Identifying supportive contexts for mindset interventions: A two-model machine learning approach

Nigel Bosch
School of Information Sciences and Department of Educational Psychology
University of Illinois at Urbana-Champaign
pnb@illinois.edu

Abstract
Growth mindset interventions (which foster students’ beliefs that their abilities can grow through effort) are more effective in some contexts than others; however, relatively few variables have been explored that could identify contexts in which growth mindset interventions are most effective. In this study, we utilized machine learning methods to predict growth mindset effectiveness in a nationwide experiment in the U.S. with over 10,000 students. These methods enable analysis of arbitrarily-complex interactions between combinations of student-level predictor variables and intervention outcome, defined as the improvement in grade point average (GPA) during the transition from high school. We utilized two separate machine learning models: one to control for complex relationships between 51 student-level predictors and GPA, and one to predict the change in GPA due to the intervention. We analyzed the trained models to discover which features influenced model predictions most, finding that prior academic achievement, intervention compliance, self-reported reasons for learning, and race/ethnicity were the most important predictors in the model for predicting intervention effectiveness. Unique to this study, we found that low intervention compliance (attempting to navigate through the intervention software without completing all steps) resulted in as much as -0.2 difference in predicted intervention effect on GPA. Our findings have implications for the design of computer-administered growth mindset interventions, especially for students who do not properly complete the intervention.

1 Introduction
Students approach learning with differing beliefs about their own abilities to learn and grow (Dweck, 2006), beliefs about specific topics (Chestnut, Lei, Leslie, & Cimpian, 2018; Leslie, Cimpian, Meyer, & Freeland, 2015), and differing reasons for learning (Yeager et al., 2014). These beliefs about learning, or learning mindsets, are related to learning outcomes. Given this connection, recent research has explored the possibility of fostering certain learning mindsets to improve learning outcomes. The study in this paper focuses on growth mindset, which is the belief that one’s ability can grow with appropriate effort and strategies (Dweck, 2006). Conversely, a fixed mindset reflects the belief that abilities and intelligence are more innate, and thus any impasses encountered while learning difficult concepts may be indicators of one’s own lack of innate ability. In this paper, we research possible contexts in which a computer-
administered intervention, designed to foster a growth mindset in students, may or may not be effective. We utilize machine learning methods to examine data from the National Study of Learning Mindsets (NSLM), a large-scale computer-administered intervention experiment (Yeager et al., 2019).

Computers are an attractive means of administering interventions directly to students at a wide scale, given their ubiquitous nature and flexibility. For example, one previous computer-based intervention study with over 3,500 students in 10 schools in the United States found that a growth mindset intervention improved 9th-grade grade point average (GPA) for lower-achieving students (Yeager et al., 2016). However, growth mindset interventions do not always produce positive results for all students (Sisk, Burgoyne, Sun, Butler, & Macnamara, 2018). It remains an important open question why such interventions are more effective for some students than others, and under what conditions interventions are likely to work.

Successful computer-based growth mindset interventions have been demonstrated in various educational settings (O’Rourke, Haimovitz, Ballweber, Dweck, & Popović, 2014; Sisk et al., 2018; Yeager et al., 2019, 2016). In one experiment, researchers modified a mathematics education game called Refraction, adding messaging designed to foster a growth mindset for students in the experimental condition, and adding neutral messaging about the importance of math in the control condition (O’Rourke et al., 2014). They found that the messaging improved students’ persistence (time spent) in the game. Similarly, researchers studying the growth mindset attitudes of students on a mobile computing learning platform found that growth mindset predicted higher quiz scores as well as longer time spent answering quizzes (Kizilcec & Goldfarb, 2019) – though not in an experimental intervention context. These findings are consistent with the theory that growth mindsets help students persist in the face of adversity (Dweck, 2006). Furthermore, previous work has shown that students who struggle more are those who are most likely to benefit from persistence and challenge-seeking behavior engendered by a growth mindset. For example, in the Refraction experiment, students who struggled most in the first few math problems later benefitted the most from the mindset intervention (O’Rourke et al., 2014). In another study comparing achievement across socioeconomic status (SES) levels, researchers found that growth mindset was correlated with SES and SES was correlated with scholastic achievement (Claro, Paunesku, & Dweck, 2016). However, the SES–achievement gap was mitigated for students from low-SES backgrounds when they had a growth mindset.

