Generalizability of Sensor-Free Affect Detection Models in a Longitudinal Dataset of Tens of Thousands of Students

Emily Jensen
University of Colorado Boulder
emily.jensen@colorado.edu

Stephen Hutt
University of Colorado Boulder
stephen.hutt@colorado.edu

Sidney K. D'Mello
University of Colorado Boulder
sidney.dmello@colorado.edu

ABSTRACT

Recent work in predictive modeling has called for increased scrutiny of how models generalize between different populations within the training data. Using interaction data from 69,174 students who used an online mathematics platform over an entire school year, we trained a sensor-free affect detection model and studied its generalizability to clusters of students based on typical platform use and demographic features. We show that models trained on one group perform similarly well when tested on the other groups, although there was a small advantage obtained by training individual subpopulation models compared to a general (all-population) model. Lastly, we perform a series of simulations to show how generalizability is affected by sample size. These results agree with our initial analysis that individual subpopulation models yield a small advantage over all-population models. Additionally, we show that training sizes smaller than 1,500 yield unstable models which make generalizability difficult to interpret. We discuss applications of this work in the context of developing large-scale affect detection models for diverse populations.

Keywords

Affect Detection, Clustering, Generalizability, Sensor-free, Online Learning

1. INTRODUCTION

Computer-enabled classrooms and online learning environments are becoming increasingly common methods of learning [12, 25]. Compared to traditional classroom settings, students must be more self-regulated when interacting with online platforms [15]. In [20], Pekrun discusses how emotion and its regulation are key factors in educational achievement. It is then important to consider student affect when developing intelligent tutors and educational platforms. A review of affect-sensitive instructional strategies, particularly for intelligent tutors [5], discusses how affect- and motivation-sensitive strategies can promote student engagement. However, the authors found that a "one-size-fits-all approach, where variants of the same strategy are indiscriminately used for all learners and in all situations" limits the overall effectiveness of these tutors in targeting individual student needs. This observation motivates a more detailed analysis of how affect detectors trained on a general population generalize across different subpopulations.

In this work, we extend previous research exploring the generalizability of sensor-free affect detectors. We trained models predicting positive and negative affective states using interaction data from an online algebra learning platform along with self-reported affect. A novel component of our work compared to previous work (e.g., [2]) is the scope of our dataset, which encompasses 69,714 students across a nine-month period. We extend the previous work in [9]. This enables a more detailed exploration of generalizability than was previously achievable.

1.1 Related Work

Reviews of issues and methods of sensor-free affect detection are covered in other work, which we summarize here. Baker and Ocumpaugh review work in sensor-free affect detection in educational software and discuss methods for collecting ground-truth labels [3]. Specific to this work, they note that student-generated responses are likely more accurate than labels from external coders. Second, [9] reviews a representative collection of sensor-free affect detection models developed in authentic classroom environments. The authors conclude that the studies show the potential success for sensor-free affect detection models in authentic environments but are limited by small sample sizes (20-646 students) from mostly homogenous samples, which limits claims or tests of generalizability.

Recent work in machine learning and prediction calls for increased awareness of how models perform for individual subpopulations in addition to overall accuracy. In [10], Kusner et al. introduce counterfactual fairness, where models should be unaware of protected attributes such as gender and race. Fair models should generalize by generating similar predictions for individuals with similar features, regardless of their protected attributes. In [26], Sculley et al. suggest slicing analysis as a method to evaluate fairness, where predictive model performance is evaluated by "slicing" along subpopulations or protected attributes. This is an alternative to measuring overall model accuracy, which can ignore disadvantaged subpopulations. In response, Gardner et al. present a framework for using slicing analysis in predictive modeling [7].

Related to this discussion on generalizability, several studies have measured how models generalize across cultural contexts. Ogan et al. [18] found differences in collaboration, engagement, and student needs between cultural groups. In [24], San Pedro et al. trained models detecting student carelessness in Philippines. They showed generalizability by testing these models on previously collected data from students in the USA. In [27], Soriano et al. compared models of help-seeking behavior. By training models on each group and testing on the other groups, they showed that models for Philippines and USA generalize to each other but not to Costa Rica.

