# Estimating Student Knowledge from Paired Interaction Data

Anna N. Rafferty
Computer Science Division
University of California
Berkeley, CA 94720
rafferty@cs.berkeley.edu

Jodi Davenport
WestEd
Oakland, CA 94612
jdavenp@wested.org

Emma Brunskill
Computer Science
Department
Carnegie Mellon University
Pittsburgh, PA 15213
ebrun@cs.cmu.edu

## ABSTRACT

Estimating students' knowledge based on their interactions with computer-based tutors has the potential to improve learning by decreasing time taking assessments and facilitating personalized interventions. Although there exist good student models for relatively structured topics and tutors, less progress has been made with more open-ended activities. Further, students often complete activities in pairs rather than individually, with no coding to indicate who performed each action. We investigate whether pair interactions with an open-ended chemistry tutor can be used to predict individual student post test performance. Using $L_1$-regularized regression, we show that student interactions with the tutor are predictive both of the average post-test score for the pair and of individual scores. Towards better understanding pair dynamics in this setting, we also find that for pairs composed of students with similar pre-test scores, we can predict the difference in students' post-test scores.

## Keywords

Collaboration, embedded assessment, supervised learning

## 1. INTRODUCTION

Computer-based educational activities have many advantages over traditional tests as a means to assess student knowledge. The function of testing is to provide information about student proficiency. If an analysis of how a student completes an activity can provide similar information, time-intensive post-tests can be eliminated, and students can have access to the immediate feedback known to support learning. Projects such as ASSISTments [13] and stealth assessment [15] have demonstrated the potential for this approach.

Both interactive activities specifically designed for assessment and traditional intelligent tutoring systems provide valuable information about student knowledge. Simulation-based activities that are designed to be assessments have proven effective for measuring science inquiry and reasoning skills (e.g., [5, 12]). Many tutoring systems use student modeling to estimate proficiency as a student works through problems in the tutor. However, estimating students' knowledge based on their work in games and more open-ended environments introduces new challenges [3]. These environments are less structured, lack explicit tags about which tasks correspond to which skills, and may offer few opportunities to practice the same skill repeatedly in a similar context. Despite these challenges, games or more open-ended environments are important as they enable different forms of learning and can be used where formal testing is impractical.

A further challenge in estimating student knowledge from computer-based activities is that in classroom environments, students often work with computers in groups. Though collaboration can improve students' learning from computer-based science activities, automatically logged data rarely captures explicit collaboration, such as which student provided any given input, or what conversations occurred in conjunction with the activity.

In this paper, we explore whether machine learning based approaches can predict student knowledge based on interactions with an open-ended chemistry tutor, ChemVLab+. Due to limitations in the number of available computers in many classrooms, students generally use ChemVLab+ in pairs, and we analyze only data from paired interactions. We investigate what predictions we can make about individual student knowledge, corroborated by a separate post-test, based on the students' interactions with ChemVLab+.

## 2. BACKGROUND

We briefly review the literature on open-ended environments and collaboration in computer-based educational activities.

### 2.1 Open-ended tutoring environments

Many computer-based educational environments have open-ended components in which students explore topics using free-form actions. One approach to understanding student learning is to identify behaviors that are correlated with high or low learning gains. For instance, the WISE platform has identified patterns of inquiry behavior that are common in more successful students [10]. Kinnebrew, Loretz, and Biswas [8] identified patterns of student actions associated with periods of productivity and analyzed which patterns were correlated with high learning gains. In contrast

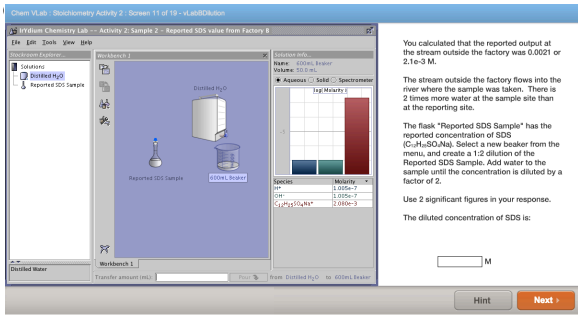


**Figure 1: A screenshot of a virtual lab activity.**

to identifying correlations between strategies and post-test performance, we focus on predicting student post-test scores given tutor interaction features.