A recent meta-analysis of growth mindset interventions in education found that overall effects on academic outcomes are weak (Sisk et al., 2018). The overall effect size for interventions was $d = 0.080 \; (p = .010)$; however, they also explored possible moderators of intervention effect. Student-level moderators included developmental stage (age), previous failed classes, experience of situational challenges (e.g., stereotype threat, moving to a new school), and SES (eligibility for free or reduced-price school lunch). Procedural moderators included type of control condition (e.g., active, do-nothing), type of intervention (e.g., in class, computer administered), and timing of outcome measure (time to measurement of academic progress). They found that students with previous failed classes and students experiencing economic disadvantage benefitted significantly more than their peers.

In the NSLM experiment (Yeager et al., 2019), researchers utilized a computer-administered intervention to encourage adoption of a growth mindset for students in their first year of high
school, a time when students are typically transitioning to a new school and may be encountering new challenges (Sisk et al., 2018). They found that students in lower-performing schools benefitted more from the intervention (compared to an active control group), in line with previous research. However, there are many possible moderators – and combinations of moderators – that could potentially relate to intervention effects. Hence, in this paper we explore a large number of possible predictors (51) gathered during the NSLM experiment, utilizing machine learning methods to predict intervention efficacy. The machine learning methods we employ here have notable advantages versus simpler statistical analyses, especially for large datasets like the NSLM dataset. Of particular importance are the ability to handle arbitrarily-complex interactions between predictors, adjustable regularization methods to help prevent over-fitting of models, and suitable cross-validation strategies to measure accuracy appropriately even when over-fitting may have occurred.

We had two objectives in training machine learning models. First, we sought to enable future improvements to the intervention by predicting when it would be less effective for improving post-intervention GPA. Second, recent advances in machine learning methods allow inspection and interpretation of complicated models, thereby discovering which variables were most predictive. This research is particularly relevant to future computer-based mindset interventions in schools. We explored detailed predictors of intervention efficacy that have never before been studied in such a large-scale randomized controlled trial (over 10,000 students from across the United States) and allow arbitrarily-complex interactions between predictors (via machine learning methods). This is the first study to include procedural predictors such as self-reported distraction during the intervention administration, compliance with intervention instructors, and others, alongside psychological measures of individual student differences. We utilized the NSLM dataset to explore two research questions, which confirmed previous findings in the dataset and uncovered new findings with relevance to practical application of growth mindset interventions. Our research questions were:

RQ1) How do student-level variables predict students’ future GPA in the control condition? By answering this question, we can estimate future GPA in absence of a mindset intervention, and thus control for this change for students in the intervention condition.

RQ2) How do student-level variables predict intervention efficacy (change in GPA due to intervention)? In this analysis we explore whether the intervention works equally well for all students, and, if not, which student characteristics predict how well the intervention will work.

2 Material and methods

This study is a secondary analysis of the dataset from the National Study of Learning Mindsets, which was a randomized controlled trial testing a computer-administered growth mindset intervention. Extensive details about the dataset and how it was collected are available in a publication from the initial analysis of the dataset (Yeager et al., 2019); in this section we summarize only key details relevant to the current analyses.

2.1 Participants

Participants were 22,695 students in 9th grade (the first year of high school), at 76 schools across the United States. Schools were chosen so that demographics were nationally representative. Of the 76 schools, 11 did not report administrative data required for the current analysis (e.g.,
student demographics). In the 65 schools that did report administrative data, students were 43% White, 24% Hispanic or Latina/o, 11% Black or African-American, 4% Asian-American, and 18% other racial or ethnic groups. A further 3 of the 65 schools did not provide both pre- and post-experiment student grades for any students. Grades were required to construct the outcome measure in this paper, and these schools were therefore removed. Thus, there were 62 schools in the final sample. Grades were available for most – but not all – students in these schools. There were 10,877 students with complete grade data that we analyzed in this study, consisting of 5,452 in the control condition and 5,425 in the intervention condition. Post-intervention GPA was defined as the mean of math, English, social studies, and science grades from Fall and Spring semesters for students who participated in the experiment near the beginning of the Fall semester, and from only the Spring semester for students who participated near the beginning of the Spring semester.