Besides cross-country generalization, several studies investigated how predictive models generalize over demographic attributes. In [17], Ocumpaugh et al. trained affect detection models on rural, suburban, and urban students. By training models on each group and testing on the other groups, they found that models for urban and suburban students generalized to each other but not to rural students. In [23], Samei et al. trained models of teacher question

asking behavior using data from urban and non-urban classrooms. They showed generalizability using the methods from [17].

Other studies measured the generalizability of predictive models over time. In [1], Baker et al. trained models detecting gaming the system behavior in a cognitive tutor. They showed generalizability by training models on data from three sessions and testing on the remaining session. In [4], Bosch et al. trained face-based affect detection models. They showed generalizability by training models on data from one day and testing on the other day.

Finally, some studies measured generalizability between different tasks or subjects. In [28], Stewart et al. compared models of mind wandering trained on students reading a scientific text or watching a narrative film. They found models trained on the narrative film dataset generalized to the scientific text dataset, but models trained on the scientific text dataset only generalized to the narrative film dataset after adjusting the predicted mind wandering rate. In [9], Hutt et al. found that models trained on data from students enrolled in Algebra 1 generalized to students enrolled in Geometry using "generic activity features" specifically designed for generalization.

1.2 Contribution of Current Study

This work contributes to the field of generalizability in sensor-free affect detection in three important ways. First, we extend beyond previous work by using data from a large, heterogeneous sample of students. Besides the noted studies that compare country-wide cultural differences, previous work relies on homogeneous samples such as individual schools, which yield sample sizes of hundreds of students. As discussed in [2], these sample sizes do not allow researchers to draw conclusions about the studied categories as a whole, so generalizability can only be tested in a minimal sense. In this study, we collected affect data from 69,174 students at 1,898 schools in the state of Florida. Because Florida closely represents the demographic composition of the United States in terms of race and ethnicity [29, 30], this allows us to study the generalizability of our models to other students in the country.

Second, we measure the generalizability of our models in terms of usage characteristics over an entire school year. In previous studies, data are collected during one or a few sessions, which overlooks long-term student behavior. This work uses interaction logs from an entire school year and measures student use over several sessions. We use clustering analysis to identify common usage patterns and show that our models generalize across these clusters.

Lastly, we provide simulation experiments to inform the number of instances needed in order to construct generalizable models. Specifically, we estimate the advantage obtained by training models on individual groups across different sample sizes.

2. DATA

We used a previously published dataset [9] but all analyses reported here are new.

2.1 Algebra Nation

Data was collected through Algebra Nation, an online math learning platform developed by Study Edge. Algebra Nation supports over 150,000 students studying Algebra 1, Algebra 2, and Geometry each semester. Students can use Algebra Nation in a variety of contexts; some teachers integrate the platform into their regular classroom time while some students only use it to study or help with homework. Students can access Algebra Nation using a mobile app or on the internet (https://www.algebranation.com/). For this study, we used data from students enrolled in Algebra 1.

In Algebra Nation, course material is organized according to state mathematics standards. Although the topics are ordered according to the curriculum, students are free to skip topics as necessary or learn the material in a different order.

For each topic, students can watch a video lecture from one of several tutors. In addition to watching videos, students can use the *Test Yourself* quiz feature for each topic, which randomly selects 10 questions aligned with state standards. After attempting a quiz, students can review feedback on their answers or watch solution videos. Lastly, students can get more help through the *Discussion Wall* where they can interact with other students and study experts hired by Algebra Nation. Students can earn *karma points* by answering questions posted by other students. However, students primarily spend time watching videos and taking quizzes rather than engaging in the social functions of the platform.

2.2 Affect Surveys

Due to the large number of students in the study and because students can use the platform in multiple contexts, we collected ground-truth affect labels using a self-report survey rather than through expert coders or human observers (see [16, 22]). These surveys were pseudo-randomly triggered based on student activity on the platform. Specifically, we manually assigned probabilities to each action so that triggered surveys were not overly intrusive and there was an adequate sampling of infrequent actions (e.g., wall posts) compared to highly frequent ones (e.g., seeking in videos).

