Applying machine learning models in multi-institutional studies can generate bias

Rebeckah K. Fussell, Meagan Sundstrom, Sabrina McDowell, and N. G. Holmes
Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA

There is increasing interest in deploying machine learning models at scale for multi-institutional studies in physics education research. Here we investigate the efficacy of applying machine learning models to institutions outside of their training set, using natural language processing to code open-ended survey responses. We find that, in general, changing institutional contexts can affect machine learning estimates of code frequencies: either previously documented sources of uncertainty increase in magnitude, new unknown sources of uncertainty emerge, or both. We also find an example where uncertainties do not change between the institution used in the training data and an institution not in the training data. Results suggest that attention to uncertainty is critical, especially when making measurements of student writing across multi-institutional data sets.

2024 PERC Proceedings edited by Ryan, Pawl, and Zwolak; Peer-reviewed, doi.org/10.1119/perc.2024.pr.Fussell

I. INTRODUCTION

Machine learning methods, including natural language processing (NLP), may be useful for physics education research (PER) studies that analyze student writing, such as open-ended survey responses. NLP can unlock the ability to analyze these rich data sources at scale, across institutions.

Much of the existing PER work has studied individual institutions one at a time, with over-sampling from institutions that primarily serve white and wealthy populations [1]. The generalizability of PER, therefore, is limited by the lack of multi-institutional studies and the under-representation of minority-serving institutions, primarily undergraduate institutions, and two-year colleges. Analyzing student writing from many diverse institutions requires a scale that is typically impractical of human coders. Machine coding may be used to include more and diverse institutions in PER studies.

The efficacy of using machine learning tools across multiple institutions, however, remains untested. Most PER literature that evaluates machine learning methods uses data from a single institution [e.g., 2–7]. Other studies aggregate students from multiple institutions into a single sample [e.g., 8–10], without investigating the effects of this aggregation.

Yet in other contexts, there are documented examples of bias in models when the population under study changes [e.g., 11]. It is therefore reasonable to hypothesize that a similar bias may arise in multi-institutional machine learning studies in PER. Here we seek a quantitative way of investigating this hypothesized population-related bias. We define population-related bias as sources of uncertainty associated with using machine learning with different student populations.

We build on prior work investigating two (non-population-related) sources of uncertainty in machine measurements [12]. The two uncertainties are *sampling error* and the *linear systematic bias effect* (abbreviated LSBE in this paper). We evaluate population-related uncertainty by studying if any uncertainty remains after accounting for these two non-population-related uncertainties. In our study, the machine measurements are estimates of the frequency of a code in a sample, denoted E_C, rather than individual assessments, because we are studying bias on the population, not student, level. An alternate approach would be to hand-code all data used for a measurement to check machine against human measurement, but this defeats the intended purpose of using machine learning for efficiency at scale. We aim to limit hand-coding to the minimum amount needed to train a model and to evaluate uncertainties.

The first source of uncertainty, sampling error, is the natural variability associated with the fact that a given sample is finite. Given a frequency estimate E_C, the standard deviation σ_{E_C} describes the spread in repeated measurements of E_C. Without the ability to take repeated measurements, the exact value of the sampling error σ_{E_C} is hidden. In the case of binomial data (e.g., applying binary codes to short survey responses), past work has empirically shown that σ_{E_C} can be approximated as:

$$\sigma_{E_C} \approx \sqrt{\frac{E_C(1-E_C)}{n}} \tag{1}$$

where n is the number of responses in a sample [12].

The second source of uncertainty is systematic error in the frequency estimate E_C captured by the LSBE. Bias in frequency estimates is not often discussed in machine learning literature, where the focus is optimizing sets of individual predictions, not frequency estimates. However, researchers have documented biases that arise in frequency estimates [13, 14]. Previous work defined this bias as a function of the frequency estimate E_C, computed as $S = E_C - E_H$, where E_H is the measurement human coders made in the same sample [12]. They found a negative linear relationship between S and E_C; the model tended to overestimate for low values of E_C and underestimate for high values of E_C. They also describe a process to measure the LSBE: hand-code a fraction of the data, take samples of this fraction of data with varying E_H, plot S versus E_C for each sample, and find the line of best fit (a function of E_C). We denote this line of best fit $LSBE(E_C)$ or $LSBE$. The parameters of the best fit line $LSBE$ fluctuate as the samples fluctuate, but the line of best fit approximates an underlying LSBE.