2.2 Intervention task

Students participated in the intervention in two sessions, separated in time by one to four weeks depending on which school they attended. Both sessions included materials designed to encourage a growth mindset, though different individual difference measures were administered at each session. The intervention consisted of messages and hypothetical scenarios intended to encourage growth mindset. For example, a message delivered to students near the beginning of the intervention read:

High school is a time when the brain can learn and grow more than almost any other time in life. The work you do in high school can actually make your brain stronger, and building a stronger brain in high school helps you in life no matter what you plan to do.

In the control condition, students completed a similar computer-administered task, though without mention of growth mindset. Instead, students learned about how the brain works. An excerpt of material delivered to students near the beginning of the control activity read:

Many people think the brain is a mystery. They don't know much about what it's made of, how it works, or what its different parts do.

The intervention was developed iteratively in previous work via qualitative and quantitative user-centered design (Yeager et al., 2016), with multiple rounds of focus groups, refinement, and testing. Some of the key features of the intervention included elements designed to 1) relate the intervention content specifically to first-year high school students, 2) promote relevance to students from community-oriented families and demographic groups, and 3) reinforce students’ internalization of growth mindset by having them communicate its importance to a hypothetical future student.

During the intervention (and control) sessions, students completed a variety of individual difference measures. These included questions about their expectations for success in math during high school (Hulleman & Harackiewicz, 2009), their reasons for learning (Stephens, Fryberg, Markus, Johnson, & Covarrubias, 2012; Yeager et al., 2014), their level of growth versus fixed mindset (Dweck, 2006), and others. The intervention software asked students procedural questions as well, such as whether they were distracted during the intervention, if
nearby students were working diligently, if there were any technical difficulties, and related questions. The software also recorded information about students’ interaction with the software, including how long they spent on each activity and how many times students’ attempts to navigate within the software were blocked because they had not yet completed the current activity (did not answer all questions or attempted to proceed too quickly without spending time on the material).

2.3 Machine learning procedure
We trained two machine learning models in this study, one for each research question. The primary focus of this study is on RQ2: prediction of the intervention’s effect on GPA from student-level variables. However, simply predicting change in GPA (i.e., post-experiment GPA – pre-experiment GPA) is not a suitable measure of intervention effect, because student GPAs may change in predictable ways even without the intervention. For example, students from low SES families may have a more (or less) difficult time transitioning from middle school to high school than their peers; thus, if we are interested in the relationship between SES and intervention efficacy, we should first subtract the relationship between SES and GPA that occurs with no intervention. Hence, we first trained a model with data from only the control condition (model 1), and inspected this model to answer RQ1. Then, we applied model 1 to the intervention condition data, predicting how much each student’s GPA in the intervention condition would change if they had not been given the intervention. Finally, we subtracted the predicted GPA change from the actual GPA change for each student in the intervention condition, yielding the residual GPA change due to the intervention after controlling for every predictor. These residual GPA changes served as labels (outcomes) to predict in the model for RQ2 (model 2), which was constructed and trained with the same procedure as model 1, detailed below.

2.3.1 Data preprocessing
The intervention software administered a battery of individual difference measures to students in both conditions. However, some measures were only provided to students in one condition. For example, we initially found the Field-specific Ability Beliefs (FAB) measure (Leslie et al., 2015) to be predictive of GPA change in the control condition; however, this measure was not administered in the intervention condition, so we removed it to maintain consistency across conditions.

We also combined multiple variables into one for measures with multiple related questions. For example, we averaged two related questions about students’ reasons for learning: whether they learn because they ultimately want to help others, and whether they learn to serve as a role model for others ($\alpha = .716$). In a measure of challenge-seeking behavior, students were asked to create their own math worksheet from an array of easy, medium, and difficult exercises. We aggregated all of their responses into one variable consisting of the number of difficult minus the number of easy exercises they chose.