The survey was displayed in a pop-up window. Students had the option to ignore surveys. To decrease the prevalence of the surveys, once a survey was triggered for a student, the student was removed from the survey pool for two weeks. Our dataset includes surveys from the 2017-2018 school year (September through May). In this time, 69,174 students responded to at least one survey. The mean number of survey responses per student was 1.94 (median = 1). Of the students that responded, the minimum number of responses was 1 and the maximum number of responses by any student was 14.

Each survey targeted one affective state, randomly selected, from the following: Anxiety, Boredom, Confusion, Contentment, Curiosity, Disappointment, Engagement, Frustration, Happiness, Hopefulness, Interest, Pride, Relief, Sadness, Surprise, Mind Wandering, Pleasantness, and Wakefulness. We chose several states because they closely relate to learning [21] while others address core dimensions of affect such as valence and arousal [11]. Mind Wandering, Pleasantness, and Wakefulness represent bipolar concepts, so we used a seven-point scale with contrasting options and presented prompts for each polarity (e.g., sleepy/awake). The other states used a five-point scale ranging from "Not at all __" to "Very __". In our analysis, we linearly scaled all survey responses to lie in a five-point range so that all states are represented equally.

2.3 Generic Activity Features

We recorded student activity on Algebra Nation using 22 features that did not depend on specific content (e.g. which video was watched or a particular quiz question). These activity features included attempting quizzes, watching videos, and interacting with the wall or discussion board. Based on our prior work [9], we counted the number of occurrences of each feature over 30-second chunks and summed the counts for each action across 5-minute window lengths preceding an affect survey. In some cases, the platform measured an unnaturally high amount of activity (e.g. playing/pausing a video 100 times within 30 seconds). We addressed these outliers by limiting each 30-second chunk to 10 recorded activities.

2.4 Usage Features

In addition to session-specific generic activity features, which were used to train the models, we were interested in investigating

generalizability to differences in how students interact with the platform over an entire school year. To do this, we defined five usage features. First, we calculated the proportion of sessions students use their mobile device compared to a desktop computer as this may indicate the context in which students are using the platform (e.g., at home or while commuting). Second, we calculated the proportion of sessions in the spring semester compared to the fall semester. We were interested in this feature because students must pass the algebra standardized exam that is offered in the spring semester in order to graduate from high school. To model how much students use the platform, we calculated the number of sessions and the average length of each session. Students may leave an active session open for long session times in their browser while switching to another task or repeatedly log into the platform without recording any meaningful interactions. We replaced these outliers with the 99th percentile value for the average length and number of logins. Finally, we calculated the mean time of day that students use the platform, which can indicate whether students primarily use the platform during the school day or at home. Usage data was available for 235,756 individual students, including those who did not receive or respond to any surveys.

2.5 Demographics

We also obtained records of demographic data of 118,177 students from the Florida Department of Education. This dataset includes students from grade 6 through grade 12, from which we defined three groups. We defined the first group as Middle School (54%), which includes grades 6 through 8. These students are often advanced and are enrolled in the Algebra course earlier than is typically expected by state standards [6]. The second group is Grade 9 (37%). We chose to keep this grade separate because it has one of the largest enrollment numbers and grade 9 is when students are enrolled in the course during the typical mathematics sequence. Lastly, we defined High School (9%) as grades 10 through 12. These students are often behind in the typical mathematics sequence and struggle to pass the course before they graduate. For gender, the available data classifies students as Male (49%) or Female (51%), which we took at face value.

This dataset records student eligibility for free or reduced-price (F/R) school lunch, which is one indicator of socioeconomic status (but see Harwell & LeBeau [8]). We defined the groups as F/R (53%) and Other (47%), with the latter reflecting those who did not qualify or did not apply. We combined free and reduced because there were so few students that qualified for a reduced-price lunch.

