



Research Report

Do Volunteers Represent the Full Population? Volunteer Bias in Clickstream-Based Learning Analytics.

Lennox Victor Ruffiner

Supervised by Dominik Glandorf

ML4ED-Lab

Lab head: Dr. Tanja Käser

`Lennox.ruffiner@epfl.ch`

October 2026

Disclaimer: This document represents an ongoing research project. The data, analysis, and findings presented herein are preliminary and subject to further development.

Abstract

Learning analytics relies on the collection and analysis of student data to optimize learning environments. However, ethical and practical constraints dictate that data is typically collected only from students who volunteer to participate in studies. This reliance on self-selected volunteers introduces the risk of non-response bias, a widely prevalent problem for educational data [11]. If volunteers systematically differ in their behaviors and academic outcomes from non-volunteers, models trained exclusively on volunteer data may not generalize to the broader student population.

This ongoing research assesses this potential bias by utilizing a unique dataset from a flipped-classroom course. By mapping anonymized clickstream data to survey participation timestamps, the study observed the digital behavior of an entire class cohort, dividing them into non-volunteers, volunteers, and super-volunteers. Preliminary results indicate that while the behavioral distributions (e.g., how students consume videos) remain largely similar across these subgroups, the academic outcomes tied to these behaviors differ. A grade prediction model trained on super-volunteers failed to accurately predict the grades of the other subgroups, despite demonstrating high statistical confidence in its predictions. This suggests the presence of a “concept shift,” where identical behaviors lead to different outcomes depending on a student’s volunteer status, highlighting a critical limitation in current educational data mining models.

1 Introduction

1.1 Context

Learning analytics is defined as the measurement, collection, and analysis of data about learners and their contexts, for the purposes of understanding and optimizing learning and the environments in which it occurs. This field has grown continuously alongside the expansion of online learning. When students use digital platforms, such as Massive Open Online Courses (MOOCs) or flipped classrooms, their digital footprints—clicks, navigation patterns, and time spent on tasks—can be logged and measured.

For ethical and practical reasons, this detailed data is usually collected only from students who have explicitly agreed to participate in a research study. These self-selected volunteers form the foundation of most predictive models used in educational research today.

1.2 The Problem: Non-Response Bias

Relying on volunteers presents a significant methodological challenge known as non-response bias. If these volunteers differ systematically in their learning behaviors and outcomes from the students who refused to participate, the resulting data will be skewed. Consequently, attempts to understand student learning or optimize educational environments may not generalize to every learner. As Robinson [9] points out, biased non-response is particularly detrimental to model performance.

Non-response bias is a fundamentally difficult problem to tackle. By definition, researchers generally cannot effectively measure this bias because the data for non-volunteers is absent. In the fields of learning analytics and Educational Data Mining (EDM), very few papers have addressed non-response bias beyond simply reporting the response rate [5]. Using “shared auxiliary variables” available for both respondents and non-respondents is considered the most effective way to evaluate and correct this bias [8]. Determining which aspects of a student’s clickstream behavior are biased, identifying if specific features can be selected to reduce this bias, and understanding how this impacts learner profiles would be of great help to the scientific community.

1.3 Objective and Approach

This study addresses this need by leveraging a unique data collection situation. By default, the course platform used at the École Polytechnique Fédérale de Lausanne (EPFL) collects anonymized clickstream data (video logs) for every student. In the context of a separate study, students were asked if they wanted to volunteer to answer occasional popup surveys regarding their study habits (e.g., “*Pourquoi avez-vous mis la vidéo en pause ?*”).

This setup made it possible to retroactively split the anonymized clickstream of the entire class into specific subgroups: those who did not volunteer, those who volunteered for the survey, and those who volunteered for the survey and shared their grades. We modeled the behavior of the entire class cohort across these subgroups, compared their behaviors, and measured the performance of a predictive algorithm trained exclusively on the volunteers.

2 Literature Review

Any study in the field of education relies on data, and that data relies on volunteers. If volunteer data fails to generalize, the validity of educational studies can be questioned. Furthermore, if institutional prediction models perform poorly on non-volunteers due to a lack of representative data, it may foster distrust. As Cofone [2] notes, algorithmic discrimination is an information problem. A failing model might push non-volunteers into deeper distrust of the institution collecting the data, making them less likely to volunteer in the future, which in turn degrades the model further—a negative spiral.