Values were missing in some cases, either because students did not respond to a question or because administrative data were unavailable. We handled these situations on a per-variable basis. For categorical variables, we added an additional category for missing values so that the models would be able to learn any important non-response patterns. For continuous and ordinal variables, we replaced missing values with the mean of that variable. In the case of ordinal variables, the means fell between ordinal ranks, thus effectively serving as a new rank.
2.3.2 Model training
We trained gradient boosting tree models via the XGBoost package (Chen & Guestrin, 2016) in R (R Core Team, 2013). Gradient boosting is a particularly powerful, state-of-the-art method for training models on high-level variables (structured data) such as the interpretable predictors in this study. Gradient boosting is a general framework that can be applied to learn different types of models. In this study, we chose to learn decision tree structures, which are well-suited to the current data because they allow for non-linear relationships, complex interactions between any number of predictors, and arbitrary distributions for each variable.

XGBoost has the capacity to fit decision boundaries as complicated as the training data itself, and thus can easily over-fit to training data. Therefore, we utilized leave-one-school-out cross-validation to evaluate prediction accuracy on held-out data, using the caret package in R (Kuhn, 2008), and tuned regularization hyperparameters to minimize over-fitting. We tuned hyperparameters including:

1) Maximum depth allowed for each tree in the model (1, 2, 3, …, 8)
2) Proportion of instances (rows) randomly sampled to train each tree (.5, .6, .7, …, 1)
3) Minimum number of instances required in each leaf of each tree (1, 2, 4, 8, …, 128)
4) Proportion of variables randomly sampled to train each tree (.5, .6, .7, …, 1)
5) Learning rate (.01, .02, .04, …, .32)
6) Minimum loss (error) reduction required to split a tree node (0, .01, .02, .04, …, 10.24)

We tuned hyperparameters with nested cross-validation in the training data only – i.e., we further split training data into train and validation subsets, trained models with various hyperparameter settings, and chose the best combination based on the validation subset. Given that there were 8 × 6 × 8 × 6 × 6 × 12 = 1,658,888 possible hyperparameter combinations to test for each of 62 cross-validation folds (one per school), we could not explore all hyperparameter combinations. Instead, we utilized coordinate descent (Wright, 2015). With this approach, we optimized each hyperparameter setting sequentially, in the order given above, iterating through the list of hyperparameters five times to allow for dependencies between hyperparameters to be partially resolved. We chose the best result based on lowest root mean squared error (RMSE) as measured via nested cross-validation within the training data1. Finally, after selecting these hyperparameters we tuned the number of trees learned in the model, from 1 to 500 (again based on training data only).

2.3.3 Model evaluation
We optimized hyperparameter settings and the models themselves based on RMSE, given that the predicted value (GPA change) was continuous. However, we also evaluated results based on correlation (Pearson’s r) between predictions and outcomes.

We inspected models to determine what they had learned about the relationships between potential predictors and GPA change. In particular, we calculated Shapley values for each feature (predictor) and instance (student). Shapley values measure the unique contribution of each feature toward the prediction of a particular instance (Lundberg & Lee, 2017), given all possible

---

1 The “best” model chosen according to lowest RMSE does not necessarily have statistically lower RMSE than all other models; rather, this is a common heuristic approach to choose a single model from among a set of candidates that may (or may not) be functionally equivalent.
combinations of other features. In the regression models trained in this study, Shapley values can be easily interpreted as the difference in predicted outcome (i.e., difference in predicted GPA change) attributable to a particular feature. Examining Shapley values across all instances thus allows graphical interpretation of non-linear effects after controlling for potential non-linear interactions with all other features.

3 Results

3.1 RQ1: How do student-level variables predict students’ future GPA in the control condition?

Mean GPA change in the control condition was -0.223 ($SD = 0.658$), indicating that, on average, students in the control condition had lower post-experiment GPAs.

3.1.1 Model 1 prediction accuracy

We calculated RMSE and $r$ for the students in each held-out school during cross-validation. Mean RMSE across all schools was 0.561, which compares favorably to the 0.658 standard deviation (which corresponds to a baseline RMSE where the mean is predicted for all instances). Mean correlation between predicted and actual GPA change was $r = .312$ ($p < .001$), indicating that student-level variables significantly predicted changes in GPA in the control condition.