Finally, this dataset includes data on race and ethnicity. We defined these groups to approximately balance group size: White (72%), Black (23%), Hispanic (32%), and Other (13%; Asian, Native American, Pacific Islander, and Mixed).

3. CLUSTERING

We clustered participants based on usage characteristics and demographics to investigate the generalizability of the affect models across clusters. To determine the number of clusters, we inspected the dendrogram generated with Ward hierarchical clustering [31] using the SciPy library (http://www.scipy.org/). For efficient clustering, we randomly sampled 1,000 instances. We then used the k-means algorithm [14] to construct the clusters using scikit-learn [19]. We chose to use all available students regardless of their participation in the surveys since our goal was to generalize over as many students as possible.

We constructed usage clusters using the five features described in Section 2.4. We first scaled each of these features to [0, 1]. The above procedure yielded five clusters (Table 1). One group (U1) showed heavy usage patterns (signified by long sessions and numerous log-ins). Two groups were defined by primarily mobile sessions and were further differentiated by sessions focused in either the fall (U4) or spring (U5) semester. Finally, two groups showed particularly light usage patterns and were differentiated by sessions focused in either the fall (U3) or spring (U2) semester.

Next, we constructed clusters using the demographic features described in Section 2.5. We dummy encoded our variables resulting in seven features indicating grade level, three features indicating lunch status, seven features indicating race/ethnicity, and one feature indicating gender. The above procedures yielded seven clusters (Table 2). Grade level largely differentiated clusters. Only

Table 1. K-means cluster centers based on typical usage. Distinguishing features are bolded.

ID	Cluster Description	Session Time (min)	Num. Sessions	Prop. Spring Use	Prop. Desktop Use	Time of Day (hour)	Prop. of Users
U1	Spring semester, heavy use	**45.46**	**25.44**	0.75	0.90	14.19	0.20
U2	Spring semester, light use	**14.26**	**3.53**	0.96	0.99	14.86	0.35
U3	Fall semester, light use	**13.81**	**3.46**	**0.11**	0.99	14.81	0.28
U4	Fall semester, mobile use	21.38	9.54	**0.19**	**0.30**	13.25	0.07
U5	Spring semester, mobile use	29.66	10.95	0.93	**0.27**	13.21	0.10

Table 2. Demographic cluster centers. For clarity, only distinguishing features are displayed and are bolded.

ID	Cluster Description	Grade 7	Grade 8	Grade 9	Grade 10	F/R Lunch	White	Black	Asian	Prop. of Users
D1	Split grades, F/R lunch, Black	**0.08**	**0.29**	**0.44**	**0.17**	**0.99**	0.03	**1.00**	0.01	0.16
D2	Grade 7, not F/R lunch, White/Asian	**1.00**	0.00	0.00	0.00	0.20	**0.80**	0.06	**0.16**	0.10
D3	Grade 8, not F/R lunch, White	0.00	**0.99**	0.00	0.00	0.00	0.87	0.09	0.07	0.22
D4	Grade 8, F/R lunch, White	0.00	**1.00**	0.00	0.00	**1.00**	0.90	0.03	0.06	0.15
D5	Grade 9, F/R lunch, White	0.11	0.00	**0.87**	0.00	**1.00**	0.91	0.01	0.03	0.16
D6	Grade 9, not F/R lunch, White/Black	0.00	0.00	**0.99**	0.00	0.04	**0.81**	**0.16**	0.04	0.16
D7	Grade 10, split F/R lunch, White/Black	0.00	0.00	0.00	**1.00**	**0.52**	**0.76**	**0.20**	0.03	0.05

cluster D1 had a significant distribution of students across grade levels. Another differentiating feature was lunch status, where three clusters (D1, D4, D5) were largely comprised of students on F/R lunch. Four clusters were differentiated by race (D1, D2, D6, D7).

4. AFFECT DETECTION MODELS

4.1 Model-building Procedure

We used scikit-learn [19] to implement a supervised learning pipeline. We chose to use the Bayesian Ridge Regression algorithm [13] since it produced good overall results in previous work on the same data [9] compared to several more complicated alternatives.