Our goal is to evaluate the responsibility of making measurements from a multi-institutional dataset using machine learning. Given these two sources of *non-population-related* uncertainty in machine measurement, we seek to understand *population-related* uncertainties in a multi-institutional dataset. We are guided by the following research question: Can we apply the same machine learning algorithm trained at one institution to data from multiple institutional contexts?

II. METHODS

A. Data sources

We use data from an open-ended survey prompt asking physics lab students about their peer interactions (adapted from [15]). Students were asked to "Please list any students in your physics lab class that you had a meaningful interaction with about course material this week." For each name they listed, they were asked in a separate text box, "What aspects of the lab material did you discuss with this person?" [16, 17]. This question was initially asked only at University 1 (U1). Coder group A developed a coding scheme for students' written explanations. They established inter-rater reliability and applied the coding scheme by hand to the 1216 responses [17]. In this paper we focus on the two most frequent codes: small group work (SGW) and concepts (CON).

The following year, this survey was administered again at U1 and, for the first time, at three additional institutions. All four courses in this study are introductory labs at R1 universities, and U4 is a Hispanic-serving institution. A summary of the number of responses from each institution is provided in

TABLE I: Data subsets used in the study. N_{total} is the total number of responses collected at that university and N_{test} is the number that were hand coded to use as test data.

	Institution	Coder group	N_{total}	N_{test}
Training data	University 1 (U1)	A	1216	-
New data	University 1 (U1)	B	339	68
	University 2 (U2)	B	80	40
	University 3 (U3)	B	234	62
	University 4 (U4)	B	476	91

Table I. Coder group B (a different team of researchers than coder group A) used the original coding scheme to code a test data set. We deliberately kept the size of this test data set small relative to the total number of available responses to limit hand-coding. Coder group B worked to code in a manner consistent with coder group A, and established inter-rater reliability on a test data set of 261 unique coded responses from these four universities.

B. Machine learning methods

We trained a Bidirectional Encoder Representations from Transformers (BERT) classifier model using the 1216 hand-coded responses from U1 (coder group A) as training data to perform the machine coding. We added a 50-neuron classifier layer to the BERT Base (110 parameter) model [18]. The codes in our coding scheme were not mutually exclusive, so we trained separate classifier models for each code, such that one model determined if the SGW code was present and a separate model determined if the CON code was present. We balanced the training by weighting the responses in the training set, as per recommendations in Ref. [12], so that there was equal probability that a sentence with or without the code (as determined by the original human coder) would be selected from a sample of training data.

C. Analysis methods

From the hand-coded test data for each institution, we took different samples i of size $n = 15$ and evaluated the relationship between the algorithm's estimate E_C and the difference between the algorithm's estimate and coder group B's estimate $S = E_C - E_H$ (as seen in [12]). The samples were selected such that E_H values were stratified between 0.1 and 0.9, in increments of 0.1. We measured the LSBE by evaluating the best linear fit to the data from each institution for each code. If the best fit line had a negative slope and the data were largely linear, we inferred that the best linear fit is a reasonable approximation of the LSBE. It is an approximation of the underlying LSBE because statistical fluctuations in the data may change the exact parameters of the linear fit.

We expected a sampling error that follows Eqn. 1. One limitation of the approximation in this equation is that a single measurement of E_C may equal 0 or 1 even if E_H is not 0 or 1, especially because our n is small, $n = 15$. In these cases, strictly following Eqn. 1 would result in an unrealistic sampling error of zero. Therefore, we computed the sampling error in these cases using $E_C = 0.99$ (or $E_C = 0.01$, which results in the same sampling error due to symmetry).

