    - Yes, participants were informed about the data collection process through a detailed consent form provided at the beginning of the study. The form outlined the purpose of the study, the types of data being collected, and how the data would be used and protected. The detailed information can be found in the study (Taback and Gibbs 2023).

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.*

    - Yes, participants provided informed consent electronically before participating in the study. The consent form explicitly stated that participation was voluntary and described how their data would be anonymized and used for research purposes. This study provides more detailed information (Taback and Gibbs 2023).

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).*

    - Participants were informed of their right to withdraw consent at any time during the study. They could contact the research team to revoke their participation, and any associated data would be excluded from the final dataset. This study includes more relevant information (Taback and Gibbs 2023).

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.*

    - An impact analysis was conducted as part of the ethical review process. It concluded that the dataset posed minimal risk to participants due to its anonymized nature and the focus on aggregated educational outcomes rather than individual performance. Detailed information is in this study (Taback and Gibbs 2023).

12. *Any other comments?*

- The data collection process adhered to high ethical and methodological standards, ensuring reliability, integrity, and the safeguarding of participant rights. This dataset serves as a model for conducting experimental research in educational settings.

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.*

    - Yes. Pre-processing involved standardizing survey responses, normalizing academic performance metrics, and encoding categorical variables such as email types and student demographics. Missing values were addressed by excluding instances with incomplete survey data, particularly for post-intervention responses. Outcome variables, such as "change in effort", were labeled based on differences between pre- and post-survey scores. This cleaning process was essential to ensure the dataset's reliability and validity for analysis.

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. Both raw data and processed data have been saved to allow for verification of cleaning steps and to enable future researchers to reprocess the data if needed. The raw data is securely stored and can be accessed via the research's GitHub repository: Nudge Data Repository.

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 GitHub repository for the project includes detailed documents used for pre-processing and cleaning the dataset. These resources provide transparency and guidance for replicating the pre-processing steps. They are available at: Nudge Data Repository.

4. *Any other comments?*

    - The meticulous preprocessing and labeling process reflects the rigorous standards employed to prepare this dataset. By preserving both raw and processed data, the research team has ensured flexibility and robustness for future use.

Uses

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

- The dataset has been used in academic projects focused on understanding the impact of personalized email nudges on student engagement and attitudes. These studies have provided new insights into the role of digital interventions in shaping educational outcomes and have been presented in academic forums and publications (Taback and Gibbs 2023).

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.*

   - The dataset is associated with research conducted by the University of Toronto's Department of Statistical Sciences. Relevant analyses, code, and findings are documented in the GitHub repository: Nudge Data Repository.

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

   - This dataset could be applied to a variety of tasks, including:

   1. Evaluating the scalability and effectiveness of behavioral interventions in educational contexts.
   2. Predictive modeling of student engagement and academic performance based on intervention types.
   3. Developing personalized learning strategies tailored to specific student demographics or behavior patterns.
   4. Conducting comparative studies on the effectiveness of in-person versus digital interventions.

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?*

   - Consumers should note that the exclusion of incomplete survey responses may introduce selection bias, potentially limiting the generalizability of findings. Additionally, the dataset's focus on only one course and specific intervention types may not translate directly to other educational settings. Researchers are advised to contextualize their interpretations and, where possible, validate findings with additional datasets.

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

   - The dataset should not be used for applications involving high-stakes decision-making, such as grading or admissions policies, as it was designed only for research purposes. Moreover, it shouldn't be used in contexts that could lead to unfair treatment of individuals or groups based on incomplete or biased data interpretations.

6. *Any other comments?*

    - This dataset provides a valuable resource for exploring digital engagement strategies in education but should be used with careful consideration of its limitations and ethical implications. Researchers are encouraged to adhere to the documented use cases and ethical guidelines outlined in the associated publications and repository.

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 itself isn't distributed commercially, but the dataset is for public distribution to researchers, educators, and academic institutions. It is accessible through open-access platforms to encourage further research in educational interventions and digital engagement strategies.

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

    - The dataset is distributed via academic channels. The dataset can be download from the GitHub repository: Nudge Data Repository. And it has the DOI.

3. *When will the dataset be distributed?*

    - The dataset has been available since it has been included in research articles and presentations at academic forums. It is actively maintained for ongoing and future research purposes.

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.*

    - Yes, the dataset is distributed under an open-access academic license. Users must agree to use the dataset solely for research, educational, or non-commercial purposes.

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.*

    - No third-party IP restrictions on the dataset. All data is owned and managed by the research team and is free from external constraints.

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.*

    - No export controls or regulatory restrictions on the dataset. It is fully compliant with institutional and international guidelines for open-access research data.

7. *Any other comments?*

    - The dataset is distributed following ethical guidelines and academic standards to guarantee its responsible use and wide availability for educational research purposes.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The dataset will be maintained by the research team at the University of Toronto's Department of Statistical Sciences. They are responsible for ensuring its continued availability and quality.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - The dataset can be accessed and inquiries directed through the project's GitHub repository at Nudge Data Repository. Contact information for the research team is also provided within the repository.

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

    - Any corrections or updates to the dataset will be documented in the GitHub repository Nudge Data Repository, where the dataset is stored, ensuring transparency and accessibility.

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)?*

    - The dataset may be updated or not. Updates to the dataset may be because of correcting labeling errors, addressing feedback, or adding additional data. These updates will be managed by the research team and communicated through release notes and update logs in the GitHub repository: Nudge Data Repository.

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.*

- Since the dataset is completely anonymized and doesn't include any personally identifiable information, there are no restrictions on its retention. However, all data is managed in accordance with institutional policies and ethical standards for data management.

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.*

    - Older versions of the dataset will continue to be supported accessible via the GitHub repository. Changes to availability or version support will be communicated through the repository's update logs.

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.*

    - Contributions and extensions to the dataset are welcomed and can be submitted via pull requests on the GitHub repository. All contributions will be reviewed and validated by the research team to ensure they meet the quality and ethical standards of the dataset.

8. *Any other comments?*

    - The research team is committed to maintaining the dataset as a valuable resource for educational research and to fostering collaboration among the academic community. Regular updates and active engagement with users ensure its continued relevance and usability.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Taback, Nathan, and Alison Gibbs. 2022a. "nudgedata." https://doi.org/10.5281/zenodo.1234.

———. 2022b. "Nudgedata." Zenodo. https://doi.org/10.5281/zenodo.7080181.

Taback, Nathan, and Alison L Gibbs. 2023. "A Randomized Study to Evaluate the Effect of a Nudge via Weekly e-Mails on Students' Attitudes Toward Statistics." *Journal of Statistics and Data Science Education* 31 (2): 134–43.