We trained regression models using 10-fold student-level cross validation. For each fold, instances for each student were included in *either* the training or testing set. This practice reduces overfitting and increases the likelihood that the model will generalize to new students. In each fold, we trained a model using the generic activity features and generated predicted survey responses on the test data. We evaluated the performance of the model using the Spearman correlation as it assumes ordinal and continuous values. We then averaged these scores across folds to obtain a final accuracy score.

We trained prediction models for positive and negative affective valence rather than the original 18 states measured in the surveys. We initially trained a model for each state and calculated the correlation between the predicted survey responses for each state. These predictions were strongly correlated within positive and negative valence. We then trained positive and negative valence models using the combined set of states and generated predictions for the individual states. The mean performance of the valence models was similar to training individual affective models, so we chose to use the valence models for parsimony. For the positive valence models, we included the following states: Arousal, Contentment, Engagement, Happiness, Hopefulness, Interest, Pleasantness, Pride, and Relief. For the negative valence models, we included the following states: Anxiety, Boredom, Confusion, Disappointment, Frustration, Mind Wandering, and Sadness. We did not include Curiosity and Surprise since their valence does not clearly align on either direction.

4.2 Preliminary Models on Cluster Membership

We first investigated whether our models discriminated using group features rather than the generic activity features. To test this, we trained models using cluster membership as the training data instead of activity features. We expected these models to perform poorly since they are not simply reflecting group differences. Indeed, we found that the average Spearman correlations were low (between 0.02 and 0.05) for both cluster models.

4.3 Generalizability

Our main analysis focused on investigating how our models, trained on activity features, generalize across different clusters. First, we considered a general model trained on the entire dataset using 10-fold student-level cross validation. We then built cluster-specific models. For each cluster, we trained and tested a model on data from that cluster. We also tested this model on the other cluster data. For example, we trained a model on U1 and tested on each of the other clusters (U2 – U5) as well as the entire dataset (All). We performed this procedure separately for the positive and negative valence states as well as for the usage and demographic clusters[1].

4.4 Results

We examined the generalizability of our models using the procedure in Section 4.3. If our models generalized well, we expect to see a model trained on one group perform similarly well when applied to other groups (Table 3). This was the case for the usage clusters, where the maximum difference between testing on one cluster and testing on another is 0.05. The demographic clusters were more varied. In this case, the greatest difference between testing on the one cluster and testing on another was 0.09.

Recent metrics proposed in slicing analysis, such as [7], apply to classification problems and not the regression task considered here. To better quantify model generalizability, we defined an *individual advantage* metric. Using the procedure from Section 4.1, we trained a model using the training set X and tested the model using the testing set Y. We represented the performance of the model, which is the average Spearman correlation, as $P_{X,Y}$. For a target group T, we defined the individual advantage metric as $(P_{T,T} - P_{All,T})/P_{All,T}$. This describes the proportion improvement over using a general model for the target group T. Therefore, a perfectly generalizable model would have an individual advantage of 0 since an individual model and general group model will have the same accuracy.

We used this metric to quantify the generalizability of our models. Both positive and negative models showed small, positive individual advantage values (mean usage 0.04; mean demographics 0.02), which indicates a small advantage to training cluster-specific models compared to a general model.

5. SAMPLE SIZE SIMULATIONS

We then investigated whether sample size affects the advantage for using individual models. Specifically, are individual advantages mitigated when more data is available? To address this question, we computed the average individual advantage metric over 10 cross validation folds for a range of sample sizes starting at 500. For each sample size, we randomly selected the appropriate number of instances from the training sets. We incrementally increased the sample size by 200 until we reached the actual group size, which varied between 1,500 and 7,100. We repeated this simulation 1,000 times and calculated the 95% confidence interval of the mean individual advantage metric at each sample size (Figure 1).

Table 3. Mean correlation of positive valence models (negative in parentheses) for usage clusters.