After measuring accuracy, we re-trained model 1 on all control condition data so that we could examine feature importance and apply the model to students in the experimental condition. We calculated aggregate feature importance as the mean of the absolute Shapley values (influence on prediction) across instances, given that we were interested in the size of influence rather than only influence in a specific direction. Given the large sample size, most variables had significantly above-zero influence on predictions; however, most were close to zero. We focus here on the five largest, in Table 1. The next largest was notably lower ($M = .024$).

3.1.2 Model 1 feature analysis

Results in Table 1 show, perhaps unsurprisingly, that past academic performance was the strongest indicator of future academic performance – both in terms of GPA as recorded in administrative records and in terms of student self-reports of their typical grades in core classes (English, math, and others). Similarly, expectations of future success in math classes indeed predicted future success. We graphically explored the relationships between variables and model 1 predictions below, to gather insight into possible directionality of effects. In feature importance figures, variance in the $y$ direction for a particular value on the $x$ axis indicates an interaction with other variables – i.e., for instances with the same value in the $x$ variable the model made different predictions based on interactions with one or more additional features.

Figure 1 shows that the relationship between pre-experiment GPA and predicted GPA was generally negative. Most notably, students with high pre-intervention GPAs (> 3) had little room for improvement; hence, predictions for these students tend to be slightly negative. Figure 2 shows an opposite trend for students’ self-reported typical grades however. Predictions were higher (above zero GPA change) for students who reported getting “Mostly A’s”, while predicted GPA change was negative for students reporting low grades. Similarly, students’ expectations for success in high school math were positively related to predicted GPA improvement (Figure 3).
Analysis of predictions based on SES, as measured by eligibility for free or reduced-price lunch, showed higher predicted GPA improvement for students from high SES households (Figure 4). Most notably, missing/not reported SES status was the most negative indicator of GPA change, though the influence on predictions was smaller for these less-important variables (as is apparent from the range of the y axis in Figure 4). Similarly, gender had relatively small influence on the model; females were predicted slightly more positively than males.

| Variable                                      | Absolute effect on prediction | M    | SD   |
|-----------------------------------------------|-------------------------------|------|------|
| Pre-experiment GPA                            | 0.184                         | 0.163|
| Self-reported typical grades                  | 0.106                         | 0.076|
| Free or reduced-price lunch eligibility (SES) | 0.072                         | 0.041|
| Expectations of success in high school math   | 0.041                         | 0.025|
| Gender                                        | 0.037                         | 0.013|
Figure 1. Shapley values for the effect of pre-experiment GPA on predicted GPA change for students in the control condition.

Figure 2. Shapley feature importance values for self-reported typical grades (a categorical variable) in the control condition.
Figure 3. Feature importance values for students' expectations of success in math.

Figure 4. Feature importance for free/reduced-price lunch eligibility (a proxy for SES).
3.2 **RQ2: How do student-level variables predict intervention efficacy (change in GPA due to intervention)?**

Mean GPA change in the intervention condition was -0.197 (SD = 0.674), indicating that students’ GPAs decreased, on average, as was the case in the control condition (M = -0.223, SD = 0.658). However, the difference between GPA change in control and intervention conditions was significant (p = .041), indicating that the intervention had a positive effect on GPA (for similar results focused on previously low-achieving students in the NSLM dataset, see Yeager et al., 2019).

We applied model 1 to the intervention condition to control for overall expected GPA change and GPA change related to individual differences. Mean predicted GPA in the intervention condition differed from actual GPA in the control condition by just 0.002, indicating that model 1 predicted similar GPAs for the intervention condition. The difference was not significant (p = .873), as expected.

Residual difference between model 1’s predicted GPA and actual GPA in the intervention condition was small (M = 0.025, SD = 0.674), as expected since the difference between conditions was small. However, even a small beneficial effect is valuable, given the brief nature of the intervention. Moreover, variance suggests that there may be contexts in which the intervention had a notably larger or smaller effect. Thus, we trained model 2 to predict these residuals attributable to the intervention.