Another area of work in open-ended learning environments focuses on recognizing strategies and goals. For example, Ha et al. [6] used Markov logic networks to identify student goals based on game behaviors. In the chemistry domain, progress has also been made on identifying semantic actions, such as a titration, based on individual behaviors in a virtual lab, such as pouring a small amount from one beaker to another [1, 4]. Strategy recognition provides descriptive information about student behaviors, but doesn't relate strategies to learning gains. Incorporate rich strategy features to predicting student understanding and post-test performance is an interesting future direction.

## 2.2 Data mining of student collaborations

Collaborative work is common in educational activities, and research provides evidence that collaborative work improves student learning [14, 21, 22]. Much work on computer-supported collaboration focuses on modeling the collaborative process and on how collaborative activities are related to learning and knowledge [18, 9]. Within this work, there is often explicit record of collaborative activities in the form of audio recordings, observation logs, or community interactions such as comments on a discussion forum [7, 16, 17]. Explicit records enable a system to learn classifiers to characterize individual student interactions and identify the roles of individual students in a collaboration [11, 19]. In contrast, our data lacks such explicit records of collaboration, a typical situation in the classroom due to practical difficulties and teacher preferences. Thus, we believe that exploring data from paired interactions in ChemVLab+ may be relevant to other educational tools as well.

# 3. CHEMVLAB+

ChemVlab+ is a collection of online activities that allow students to apply their chemistry knowledge in authentic, real-world contexts [2]. Each activity involves a separate problem, such as whether factories are reporting accurate pollution levels, and consists of a series of pages. ChemVLab+ activities include both freeform actions, such as virtual labs (see Figure 1), and more constrained actions, such as multiple choice questions.[1] The virtual labs enable similar actions as in a real chemistry lab: students manipulate beakers and use chemical instruments. These virtual labs are a key part of ChemVLab+ as they allow students to plan and execute experiments to investigate the real-world problem.

The data we analyze were collected from three schools during the 2011-2012 school year. Students first completed a paper and pencil pre-test. They then completed four ChemVLab+ stoichiometry activities using computers in the classroom; activities were completed in the same order by all students and over at least four class periods. Shortly after the final use of the ChemVLab+ activities, the students completed a post-test, which was identical to the pre-test. The test contained multiple choice and numerical free response items; the topics covered were similar to those in ChemVLab+, and there were a total of 30 points on the test. In the data, all 266 students completed the activities in assigned groups of two students, but completed pre- and post-test individually.

# 4. PREDICTING POST-TEST SCORES

We now explore what we can learn about students' knowledge based on their interactions with ChemVLab+, beginning with prediction of post-test scores. Paired performance data cannot necessarily tell us about individuals: intraclass correlation shows that for our data, the two post-test scores in the pair are not significantly correlated ($r = 0.12$, $p = .08$, $n.s.$). However, we will shortly see we still can predict some interesting aspects of individual and pair performance.

## 4.1 Methods

For all analyses, we use methods based on lasso ($L_1$-norm) regularization [20]. Lasso regularization is a popular machine learning method that adds a penalty term $\lambda||\beta||_1$ to the objective in traditional linear regression approaches, where $\beta$ is the vector of predictor coefficients and $\lambda$ is a scaling factor. This term favors solutions where many features have weight zero, even if this results in some increase in error; larger values of $\lambda$ favor using fewer predictors. Thus, feature selection is performed as part of the regression algorithm.