	Test U1	Test U2	Test U3	Test U4	Test U5	Test All
Train U1	**0.25**	0.23	0.21	0.19	0.21	0.21
	(0.19)	(0.22)	(0.19)	(0.19)	(0.18)	(0.20)
Train U2	0.21	**0.26**	0.22	0.21	0.22	0.22
	(0.18)	**(0.23)**	(0.19)	(0.20)	(0.19)	(0.20)
Train U3	0.21	0.24	**0.23**	0.20	0.20	0.22
	(0.17)	(0.21)	**(0.20)**	(0.19)	(0.17)	(0.19)
Train U4	0.20	0.25	0.21	**0.21**	0.21	0.22
	(0.17)	(0.22)	(0.19)	**(0.20)**	(0.19)	(0.21)
Train U5	0.23	0.25	0.21	0.20	**0.24**	0.22
	(0.17)	(0.22)	(0.18)	(0.19)	**(0.21)**	(0.20)
Train All	0.22	0.25	0.22	0.21	0.22	**0.22**
	(0.18)	(0.22)	(0.20)	(0.20)	(0.19)	**(0.21)**

[1] Similar results for other slices can be found using this code (link).

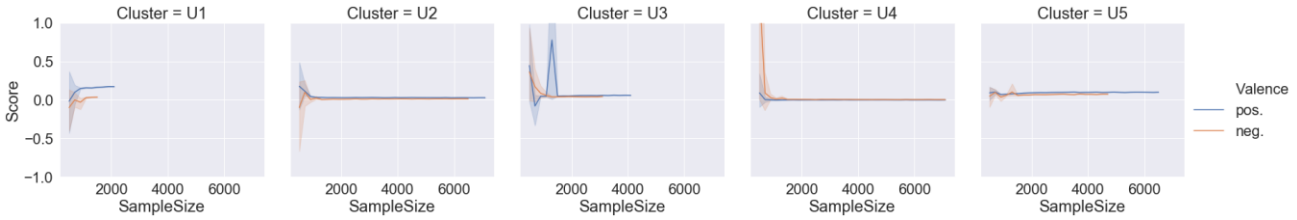

Figure 1. Averaged individual advantage simulation scores for usage clusters

We first noted that the scores for models using sample sizes less than 1,500 varied wildly, as indicated by the width of the confidence intervals in this region. As such, we can only conclude that results obtained from small samples might not be reliable. This is concerning since previous work used sample sizes ranging from 20 to 646 students [9]. As expected, for larger sample sizes, the models quickly stabilized and produced more reliable scores.

For most clusters, the individual advantage scores stabilized to a value of 0.10 or less, which indicates a small advantage of training cluster-specific models. The scores for clusters D1 and U1 seemed to increase as the sample size increased, but we cannot make strong conclusions since the sample size of these clusters was small.

6. DISCUSSION

In an attempt to answer the call for predictive generalizability [7, 10, 26], we used interaction data from 69,174 students over an entire school year to study the extent to which sensor-free affect detectors generalize across usage and demographic clusters.

6.1 Main Findings

We found that students primarily differed in their interaction rate, active semester, and primary device. The demographic clusters were primarily discriminated by grade level, F/R lunch eligibility, and (to a smaller extent) race. Using cluster membership as the only training feature resulted in near-zero results, which shows students in a particular cluster are not generally predisposed to certain affective states. We must then consider the context of a student's activity when predicting their immediate affective state.

Similar to previous work [1, 4, 9, 17, 23, 24, 27, 28], we examined the generalizability of our models by training cluster-specific models and testing them on the other clusters. We found that cluster-specific models perform slightly better on the target cluster, with a maximum difference of 0.05 for the usage clusters and 0.09 for the demographic clusters. We expanded this analysis by introducing an individual advantage metric, which measures the advantage given to a target group compared to a general (entire population) model. This metric agreed with our initial analysis by showing a small advantage given by training a cluster-specific model. The maximum advantage was 0.14 for the usage clusters and 0.11 for the demographic clusters. Although these results provide evidence that cluster-specific models are better at predicting affective valence, it is not clear what difference is meaningful in practice.

