# Student Behavioral Embeddings and Their Relationship to Outcomes in a Collaborative Online Course

Renzhe Yu
UC Irvine
renzhey@uci.edu

Zachary Pardos
UC Berkeley
pardos@berkeley.edu

John Scott
UC Berkeley
jmscott212@berkeley.edu

## ABSTRACT

In online collaborative learning environments, prior work has found moderate success in correlating behaviors to learning after passing them through the lens of human knowledge (e.g., hand labeled content taxonomies). However, these manual approaches may not be cost-effective for triggering in-time support, especially given the complexity of interpersonal and temporal behavioral patterns under rich interactions. In this paper, we test the hypothesis that a neural embedding of students that synthesizes their event-level course behaviors, without hand labels or knowledge about the specific course design, can be used to make predictions of desired outcomes and thus inform intelligent support at scale. While our student representations predicted student interactivity (i.e., sociality) measures, they failed to better predict course grades and grade improvement as compared to a naive baseline. We reflect on this result as a data point added to the nascent trend of raw student behaviors (e.g., clickstream) proving difficult to directly correlate to learning outcomes and discuss the implications for big education data modeling.

## Keywords

Collaborative learning environment, neural embedding, skip-gram, online course, higher education, behavior, predictive modeling

## 1. INTRODUCTION

Representation of collaborative learning behaviors in their raw formats has been challenging due to the complicated internal dependencies. Theory-driven approaches can extract some conceptually important measures of these learning processes but might not give good grounds for real-time learner support due to the human effort required. In this paper, we examine an aggregate, unsupervised representation of these collaborative learning behaviors in the context of a formal course that features sharing, remixing and interacting with student artifacts. We use a connectionist, neural network approach to representing a student as a function of a co-interaction network temporally formed by peers interacting in different ways in different weeks of the course. In reflection of the prior empirical work, we test the correspondence of these representations to learning outcomes. First, we investigate if the sociality of a student, or how much she is involved in the collaborative community, can be predicted from these low-level behavioral representations, as this is a direct goal of the special course design we analyze. Second, given the moderate relationship between interpersonal connections and learning performance in the literature, we test whether these vector representations are indicative of their final course performance. This exploration has strong pedagogical implications because an unsupervised student-level representation that captures signals of effective learning can be further deployed in intelligent systems to give just in-time feedback/interventions in the face of interconnected behavioral streams.

### 1.1 Collaborative learning behavior and outcomes

Generations of learning theories and pedagogies have highlighted the benefits of social processes for effective learning [15, 13]. Accordingly, there has been a multitude of studies that characterize these processes and examine how they relate to learning outcomes from granular learning behavior data [2]. One typical context of these studies is collaborative learning environments where students are required to work together in one way or another. As the interpersonal and temporal dependencies complicate the social processes, multiple methodological paradigms have been adopted to represent students' collaborative learning behavior.

To model the structures of interpersonal connections, social network analysis (SNA) conceptualizes learners as nodes and their various formats of interaction as edges and typically identifies global or local structures. Some studies are concentrated on the discovery of global structures such as core-periphery structures [6] and cohesive groups [3], while a number of others take more local perspectives and find the predictive power of network positions for learning outcomes [1, 5]. An alternative paradigm is the extension of psychometric or knowledge tracing models to collaborative settings, where collaboration status or group membership information is used to construct additional terms in the original functions [16, 9]. These adapted models have shown improved predictive power of students' learning performance.

The approaches above represent students' collaborative learning behaviors via theory-based or human-engineered models and each captures some dimensions that are predictive of various learning outcomes. At the same time, they might run the risk of misidentifying the model forms and leave some of the behavioral signals unattended, compared to more bottom-up, data-driven methods.

### 1.2 Connectionist student representation

In domains where it is difficult to enumerate and give values to features that satisfactorily represent the items, distributional approaches to modeling them might be useful. For example, the meaning of words in a lexicon is socially mediated and does not lend themselves well to description through feature engineering. Thus, the connectionist representation approach, which uses neural networks to learn a continuous feature vector representing all of the contexts of a word in a corpus, has become popular [8]. Similar challenges are present when it comes to positioning students based on their fine-grained behavior in open-ended learning environments. In response, recent research has attempted to apply connectionist models to learn a continuous vector of a student which represents all the contexts of her raw behaviors. For example, sequences of student responses in intelligent tutoring systems or student actions in MOOCs are used to map students to continuous vector spaces [14, 11]. Co-enrollment sequences with other students, although not in micro-level learning context, are used to represent undergraduate students throughout their degree [7]. While low-level behavioral embeddings have been used in non-social contexts to predict student performance, applying these techniques to collaborative settings may offer further insight into the complicated social processes.