---

**Figure 5. Feature importance for gender.**
3.2.1 Model 2 prediction accuracy
As in model 1 analysis, we calculated RMSE and \( r \) from held-out predictions during cross-validation for model 2. RMSE was 0.614, which compares favorably to the standard deviation of the labels (i.e., residuals). Mean \( r \) was small (.065) but significantly above chance (\( p = .012 \)). Feature importance values for model 2 were smaller than those for model 1, since the mean of labels was closer to 0 and feature importance values are on the same scale as the labels. Hence, in the next section, we discuss variables with mean absolute Shapley values of 0.01 or higher.

3.2.2 Model 2 feature analysis
Feature importance results in Table 2 indicate that pre-experiment GPA was the most important predictor of the intervention effect (\( M = 0.037, SD = 0.028 \)), despite controlling for overall relationships between past GPA and future GPA effects by applying model 1. This indicates that GPA was important for predicting the intervention effect itself. Also important was the count of blocked navigation events (\( M = 0.019, SD = 0.036 \)), which are events where students attempted to proceed through the intervention software without answering required questions or without spending sufficient time reviewing intervention materials. The reason for learning variable (\( M = 0.012, SD = 0.007 \)) refers to an individual difference measure of extrinsic purpose for learning (see Section 2.3.1), while race/ethnicity (\( M = 0.012, SD = 0.005 \)) was obtained from administrative data.

We explored non-linear feature importance effects, as for model 1, by comparing the effect each feature had on model 2 predictions across all students. Figure 6 shows the relationship between pre-experiment GPA and predicted intervention effect, highlighting that the prediction intervention efficacy was higher for lower-GPA students. Additionally, predictions varied more for lower-GPA students, indicating a greater degree of interaction with other variables.

The count of blocked navigation events shows a clear negative relationship with predicted intervention effect. There were few students who attempted a large number of blocked navigation events; however, for those who did, predicted intervention effect was as low as \(-0.22\).

Reason for learning had little effect on predictions. However, Figure 8 shows that there was some variance on the y axis for high values (i.e., when reason for learning was focused on being a role model and helping others), which indicates interactions with other features. Race/ethnicity had similarly small effects, though what effect exists suggest that predictions were highest for White students, and lowest for Black/African American students.

| Variable                | Absolute effect on prediction |
|-------------------------|-------------------------------|
|                         | \( M \) | \( SD \)       |
| Pre-experiment GPA      | 0.037 | 0.028          |
| Blocked navigation count| 0.019 | 0.036          |
| Reason for learning     | 0.012 | 0.007          |
| Race/ethnicity          | 0.012 | 0.005          |
Figure 6. Feature importance plot for pre-experiment GPA in model 2.

Figure 7. Feature importance plot for the count of blocked navigation events. Most students experienced very few blocked navigation events, though it was an important feature for those few students who did attempt many such events.
Figure 8. Feature importance plot for self-reported reasons for learning. Jitter on the x axis added to aid visualization.

Figure 9. Feature importance plot for student race/ethnicity. Note the small y axis range, indicating little relationship between race/ethnicity and predictions.
4 Discussion
In this study, we were interested in discovering possible indicators of when the NSLM intervention would work, including as many predictive variables as possible to identify indicators that have not been previously explored. We utilized machine learning methods to account for complex variable interactions and to perform automatic feature selection, and measure the importance of features to ascertain relationships between variables and predictions. In this section we discuss our main findings from these analyses, implications of these findings, and possibilities for future work.

4.1 Main findings
We expected that prior GPA would predict whether or not student GPA would improve in the control condition, which was indeed the case. In general, GPA decreased as students transitioned to high school, and students with higher GPAs (and little room to improve) were predicted to experience the largest GPA decreases. However, students’ self-reported grades showed the opposite trend, and closely matched their expectations of success. This contrast between the effects of GPA and self-perceptions on predicted outcome may indicate differences in mindset, which merits further research. Students who have high expectations for their abilities may perceive their grades as higher, and may adopt a stronger growth mindset toward challenges in transitioning to high school.