2.1 The Nature of Non-Response Bias

Research shows that non-response bias is real and can sometimes be large [4]. Historically, the most common prescription for survey researchers has been to simply minimize non-response rates. However, Groves demonstrates that the raw non-response rate is actually a weak predictor of the magnitude of non-response bias. Instead, non-response bias is a function of the covariance between the propensity to respond (how likely someone is to volunteer) and the variables involved in the survey estimate.

Importantly, covariances are highly variable across different items within the same survey. As Groves & Peytcheva [3] explicitly state, “nonresponse bias is a property of a specific statistic, not of the survey as a whole.” This mathematical reality justifies examining specific variables to see if certain behavioral features are “non-response bias proof” (RQ3). Checking how respondents differ from the frame using logged data is far more informative than guessing from response rates alone.

2.2 Differences Between Volunteers and Non-Volunteers

Several studies have documented differences between those who consent to research and those who do not. In medical education, Callahan [1] found that volunteers are more likely to outperform others on some performance measures, and that this difference, while small, is statistically significant and stable over decades. Similarly, Callan et al. [?] found that students who provided active informed consent in educational research demonstrated significantly higher academic achievement (GPA and standardized course examination grades) with moderate to large effect sizes. Verhulst [12] observed similar trends in residency follow-ups, interpreting that lower-performing students disproportionately withhold permission because of their poor performances.

Regarding attitudes, Rogelberg [10] found that “active non-respondents” (those who purposefully and consciously make a decision not to respond) generally exhibit lower organizational commitment, lower satisfaction, and more negative beliefs about how an organization handles data. Interestingly, these non-respondents did not differ from volunteers in demographics or work context (gender, education, tenure), suggesting that attitude and behavior, rather than basic demographics, drive non-response. Li et al. [6] similarly identified that a student’s willingness to volunteer is influenced by their degree of trust in the institution and their comfort with instructors using data to gain insights.

2.3 Predictive Features in Digital Learning

To measure student behavior in digital environments, researchers look to specific interactions. Marras [7] demonstrated that predicting success in flipped classes requires granular, detailed video interaction data, whereas MOOCs typically rely on broad study regularity. Both formats universally rely on quiz data when available. Our study utilizes these established granular video interaction features to compare subgroups.

2.4 Hypotheses and Objectives

Based on the literature, we formulated three main research questions (RQs):

- **RQ1:** Does behavior differ between volunteers and non-volunteers?
- **RQ2:** Does a model trained on volunteers generalize to the entire population?
- **RQ3:** What specific behavior is biased, and how can one reduce self-selection bias through feature selection?

We hypothesized that volunteers do not perfectly represent the whole population. Specifically, we hypothesized they are more organized and more likely to follow schedules (altering *when* they consume course materials). However, we hypothesized there would be no significant bias in *how* they consume videos (e.g., video speed, pauses). Finally, aligning with prior literature [1, ?], we hypothesized that the grades of volunteers would be higher than those of non-volunteers.

3 Materials and Methods

3.1 Subjects

The subjects of this study were 247 first-year students enrolled in a linear algebra flipped-classroom course at EPFL.

3.2 Data Collection and Subgroup Mapping

The base data consisted of course-spanning clickstream logs (records of students playing, pausing, seeking, and changing video speeds) collected anonymously for all students.

Concurrently, an Ecological Momentary Assessment (EMA) study asked these students if they wanted to volunteer to answer popup surveys triggered by video events. For example, pausing a video might trigger the prompt: “*Pourquoi avez-vous mis la vidéo en pause ?*” (Why did you pause the video?) with options such as “Pour comprendre le contenu” (To understand the content) or “Pour résoudre un problème / faire un exercice” (To solve a problem / do an exercise). Similar prompts were coded for rewinding, skipping, and speed changes.

Students were further asked if they would consent to share their final grades for research. This process split the class into three distinct subgroups of active responders/non-respondents:

1. **Non-volunteers:** Students who refused to participate in the popup study and refused to share their grades.
2. **Volunteers:** Students who agreed to participate in the popup study but refused to share their grades.
3. **Super-volunteers:** Students who agreed to participate in the study and agreed to share their grades.

Because the EMA survey answers were deanonymized to the researchers, we possessed the exact timestamps of when a volunteer or super-volunteer hit “pause” to trigger a survey. By mapping these specific timestamp IDs back to the completely anonymized total clickstream logs, we established an injective mapping. This allowed us to categorize the anonymous, total-course clickstream data into the three volunteering statuses, providing shared variables across all groups.