We computed features for each pair of students based on their behavior in the four ChemVLab+ activities. Twelve features were used for each activity, including four features based on help seeking and submission behavior on each page, four features for activity in the virtual labs, and four features capturing holistic behavior in the activity (e.g., total time on task). In some cases, a pair did not complete any pages in an activity, generally due to being absent from school. In these cases, we set the value of the feature for number of pages completed in the activity to zero. For all other features in that activity, we use feature imputation and set their values to the average value of that feature for other students.[2]

As students use ChemVLab+ in pairs but take the post-test separately, we predict three possible quantities using the interaction data: the higher of the two post-test scores, the lower of the scores, and the average of the scores. Note that if we use only pair interaction data, we can predict the higher of the two scores, but not which student will get which score. For each analysis, we want to maximize the proportion of the data that can be used for training while

[1] Activities can be found at `http://www.chemvlab.org`.

[2] We also tried imputation using data only from pairs who were similar to the current pair on the other activities; this did not significantly affect predictive performance.

| Prediction task | MAD by features for regression | |
|---|---|---|
| | ChemVLab+ | Pretest |
| Avg. post-test score | 2.6 | 2.5 |
| Higher post-test score | 3.0 | 2.9 |
| Lower post-test score | 3.2 | 3.3 |

**Table 1: Regression error for post-test predictions.**

minimizing overfitting. Given the relatively small dataset of 133 pairs of students, we use linear regression, which does not include interactions among features. By excluding interactions, we limit some of the risk of overfitting due to chance relationships among features. We fit the regression using 10-fold cross validation and limit the maximum number of features that can have non-zero coefficients to 20.

Accuracy is measured as the mean absolute deviation (MAD) for predictions for all pairs: $\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|\hat{Y} - Y|$, where $\hat{Y}$ is the predicted post-test score, $Y$ is the true post-test score, and $n = 133$ is the number of pairs. Lower MAD values indicate more accurate performance. We compare the performance of regression using the features based on tutor-student interactions versus using only pre-test features. The pre-test is highly correlated with the post-test score ($r(265) = 0.67$, $p < .001$), so we would expect pre-test scores to be relatively accurate predictors of post-test scores. To predict the average post-test score using pre-test features, we have a feature for the higher pre-test score in the pair and the lower score. For predicting individual post-test scores using pre-test features, we use the student's pre-test score.

## 4.2 Results

As shown in Table 1, the tutor-student interaction features achieve comparable predictive performance to using the pre-test features to predict student performance, and both provide quite accurate estimates. This suggests that even without prior information about the students, interaction data alone can provide useful indicators of student knowledge, despite the additional challenge that all interaction data comes from paired performance. Both sets of features are slightly better at predicting the average post-test score for the pair, which has somewhat less variance, and both are slightly worse at predicting the lower score. The decrease in accuracy for the pre-test features on the latter target is likely because the lower score has a smaller correlation with the pre-test score than the higher score ($r(132) = 0.48$ versus $r(132) = 0.71$; for both, $p < .001$). For all analyses, we also examined using both interaction and pretest features, but this did not significantly improve performance, suggesting that the two types of features capture similar information.

Lasso regression favors sparse solutions: the regression models used between 9 and 14 features, with the model for predicting the lower post-test score having the fewest features and the model for predicting the higher score having the most. All models included features from each activity as well as virtual lab features. Overall, these results demonstrate that the interaction data are relatively accurate predictors of post-test scores, despite the variety of tasks and the lack of a model of learning in the tutor. We also explored predicting learning gains based on the interaction data, but had less success, probably due to the choice of features. Our features captured behavior averaged across pages, but did not take into account changes in behavior from page to page.

# 5. TOWARDS RECOGNIZING HOW PAIRINGS AFFECT LEARNING

The previous section demonstrates the potential for using the ChemVLab+ activities as embedded assessments. We now explore what we can learn about pairs as a unit by predicting the difference between the two post-test scores in the pair. When restricted to pairs with similar pretest scores, large differences in post-test scores may signal a lack of collaboration, which could be used to drive interventions. Predicting differences in post-test scores may also reveal interaction features related to collaboration.