Lastly, we investigated how model generalizability changes in response to sample size. We performed a series of simulations that trained affect-detection models and systematically varied sample sizes. Models trained on 1,500 samples or less did not generate stable scores or predictions, even after 1,000 iterations. When considering sample sizes greater than 1,500 that yielded reliable scores, we found that the individual advantage scores stabilize as sample size increases at a value of 0.10 or less, which is consistent with our initial analysis. This suggests that generalizability is not greatly affected by sample size beyond the 1,500-sample threshold.

6.2 Limitations and Future Work

The greatest area of improvement is the overall model performance. As discussed in [9], the average performance corresponds to a small-sized effect. This is likely caused by the limited number and extreme generality of the training features. Future work can address this by introducing more platform-specific features, such as which quiz a student was attempting. We can then see if our models have the power to distinguish between individual affective states rather than simply identifying positive or negative valence. Of course, the use of these features will result in more platform-specific models, which limits their generalizability to different platforms or even to other domains within the sample platform.

Our analysis of generalizability was limited to demographic features and overall interaction with the Algebra Nation platform. This analysis should be extended to include other academic subjects, time frames, and regional groups. For example, while Florida does reflect the overall demographic composition of the United States, other states do not. It would be interesting to see how our models generalize to other populations. With respect to subject generalizability, while [9] showed generalizability between Algebra and Geometry, we could see how a model for mathematics generalizes to unrelated subjects such as chemistry or music.

There are several exciting opportunities to apply these large-scale sensor-free affect detectors. First, we will be able to develop real-time interventions based on predictions of a student's affective state and promote more a more engaging experience with the curriculum. In addition, as we collect data from different regions and over longer time periods, we can more directly investigate the relationship between engagement and end-of-course scores.

Lastly, it is important to understand the impacts of a one-size-fits-all model on long-term student achievement. When developing interventions, one should consider possible effects if predictions of affect are incorrect. In this case, the intervention should not have any negative consequences for the student receiving it.

7. CONCLUSION

Sensor-free affect detection models provide the opportunity to provide personalized experience for large populations of students. In this work, we answered the call to investigate how these predictive models generalize between different subpopulations in the training data. We did this using a longitudinal dataset of student interaction with an online math learning platform with our groups of interest being clusters based on typical usage on the platform and demographic features. We showed that while models trained on one cluster perform similarly well when applied to the other clusters, there is a small advantage to use individual subpopulation models rather than one general population model. It is important to consider these models' differential performance and impact when deploying large-scale platforms that adapt to sensor-free predictions of individual students' affective states.

8. ACKNOWLEDGMENTS

This work was supported by the Institute of Education Sciences, U.S. Department of Education, through grant R305C160004 and Intel Research. The opinions expressed are those of the authors and do not represent views of the Institute, the U.S. Department of Education, or Intel.

REFERENCES

[1] Baker, R.S.J. d. et al. 2008. Developing a generalizable detector of when students game the system. *User Modeling and User-Adapted Interaction*. 18, 3 (Aug. 2008), 287–314.

[2] Baker, R.S.J. d. and Gowda, S.M. 2010. An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools. *Proceedings of the 3rd International Conference on Educational Data Mining* (Pittsburgh, Pennsylvania, 2010), 11–20.

[3] Baker, R.S.J. d. and Ocumpaugh, J. 2014. Interaction-based affect detection in educational software. *The Oxford Handbook of Affective Computing*. R.A. Calvo et al., eds. Oxford University Press.

[4] Bosch, N. et al. 2015. Temporal generalizability of face-based affect detection in noisy classroom environments. *Artificial Intelligence in Education* (2015), 44–53.

[5] D'Mello, S. et al. 2014. I feel your pain: a selective review of affect-sensitive instructional strategies. *Design Recommendations for Intelligent Tutoring Systems - Volume 2 Instructional Management*. R.A. Sottilare et al., eds. U.S. Army Research Laboratory. 35–48.