3.3 Data Processing and Feature Engineering

To ensure the data represented active learning, we filtered the clickstream to remove students who had dropped out. A student was removed if they met any of the following criteria: inactive after week 12, less than 500 total minutes of watch time, or fewer than 50 total clickstream logs.

For the remaining students, we computed 20 behavioral features. These metrics were chosen based on prior research demonstrating their efficacy for grade prediction in flipped classrooms [7], focusing on granular interaction data. The methodology also allows for the future computation of multidimensional student profiles.

3.4 Research Procedure and Statistical Design

To analyze the data, we conducted the following procedures:

1. **Distribution Tests:** We performed permutation tests to compare the distributions of each individual behavioral feature between the subgroups.
2. **Multivariate Tests:** We ran multivariate tests to observe the distribution differences of the feature space as a whole between subgroups.
3. **Covariance Tests:** We tested the covariance of specific features against volunteering status.
4. **Volunteer Prediction:** We trained a logistic regression model to determine if volunteering status could be predicted solely based on behavioral features.
5. **Grade Prediction Model (RQ2):** We trained a Gaussian Process (GP) using the behavioral features and grades of the “super-volunteers.” The GP was chosen because it returns a probability distribution with a standard deviation, allowing us to assess how confident the model is in its predictions.
6. **Model Evaluation:** We ran the trained GP model on the behavioral data of the non-volunteers and standard volunteers, comparing the distribution of the predicted grades against the actual real-world grade distribution of the course.

4 Analysis and Findings

Note: The study is ongoing, and these results are preliminary. Profiling has not yet been implemented, and feature extraction has not yet been exhaustively run on all available data.

4.1 Academic Outcomes Contrasting with Literature

Contrary to standard findings in educational and medical literature which suggest volunteers perform better academically [1, ?, 12], our preliminary data showed that the grades of volunteers in this specific dataset were, on average, *lower* than the grades of non-volunteers.

It is important to note that in Verhulst’s study, non-response bias was attributed to lower-performing students withholding permission for follow-ups *because* of their poor performances. In our dataset, students had to make the active decision to volunteer or decline *before* knowing their performance. This suggests that in a demanding engineering context, high-performing students might decline to volunteer at the start of the semester to avoid losing study time.

4.2 Behavioral Distributions

When examining the clickstream behaviors, the distribution between volunteers and non-volunteers did not appear to differ significantly.

- **Permutation Tests:** Out of the 20 behavioral features tested, only one feature (“course anticipation,” scored 0 to 1 based on watching videos prior to deadlines) yielded a statistically significant p-value. Given that this was only 1 feature out of 20, it could be a random statistical artifact.
- **Multivariate Tests:** The multivariate tests comparing the distributions between subgroups were not statistically significant, supporting the observation that digital behavior is uniform across groups.

4.3 Predicting Volunteer Status

We trained a logistic regression model to predict whether a student was a volunteer based on their clickstream behavior. The model achieved a Precision-Recall Area Under the Curve (PR AUC) of 0.63, with a standard deviation of 0.05. This indicates that predicting volunteer status based purely on behavior is only slightly better than random chance.

4.4 Model Generalization and Confidence

The results of the grade prediction task (RQ2) yielded the most significant findings:

- **Performance on Super-Volunteers:** When the Gaussian Process (GP) model was tested on super-volunteers, the distribution of the predicted grades versus their real grades yielded an insignificant p-value (around 0.5), meaning the distributions matched closely.
- **Failure on the General Population:** When the same model was applied to the non-volunteers and standard volunteers, it failed to generalize. Comparing

the predicted grades for this group against the known course distribution resulted in a statistically significant p-value of 0.03.

- **Misplaced Model Confidence:** The standard deviation of the model’s output was, on average, the same regardless of whether the input data belonged to a volunteer, super-volunteer, or non-volunteer. The model was as confident in its incorrect predictions for non-volunteers as it was in its correct predictions for super-volunteers.

5 Conclusion

This research aims to directly assess non-response bias in learning analytics by analyzing shared behavioral variables among volunteers and non-volunteers.

The preliminary findings show that while the distribution of behavioral features and their covariance remains largely identical across subgroups, a predictive model trained on these behaviors fails when applied to the broader population. The model successfully predicts outcomes for super-volunteers but confidently fails to predict the grades of non-super-volunteers.