Guided by our research question, we sought to evaluate whether these two uncertainties–the LSBE and binomial sampling error–were similar when applying the machine coding across institutions. We used a weighted chi-squared test to evaluate whether the LSBE for U1, represented by $LSBE_{U1}(E_C)$, fit the other institutional data within uncertainties. The weighted chi-squared test statistic is

$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{(S_i - LSBE(E_{C,i}))^2}{(\sigma_{E_{C,i}})^2} \tag{2}$$

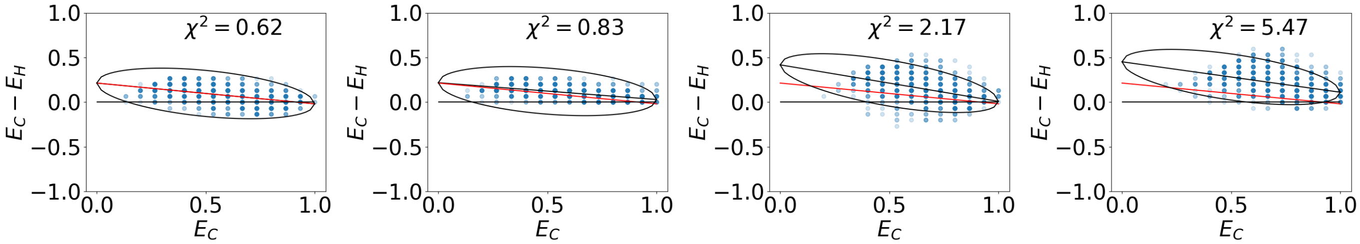
where N is all samples in the plot (adapted from [19]).

The chi-squared test statistic corresponds to the square of the number of standard deviations between the data points $(E_{C,i}, S_i)$ and the predictions at these points computed by the best fit function $LSBE(E_{C,i})$. Therefore, if $\chi^2 \lesssim 1$, we interpret this to mean the $LSBE$ function is sufficient to describe the relationship between E_C and S, accounting for sampling error. If $\chi^2 >> 1$, we interpret this to mean that the $LSBE$ function is insufficient to describe the relationship, accounting for sampling error. In this case, either changes to the $LSBE$ function are required or additional sources of uncertainty need to be considered.

We test all institutions against the LSBE fit for the test data from U1 because this was the same institution as the training data. The LSBE fit for U1, therefore, provides a measure of the non-population-related systematic uncertainty associated with the machine coding. If, from this chi-squared analysis, we determine that the LSBE and sampling error are sufficient to describe the uncertainty in machine coding for a particular data set, we infer that the non-population-related uncertainties sufficiently capture any systematic variability in the data, accounting for sampling error. This does not mean that other sources of uncertainty are not present; it just means that any other sources of uncertainty are small relative to these two sources of uncertainty. If we instead find that the LSBE and sampling error are not sufficient to describe the uncertainty in machine coding, we then infer not only that other sources of uncertainty are present in that data, but also that these uncertainties are large enough that it is necessary to change the methods or further investigate the uncertainties. Analysis code is available on GitHub [20].

III. RESULTS

In this section, we describe the results for the SGW code followed by the CON code. We will first evaluate the LSBE for U1 ($LSBE_{U1}$) and then evaluate whether or not $LSBE_{U1}$ applies to the other institutions.



(a) University 1 (b) University 2 (c) University 3 (d) University 4

FIG. 1: Systematic error vs. computer estimate for the SGW code at each institution. The black line represents the line of best fit and the black oval is two times the statistical uncertainty. The red line represents the best fit line for the U1 data (first panel). The chi-squared value reported is the weighted chi-squared between the data and the red line from U1.

The U1 data has a negative linear trend with best fit line $S(E_C) = -0.23E_C + 0.21$ (denoted $LSBE_{U1,SGW}$) for the SGW code (Fig. 1a). This means that if SGW is less common in a sample, the BERT model tends to over-estimate the measurement and the magnitude of this overestimating gets smaller as SGW becomes more common (E_C increases). The chi-squared value is 0.62, suggesting a good fit to the data within sampling error and, thus, that the best fit line is a good approximation of the underlying LSBE for the model. If there are sources of uncertainty in addition to the LSBE and sampling error, they are not large enough to matter in comparison.