## 5.1 Methods

Lasso regression is again used for prediction and feature selection, with the same 48 features as in the previous analysis. 10-fold cross validation is used to fit the model, and the regression is limited to 20 features with non-zero weights. In analyses with pretest features, these features are the highest pretest in the pair, the lowest pretest in the pair, and the difference between the two pre-test scores.

## 5.2 Results

We first predicted differences in post-test scores for all pairs. The average difference in post-test scores was 6.0 points, with a standard deviation of 4.8 points. As shown in Table 2, prediction is relatively poor, and including both tutor interaction features and pre-test features did not increase performance. Due to concerns about overfitting, we limited the regression to linear features, which means the weight of each tutor feature is the same regardless of pretest-score. However, we might expect that these weights should be dependent on the pre-test scores. For instance, in a pair with dissimilar pre-test scores, high rates of hint reading might be indicative of a lack of collaboration. In pairs with similar pre-test scores, rates of hint reading might be less predictive because both students are likely to benefit from the hints.

To address this issue, we restricted the regression to the 43 pairs who had pre-test scores that were within two points of one another. The average difference in post-test score for these pairs was 4.9 points ($SD$= 4.1), and only about one-third of the pairs have post-test scores that are within two points of one another. Regressing on pairs with similar pre-test scores results in substantially lower prediction error than when all pairs are included (Table 2). Prediction is much more accurate than the standard deviation, and the interaction features result in more accurate predictions than the pre-test scores. For the analysis using the interaction features, twelve of these features had non-zero coefficients, including six features based on behavior in the virtual lab.

The previous analysis showed that we can predict differences in post-test score for pairs with similar initial knowledge. However, it does not tell us how initial knowledge and collaboration interact. Just as features and weights for predicting differences in post-test scores may differ for pairs based on the similarity of their pre-test scores, the regression may differ for pairs with different levels of initial knowledge. To explore this issue, we performed two additional analyses: predicting post-test scores for only those pairs where both students had below-average pre-test scores (*low* pairs) and predicting post-test scores for only those pairs where both

| Pairs included | MAD by features for regression | |
|---|---|---|
| | ChemVLab+ | Pretest |
| All | 3.9 | 3.7 |
| Similar pre-test | 2.0 | 3.4 |
| *High* pre-test | 3.0 | 3.2 |
| *Low* pre-test | 2.7 | 3.6 |

**Table 2: Regression error for predicting differences between the post-test scores for students in the pair.**

students had above-average pre-test scores (*high* pairs). The average pre-test score was 12.5 points out of 30.

The 35 *high* pairs had an average post-test score difference of 5.5 points ($SD$= 3.8). As shown in Table 2, this difference can be predicted relatively accurately. The most notable thing about this analysis, though, is that only two features are given non-zero weights. The small number of features suggests that when students have high initial knowledge, few features are indicative of the quality of collaboration.

In contrast, eight features have non-zero weight when predicting differences in post-test for the 43 *low* pairs. These pairs had an average post-test score difference 4.6 points ($SD$= 4.3), and the interaction features are more accurate predictors than the pretest features (Table 2). The features with non-zero weight included three lab features and at least one feature from each activity. One feature associated with smaller differences in post-test scores, due to having a relatively large negative weight, was the average number of submissions per page in Activity 2. This activity was difficult for students, and a lower number of submissions may have indicated that students were combining their knowledge, which is likely to result in more similar post-test scores.

# 6. CONCLUSIONS

Given differences in classroom implementations and the pedagogical benefits of more open-ended tutors, there are many advantages to predicting student performance based on real-world use of these systems. In this paper, we examined data from a series of chemistry activities that students completed in pairs, and found that pairs' interactions with the activities were predictive of individual post-test scores. Though we could make some predictions about differences in post-test scores for a pair, there is likely to be a limit on how well we can perform this task given the lack of data about individuals within the pair. We plan to explore how limited data about individual behavior, collected via classroom observation, can be used to create more accurate models of collaboration, and whether explicitly modeling control of the computer as a latent variable can improve performance. We would also like to explore a broader feature set, including features that capture changes in performance over time and more fine-grained virtual lab features (e.g., from pattern-mining [8]). We see this work as a first step in showing the potential of data mining techniques to transform collaborative educational activities into embedded assessments, even when activities are not designed for this purpose.