We also expected demographic differences in both predicted GPA change (model 1) and predicted intervention effect (model 2), given previous research on race, ethnicity, gender identity, SES, and other demographic factors (Akos & Galassi, 2004; Benner, 2011; Benner & Graham, 2009; Bian, Leslie, Murphy, & Cimpian, 2018; Claro et al., 2016). In our results, SES and gender effects were notable for model 1 predictions, and race/ethnicity was for model 2. Directionality of these effects largely matches those in previous research, though the effect of SES in model 1 is notable. In particular, students for whom free/reduced-price lunch eligibility information was not available were predicted as having more negative GPA change than the other groups. This may be a function of systematic differences between schools for those that provided administrative data on this SES proxy, though it is unclear without additional data about those schools. Perhaps most importantly, demographic variables were not particularly strong predictors of intervention efficacy according to model 2, compared to pre-experiment GPA and the blocked navigation count procedural measure.

4.2 Implications for computer-administered learning mindset interventions
If prediction accuracy for model 2 was nearly perfect, the model could – in theory – be applied at the individual student level to identify those for whom the intervention would provide no benefit, and thus avoid wasting student time and computing resources on the intervention. In practice, however, prediction accuracy was low (though better than chance level), which is expected given the inherent difficulty of predicting a small intervention effect.

However, there are implications that can be drawn from model 2. First, the intervention appears, in general, to be more effective for lower-achieving students. This finding confirms previous work using other methods on the same dataset and other datasets (Sisk et al., 2018; Yeager et al., 2019). Thus, if a growth mindset intervention is to be targeted to a specific population, the best target is likely students with lower prior academic achievement. Second, demographic variables
were not especially important for model 2 predictions. Notably, the strongest demographic predictor (race/ethnicity) influenced predictions primarily less than 0.02 in either direction, which is less than the magnitude of the intervention effect. Thus, model 2 suggests that the intervention works – if not necessarily equally well – across demographic groups. Third, the count of blocked navigation attempts was a predictor of ineffective interventions for some students. Few students experienced many blocked navigation attempts, though predicted intervention effect was notably lower for those students. This result is especially crucial for administering computer-based mindset interventions, and merits future work with these students to discover their reasons for non-compliance with intervention procedures, and to develop methods that are more suited to their particular needs. Interventions may also benefit from increased teacher involvement if administered in classroom environments, specifically for the purpose of encouraging students to remain engaged with the intervention task.

Our findings also highlight the importance of selecting an appropriate imputation method for missing data when analyzing the intervention results. In particular, adding a unique category for missing data was key to the predictive value of the SES proxy variable (eligibility for free/reduced-price lunch). Strategies such as replacement with the mode or averaging across multiple branches of decision trees would diminish this effect.

4.3 Limitations and future work
There are a few limitations of the analyses in this paper that should be addressed in future work. First, the intervention effect (outcome) we focused on was limited to change in GPA. However, there may be other benefits. For example, adopting a growth mindset may encourage students to seek more challenging coursework in the future (Yeager et al., 2019), thereby promoting learning but not necessarily improving GPA. Our analyses could be repeated in future work with alternative outcomes including future course-taking behaviors and change in mindset itself, which may be beneficial in non-academic contexts as well (Dweck, 2006).

Second, the analysis of features in both machine learning models is abductive in nature. It is possible that there are multiple explanations for the predictive patterns found by the models (e.g., noise), since accuracy was far from perfect. However, given that we employed multiple forms of regularization tuned via nested cross-validation, trained on a nationally-representative dataset, and found effects in line with previous research, it appears likely that patterns captured by the models are not spurious. Future research is especially needed to systematically explore the connection between blocked navigation events and lower predicted intervention efficacy, since this is a novel finding with implications for computer-administered mindset interventions.

Third, this study focused on student-level predictors of intervention efficacy. However, there may also be school-level predictors that contribute to explaining variance in intervention efficacy, and which would be helpful to understand. For example, student-to-teacher ratio may relate to how much support students receive when they most need it, and race/ethnicity could be more important in some schools than other (e.g., when a student is a member of a minority group consisting of 2% of the student body versus 25% of the student body). Future work may thus offer additional insights by incorporating school-level predictors.
4.4 Concluding remarks

Computer-administered growth mindset interventions can be beneficial for learning, as results have shown. However, not every student benefits. It is important to understand which students and groups of students benefit, to avoid inequitable outcomes and enable more judicious use of resources. Thus, in this study we considered a large number of possible predictors of intervention efficacy and found that having a need for intervention (i.e., lower GPA) is important, but also that the mindset intervention was – unsurprisingly – ineffective when students did not follow intended procedures while interacting with the intervention software. Our findings will influence future research on compliance during computer-administered mindset intervention; ultimately, this will lead to improved mindsets for students, persistence in learning, and more learning.