Next, we evaluate if the function $LSBE_{U1,SGW}$ can be generalized to other institutions. $LSBE_{U1,SGW}$ fits the data from U2 well, with a chi-squared value of 0.83 (Fig. 1b). Additionally, the best fit line for the data from U2 is quite similar (both slope and intercept) to the parameters for $LSBE_{U1,SGW}$. Almost all points fall within two uncertainty intervals of the respective best fit line for Universities 1 and 2 (black ovals; Figs. 1a and 1b). Any uncertainties in addition to LSBE and sampling error must be small in this case.

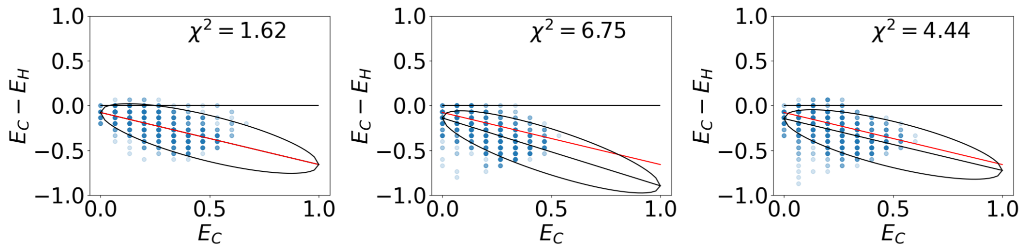
The data from Universities 3 and 4, in contrast, are not well fit by $LSBE_{U1,SGW}$. The best fit line for U3 data is steeper and has a much higher intercept compared to $LSBE_{U1,SGW}$ (Fig. 1c). The chi-squared value between the U3 data and $LSBE_{U1,SGW}$ is 2.17, suggesting concerns with fit. This chi-squared value may also be an underestimate because we were not able to obtain as many samples in the region of greatest disagreement between the two lines (E_C<0.4). The best fit line for U4 is slightly steeper than $LSBE_{U1,SGW}$ and has a much higher intercept. The chi-squared value between the data and $LSBE_{U1,SGW}$ is 5.47, suggesting a poor fit (Fig. 1d). Both of these fits indicate that the underlying LSBEs for U3 and U4 are different than $LSBE_{U1,SGW}$. Additionally, there are many samples that do not fall within two uncertainty intervals of the respective best fit line in both Universities 3 and 4 (black ovals; Figs. 1c and 1d). Together, this evidence suggests the presence of additional sources of uncertainty, beyond the LSBE and sampling error.

We now turn to the CON code as another example. Again, the U1 data has a negative linear trend, now with best fit line $S(E_C) = -0.58E_C - 0.08$ (denoted $LSBE_{U1,CON}$) and almost all data points below the $S = 0$ line (Fig. 2a). This means that if CON is less common in a sample, the BERT model tends to underestimate the measurement and the magnitude of this underestimate gets larger as CON becomes more common (E_C increases). The chi-squared value is 1.62, greater than 1, suggesting uncertainties in addition to the LSBE and sampling error. In particular, some data fall outside of the sampling error of the best fit line (black oval; Fig. 2a), which we will investigate in future. For now, $LSBE_{U1,CON}$ may be an acceptable approximation of the underlying LSBE because these outliers lie both above and below the oval.

Next, we evaluate if $LSBE_{U1,CON}$ can be generalized to other institutions. We could not evaluate the CON code for U2 as it is less frequent than SGW and we had insufficient test data (Table I). Figure 2 shows that $LSBE_{U1,CON}$ does not fit the data from Universities 3 and 4. $LSBE_{U1,CON}$ has a chi-squared value of 6.75 when used to model the U3 data (Fig. 2b). The slope of the best fit line for U3 is slightly steeper and the intercept is more negative than $LSBE_{U1,CON}$. $LSBE_{U1,CON}$ has a chi-squared value of 4.44 when used to model the U4 data (Fig. 2c). The intercept of the best fit line for U4 is more negative than $LSBE_{U1,CON}$, though the slopes are similar. Both institutions have samples far enough away from their respective lines of best fit (points outside the black ovals) to suggest that additional sources of uncertainty–beyond sampling error and LSBE–are at play, even more so than we saw with U1 (Fig. 2).