[6] Education, F.D. of 2014. Mathematics florida standards (mafs).

[7] Gardner, J. et al. 2019. Evaluating the fairness of predictive student models through slicing analysis. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (2019), 225–234..

[8] Harwell, M. and Lebeau, B. 2010. Student eligibility for a free lunch as an ses measure in education research. *Educational Researcher*. 39, 2 (2010), 120–131.

[9] Hutt, S. et al. 2019. Time to scale : generalizable affect detection for tens of thousands of students across an entire school year. *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)* (2019).

[10] Kusner, M. et al. 2017. Counterfactual fairness. *31st Conference on Neural Information Processing Systems (NIPS 2017)* (2017), 4066–4076.

[11] Linnenbrink, E.A. 2007. The role of affect in student learning: a multi-dimensional approach to considering the interaction of affect, motivation, and engagement. *Emotion in Education*. P.A. Schutz and R. Pekrun, eds. Elsevier Inc. 107–124.

[12] Liyanagunawardena, T.R. et al. 2013. MOOCs : a systematic study of the published literature 2008-2012. *International Review of Research in Open and Distance Learning*. (2013).

[13] MacKay, D.J.C. 1992. Bayesian interpolation. *Neural Computation*. 4, 3 (May 1992), 415–447.

[14] MacQueen, J. 1967. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Berkeley, California, 1967), 281–297.

[15] Moos, D.C. 2014. Setting the stage for the metacognition during hypermedia learning: what motivation constructs matter? *Computers and Education*. 70, (2014), 128–137.

[16] Nicaud, J.F. et al. 2006. Experiments with aplusix in four countries. *International Journal for Technology in Mathematics Education*. 13, 2 (2006), 79–88.

[17] Ocumpaugh, J. et al. 2014. Population validity for educational data mining models: a case study in affect detection. *British Journal of Educational Technology*. 45, 3 (May 2014), 487–501.

[18] Ogan, A. et al. 2012. Collaboration in cognitive tutor use in latin america: field study and design recommendations. *2012 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2012)* (New York, New York, USA, 2012), 1381–1390.

[19] Pedregosa, F. et al. 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*. 12, (2011), 2825–2830.

[20] Pekrun, R. 2017. Emotion and achievement during adolescence. *Child Development Perspectives*. 11, 3 (2017), 215–221.

[21] Pekrun, R. 2007. Emotions in students' scholastic development. *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*. R.P. Perry and J.C. Smart, eds. Springer. 553–610.

[22] Porayska-Pomsta, K. et al. 2013. Knowledge elicitation methods for affect modelling in education. *International Journal of Artificial Intelligence in Education*. 22, 3 (2013), 107–140.

[23] Samei, B. et al. 2015. Modeling classroom discourse: do models that predict dialogic instruction properties generalize across populations? *Proceedings of the 8th International Conference on Educational Data Mining* (2015), 444–447.

[24] San Pedro, M.O.C.Z. et al. 2011. Detecting carelessness through contextual estimation of slip probabilities among students using an intelligent tutor for mathematics. *Artificial Intelligence in Education* (2011), 304–311.

[25] Sandeen, C. 2013. Integrating moocs into traditional higher education: the emerging "mooc 3.0" era. *Change: The Magazine of Higher Learning*. 45, 6 (2013), 34–39.

[26] Sculley, D. et al. 2018. Winner's curse? on pace, progress, and empirical rigor. *6th International Conference on Learning Representations* (2018).

[27] Soriano, J.C.A. et al. 2012. A cross-cultural comparison of effective help-seeking behavior among students using an its for math. *Intelligent Tutoring Systems* (2012), 636–637.

[28] Stewart, A. et al. 2017. Generalizability of face-based mind wandering detection across task contexts. *Proceedings of the 10th International Conference on Educational Data Mining* (2017), 88–95.

[29] U.S. Census Bureau 2018. *Florida QuickFacts*.

[30] U.S. Census Bureau 2018. *United States QuickFacts*.

[31] Ward, J.H. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*. 58, 301 (Mar. 1963), 236–244.