References

Akos, P., & Galassi, J. P. (2004). Gender and race as variables in psychosocial adjustment to middle and high school. The Journal of Educational Research, 98(2), 102–108. https://doi.org/10.3200/JOER.98.2.102-108

Benner, A. D. (2011). The transition to high school: Current knowledge, future directions. Educational Psychology Review, 23(3), 299–328. https://doi.org/10.1007/s10648-011-9152-0

Benner, A. D., & Graham, S. (2009). The transition to high school as a developmental process among multiethnic urban youth. Child Development, 80(2), 356–376. https://doi.org/10.1111/j.1467-8624.2009.01265.x

Bian, L., Leslie, S.-J., Murphy, M. C., & Cimpian, A. (2018). Messages about brilliance undermine women’s interest in educational and professional opportunities. Journal of Experimental Social Psychology, 76, 404–420. https://doi.org/10.1016/j.jesp.2017.11.006

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chestnut, E. K., Lei, R. F., Leslie, S.-J., & Cimpian, A. (2018). The myth that only brilliant people are good at math and its implications for diversity. Education Sciences, 8(2), 65. https://doi.org/10.3390/eduscience8020065

Claro, S., Paunesku, D., & Dweck, C. S. (2016). Growth mindset tempers the effects of poverty on academic achievement. Proceedings of the National Academy of Sciences, 113(31), 8664–8668. https://doi.org/10.1073/pnas.1608207113

Dweck, C. S. (2006). Mindset: The New Psychology of Success. New York, NY: Random House.

Hulme, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. Science, 326(5958), 1410–1412. https://doi.org/10.1126/science.1177067

Kizilcec, R. F., & Goldfarb, D. (2019). Growth mindset predicts student achievement and behavior in mobile learning. Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale, 8:1–8:10. https://doi.org/10.1145/3330430.3333632

Kuhn, M. (2008). Building predictive models in R using the caret package. Journal of Statistical Software, 28(5), 1–26.

Leslie, S.-J., Cimpian, A., Meyer, M., & Freeland, E. (2015). Expectations of brilliance underlie gender distributions across academic disciplines. Science, 347(6219), 262–265. https://doi.org/10.1126/science.1261375
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 (pp. 4765–4774). Curran Associates, Inc.

O’Rourke, E., Haimovitz, K., Ballweber, C., Dweck, C., & Popović, Z. (2014). Brain points: A growth mindset incentive structure boosts persistence in an educational game. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 3339–3348. https://doi.org/10.1145/2556288.2557157

R Core Team. (2013). R: A language and environment for statistical computing.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. Psychological Science, 29(4), 549–571. https://doi.org/10.1177/0956797617739704

Stephens, N. M., Fryberg, S. A., Markus, H. R., Johnson, C. S., & Covarrubias, R. (2012). Unseen disadvantage: How American universities’ focus on independence undermines the academic performance of first-generation college students. Journal of Personality and Social Psychology, 102(6), 1178–1197. https://doi.org/10.1037/a0027143

Wright, S. J. (2015). Coordinate descent algorithms. Mathematical Programming, 151(1), 3–34. https://doi.org/10.1007/s10107-015-0892-3

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., … Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves achievement. Nature, 1–6. https://doi.org/10.1038/s41586-019-1466-y

Yeager, D. S., Henderson, M. D., D’Mello, S. K., Paunesku, D., Walton, G. M., Spitzer, B. J., & Duckworth, A. L. (2014). Boring but important: A self-transcendent purpose for learning fosters academic self-regulation. Journal of Personality and Social Psychology, 107(4), 559–580. https://doi.org/10.1037/a0037637

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., … Dweck, C. S. (2016). Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. Journal of Educational Psychology, 108(3), 374–391. https://doi.org/10.1037/edu0000098