IV. DISCUSSION

We have found that applying models to new institutions without retraining can miss sources of error. In some contexts this can result in biased conclusions. Our data show that it could be responsible to use this BERT model to make measurements of the SGW code on the data from U1 and U2. The LSBE function, accounting for sampling error, sufficiently described the relationship between S and E_C for those two institutions, indicating that the non-population-related sources of uncertainty are likely the major sources of uncer-



(a) University 1 (b) University 3 (c) University 4

FIG. 2: Systematic error vs. computer estimate for the CON code at each institution.The black line represents the line of best fit and the black oval is two times the sampling error. The red line represents the best fit line for the U1 data (first panel). The chi-squared value reported is the weighted chi-squared between the data and the red line from U1.

tainty in the machine coding. Our analysis, however, suggested there are additional sources of uncertainty when applying the SGW code to data from Universities 3 and 4 and the CON code to data from Universities 1, 3, and 4. Therefore, it would be irresponsible to make measurements for these codes at these universities without additional work to understand or eliminate these uncertainties. We elaborate on these findings below and identify recommendations for future work.

First, our analyses suggest that changing institutional context may change the underlying linear systematic bias effect. We saw one example (U2, SGW code) where the equation for the LSBE was essentially unchanged from U1 (Fig. 1b). For the other two institutions, the LSBE from U1 was not sufficient to account for systematic errors in machine coding of SGW (Figs. 1c and 1d). In particular, the best fit line parameters changed for Universities 3 and 4; the intercepts moved away from zero and the slopes became steeper. The magnitude of systematic error, in general, got worse. We would need to analyze aspects of the populations from each institution to understand what factors contributed to $LSBE_{U1,SGW}$ applying to some datasets and not others.

In addition to the changes to the LSBE parameters, we observed samples that were more than two standard sampling errors away from the linear systematic effect (some data points are outside the black ovals and the linear fits do not necessarily represent the data). This suggests the presence of additional uncertainties not accounted for by the paradigm laid out in prior work [12]. We consider two possible explanations.

First, the additional variance in machine measurement may be because of features in the data at these institutions that were missing from the training data. For example, students in different instructional contexts may have identified ways of discussing small group work (SGW) that were not represented in the U1 data. Second, factors other than institutional context may have led to this additional uncertainty. Many samples of the CON code from U1, for example, also fell outside two standard sampling errors from the line of best fit (Fig. 2a), despite being from the same institution as the training data. Possible explanations include the change in year between the training and test data or the change in coder group. In some cases, this additional variability may be systematic and not random. For example, more samples fall below the oval than above the oval for the U3 CON code data (Fig. 2c). We would need to analyze the responses in more detail, not blind to institution, to disentangle these effects.

It is not necessary to understand every potential source of uncertainty if the uncertainty can instead be sufficiently reduced. For example, we could have coder group B recode the training data for the CON code to eliminate the potential effect of coding group on the uncertainty. Alternatively, we could retrain the model with samples from each institution to reduce the effect of instructional context on the uncertainty. Each of these place limitations on the scalability of machine coding. Ultimately, it is important to check if a machine learning algorithm trained at one institution and coded by one group of researchers can be responsibly applied to new populations of students without additional human coding.

V. CONCLUSION

We have found that uncertainties can arise when applying machine learning models to different institutions. We expect that similar uncertainties also exist in coding schemes *applied by humans*. Similar methods could be used in future to evaluate such biases.This analysis could complement or replace the calculation of reliability metrics (e.g., Cohen's kappa) because these metrics do not reveal the direction of discrepancies (if a measurement will be over- or under-estimated), only the presence of discrepancies. Relying on human coding instead of machine coding is not a solution to the problem of systematic bias. We encourage all researchers to carefully consider sources of uncertainty in all studies.