**Acknowledgements.** This research was supported an NDSEG Fellowship to ANR and by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A100069 to WestEd. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

# 7. REFERENCES

[1] O. Amir and Y. Gal. Plan recognition in virtual laboratories. In *IJCAI*, 2011.

[2] J. L. Davenport, A. N. Rafferty, M. J. Timms, D. Yaron, and M. Karabinos. ChemVLab+: Evaluating a virtual lab tutor for high school chemistry. In *ICLS*, 2012.

[3] S. De Freitas and M. Oliver. How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Comput Educ*, 46(3):249–264, 2006.

[4] Y. Gal, E. Yamangil, S. M. Shieber, A. Rubin, and B. J. Grosz. Towards collaborative intelligent tutors: Automated recognition of users' strategies. In *ITS*, 2008.

[5] J. Gobert, M. Sao Pedro, R. S. Baker, E. Toto, and O. Montalvo. Leveraging educational data mining for real time performance assessment of scientific inquiry skills within microworlds. *J Educ Data Mining*, 4:153–185, 2012.

[6] E. Y. Ha, J. P. Rowe, B. W. Mott, and J. C. Lester. Goal recognition with Markov logic networks for player-adaptive games. In *AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 2011.

[7] M. Kapur. Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *IJCSCL*, 6:39–56, 2011.

[8] J. Kinnebrew, K. Loretz, and G. Biswas. A contextualized, differential sequence mining method to derive students' learning behavior patterns. *J Educ Data Mining*, in press.

[9] Y. Li, J. Wang, J. Liao, D. Zhao, and R. Huang. Assessing collaborative process in CSCL with an intelligent content analysis toolkit. In *ICALT*, 2007.

[10] K. W. McElhaney and M. C. Linn. Impacts of students' experimentation using a dynamic visualization on their understanding of motion. In *ICLS*, 2008.

[11] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaïane. Clustering and sequential pattern mining of online collaborative learning data. *IEEE Transactions on Knowledge and Data Engineering*, 21(6):759–772, 2009.

[12] E. S. Quellmalz, J. L. Davenport, M. J. Timms, G. DeBoer, K. Jordan, K. Huang, and B. Buckley. Next-generation environments for assessing and promoting complex science learning. *J Educ Psychol*, in press.

[13] L. Razzaq, M. Feng, G. Nuzzo-Jones, N. T. Heffernan, K. R. Koedinger, B. Junker, et al. The Assistment project: Blending assessment and assisting. In *ITS*, 2005.

[14] B. B. Schwarz, Y. Neuman, and S. Biezuner. Two wrongs may make a right... if they argue together! *Cognition Instruct*, 18(4):461–494, 2000.

[15] V. Shute. Stealth assessment in computer-based games to support learning. *Computer Games and Instruction*, 2011.

[16] A. Soller, J. Wiebe, and A. Lesgold. A machine learning approach to assessing knowledge sharing during collaborative learning activities. In *CSCL*, 2002.

[17] G. Stahl. Meaning and interpretation in collaboration. In *CSCL*, 2003.

[18] G. Stahl, T. Koschmann, and D. Suthers. Computer supported collaborative learning: An historical perspective. In R. K. Sawyer, editor, *Cambridge Handbook of the Learning Sciences*. Cambridge University Press, 2006.

[19] L. Talavera and E. Gaudioso. Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *AI in CSCL at ECAI*, 2004.

[20] R. Tibshirani. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B Met*, pages 267–288, 1996.

[21] F. I. Winters and P. A. Alexander. Peer collaboration: The relation of regulatory behaviors to learning with hypermedia. *Instr Sci*, 30(4):407–427, 2011.

[22] F. I. Winters and R. Azevedo. High-school students' regulation of learning during computer-based science inquiry. *J Educ Comput Res*, 33(2):189–217, 2005.