This indicates the presence of a **concept shift**:

$$P(y|x, \text{super-volunteers}) \neq P(y|x, \text{non-super-volunteers}) \quad (1)$$

The exact same digital behavior (e.g., pausing and rewinding) leads to a different academic outcome depending on the volunteering status of the student being observed.

Because the underlying behavioral data looks the same, the algorithm cannot distinguish between the contexts. We term this a “silent failure.” The predictive model fails to generalize to non-volunteers, yet its internal confidence metrics remain high. If such a model were deployed, it would confidently misclassify non-volunteers, potentially depriving them of necessary interventions and initiating the negative spiral of algorithmic discrimination and distrust outlined by Cofone [2]. This confirms that studies relying strictly on consenting cohorts operate under a bias that cannot be solved simply by reporting the non-response rate.

However, consistent with Groves & Peytcheva [3], non-response bias is a property of specific statistics. Finding that certain features (and not others) carry this bias gives hope that specific feature selection may eventually mitigate self-selection bias (RQ3).

6 Future Works

To further validate these preliminary findings, the following future research steps are planned:

1. **Longitudinal Data Collection:** Collect additional data from the same linear algebra course in a subsequent academic year to verify the stability of these findings.

2. **Master's Level Analysis:** Apply the exact same methodology to data from a Master's level course to observe if academic maturity alters the concept shift between volunteers and non-volunteers.
3. **Demographic Integration:** Collect and integrate demographic data for both volunteers and non-volunteers to investigate whether specific demographics drive active non-response. If demographics do not differ (as suggested by Rogelberg's organizational survey findings), it reinforces that attitude and behavior are the primary drivers.
4. **Multidimensional Clustering:** Implement multidimensional student profiling to move beyond simple behavioral features. The goal is to determine if more complex learner profiles are less susceptible to concept shift and can be used to reduce bias.
5. **Methodological Refinement:** Continue to refine the statistical methods to ensure rigor in testing covariance and evaluating bias reduction techniques.

References

- [1] Clara Callahan, Mohammadreza Hojat, and Joseph Gonnella. Volunteer bias in medical education research: an empirical study of over three decades of longitudinal data. *Medical Education*, 41:746–753, 07 2007.
- [2] Ignacio Cofone. Algorithmic discrimination is an information problem. *Hastings Law Journal*, 70:1389, 2019.
- [3] Robert Groves and Emilia Peytcheva. The impact of nonresponse rates on nonresponse bias: A meta-analysis. *Public Opinion Quarterly*, 72:167–189, 05 2008.
- [4] Robert M. Groves. Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5):646–675, 01 2006.
- [5] J. R. Halbesleben and M. V. Whitman. Evaluating survey quality in health services research: A decision framework for assessing nonresponse bias. *Health Services Research*, 48(3):913–930, 2013.
- [6] Warren Li, Christopher Brooks, and Florian Schaub. The impact of student opt-out on educational predictive models. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, page 411–420, New York, NY, USA, 2019. Association for Computing Machinery.
- [7] Mirko Marras, Julien Tuan Tu Vignoud, and Tanja Käser. Can feature predictive power generalize? benchmarking early predictors of student success across flipped and online courses. In *Proceedings of the 14th International Conference on Educational Data Mining*, 2021.
- [8] Thomas O’Toole, Alexandru Cernat, Nikos Tzavidis, Natalie Shlomo, Joseph Sakshaug, and Peter Barlow. Survey practice guide 5: Using integrated non-survey data for evaluating and correcting for non-response, February 2026.
- [9] T. S. Robinson, N. Tax, R. Mudd, et al. Active learning with biased non-response to label requests. *Data Mining and Knowledge Discovery*, 38:2117–2140, 2024.
- [10] Steven Rogelberg, Alexandra Luong, Matthew Sederburg, and Dean Cristol. Employee attitude surveys: Examining the attitudes of noncompliant employees. *Journal of Applied Psychology*, 85:284–293, 04 2000.
- [11] Yajuan Si, Roderick J. A. Little, Ya Mo, and Nell Sedransk. A Case Study of Nonresponse Bias Analysis In Educational Assessment Surveys. *arXiv e-prints*, page arXiv:2104.04432, April 2021.
- [12] S. J. Verhulst and L. H. Distlehorst. Examination of nonresponse bias in a major residency follow-up study. *Academic Medicine*, 68(2 Suppl):S61–S63, 1993.