ACKNOWLEDGMENTS

Thank you to L. N. Simpfendoerfer, and to Sarah McHale, Olive Ross, and Emily M. Stump for thoughtful comments. We thank Eric Burkholder, Andrew Loveridge, and Danny Doucette for data collection efforts. We acknowledge funding from NSF GRFP DGE-2139899 and NSF DUE-1836617.

[1] S. Kanim and X. Cid, Demographics of physics education research, Physical Review Physics Education Research **16**, 020106 (2020), publisher: American Physical Society.

[2] J. Wilson, B. Pollard, J. M. Aiken, M. D. Caballero, and H. Lewandowski, Classification of open-ended responses to a research-based assessment using natural language processing, Physical Review Physics Education Research **18**, 010141 (2022).

[3] R. Fussell, A. Mazrui, and N. G. Holmes, Machine learning for automated content analysis: characteristics of training data impact reliability (2022) pp. 194–199, iSSN: 2377-2379.

[4] M. Thomas, S. Bagley, and M. Urban_Lurain, Using Machine Learning Algorithms to Categorize Free Responses to Calculus Questions, Proceedings of the Twenty-first Annual Conference on Research in Undergraduate Mathematics Education (2018).

[5] A. Bralin, J. W. Morphew, C. M. Rebello, and N. S. Rebello, Analysis of student essays in an introductory physics course using natural language processing, in *2023 Physics Education Research Conference Proceedings* (American Association of Physics Teachers, Sacramento, CA, 2023) pp. 58–63.

[6] J. Campbell, K. Ansell, and T. Stelzer, Evaluating IBM's Watson natural language processing artificial intelligence as a short-answer categorization tool for physics education research, Physical Review Physics Education Research **20**, 010116 (2024), publisher: American Physical Society.

[7] J. M. Geiger, L. M. Goodhew, and T. O. B. Odden, Developing a natural language processing approach for analyzing student ideas in calculus-based introductory physics (2022) pp. 206–211, iSSN: 2377-2379.

[8] T. D. Ullmann, Automated Analysis of Reflection in Writing: Validating Machine Learning Approaches, International Journal of Artificial Intelligence in Education **29**, 217 (2019).

[9] J. M. Rosenberg and C. Krist, Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations, Journal of Science Education and Technology **30**, 255 (2021).

[10] C. M. Nakamura, S. K. Murphy, M. G. Christel, S. M. Stevens, and D. A. Zollman, Automated analysis of short responses in an interactive synthetic tutoring system for introductory physics, Physical Review Physics Education Research **12**, 010122 (2016).

[11] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, A Survey on Bias and Fairness in Machine Learning, ACM Computing Surveys **54**, 115:1 (2021).

[12] R. K. Fussell, E. M. Stump, and N. Holmes, Method to assess the trustworthiness of machine coding at scale, Physical Review Physics Education Research **20**, 010113 (2024), publisher: American Physical Society.

[13] B. C. Wallace and I. J. Dahabreh, Improving class probability estimates for imbalanced data, Knowledge and Information Systems **41**, 33 (2014).

[14] N. T. Young and M. D. Caballero, Predictive and Explanatory Models Might Miss Informative Features in Educational Data, Journal of Educational Data Mining **13**, 31 (2021).

[15] J. P. Zwolak, M. Zwolak, and E. Brewe, Educational commitment and social networking: The power of informal networks, Phys. Rev. Phys. Educ. Res. **14**, 010131 (2018).

[16] M. Sundstrom, A. Schang, A. B. Heim, and N. G. Holmes, Understanding interaction network formation across instructional contexts in remote physics courses, Phys. Rev. Phys. Educ. Res. **18**, 020141 (2022).

[17] L. N. Simpfendoerfer, M. Sundstrom, M. Dew, and N. G. Holmes, What topics of peer interactions correlate with student performance in physics courses?, European Journal of Physics **45**, 035704 (2024).

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019), arXiv:1810.04805 [cs].

[19] J. R. Taylor, *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, 2nd ed. (University Science Books, 1996).

[20] R. K. Fussell, PERC 2024 multi institutional bias (2024).