## 2. DATASET

In this study, we analyzed a fully online course offered to residential students at a four-year public university in the United States. The course was focused on sociocultural aspects of literacy and global education. To facilitate collaborative learning, the course design featured a number of activities related to sharing, discussing, remixing and composing media with peer students. These activities were enabled by SuiteC, a toolkit that was integrated into the Canvas learning management system (LMS) [4]. There were three main components of SuiteC:

- Asset Library is a social platform where students contribute and share various media content in the form of "assets," and interact with peer assets by viewing, liking and commenting on them. Figure 1a shows the gateway page of the Asset Library with the feed of recently contributed asset.

- Whiteboards is an authoring tool that allows students to work individually or collaboratively on designing multimedia artifacts. Students can import assets as whiteboard elements and export finished whiteboards as assets for peer interaction. Figure 1b illustrates the interface when students collaborate on a whiteboard.

- Engagement Index is a gamification tool that tracks and evaluates student engagement in the SuiteC tools and provides a leaderboard for social comparison.

The course lasted for 14 weeks in Spring 2016. Each week except for the spring break, students worked through five activity phases that involved sharing, commenting, and creating assets and whiteboards, organized under course hashtags that students included in their posts. These SuiteC activities accounted for 25% of the final grade. whereas another 55% came from two long-form written papers that required students to integrate course readings. These two major assessments occurred around the middle and the end of the semester, respectively. The remaining 20% of the course grade consisted of eight ethnographic field notes authored by students on their site visits.

We acquired all the time-stamped click events within SuiteC for this course, with a total of 684,095 entries. Each entry recorded a granular action that a student performed on the foregoing tools, e.g. view an asset, add element to a whiteboard, etc. Attributes of the action included event type, timestamp, associated asset/whiteboard id, anonymized user id, user role, among others. After removing events that fell out of the normal period of the semester and that were not triggered by a student, we kept 658,967 entries for our analysis. In addition, the gradebook which contained scores for the two major assessments and the final course grades was also available.

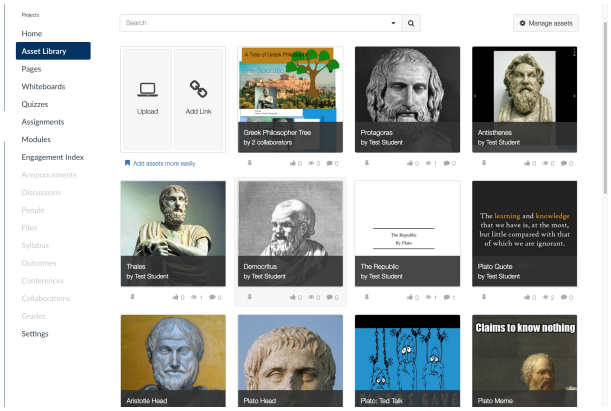
## 3. METHODS

### 3.1 Student representation using the skip-gram model

In this section, we describe our methodological approach to unsupervised student feature learning by way of neural embeddings. We model our student representation after [8], who used a neural network architecture called a skip-gram to learn word representations from their context distributions in a corpus. Given a word sequence $\{w_1, w_2, \ldots, w_T\}$, this model maximizes the average log probability of contextual words:

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t) \quad (1)$$

where $c$ is the contextual window size and the conditional probability $p(w_{t+j}|w_t)$ is computed using a softmax function over all possible words in the corpus for each given $w_t$. Because words that share meanings are more likely to occur in similar contexts, the word vectors of synonyms learned via this model should be in proximity in the high-dimensional space. Moreover, the learnt word vectors encode semantic relationships into interesting yet simple mathematical properties. For example,$\mathbf{v}_{Paris}$ is closest to $\mathbf{v}_{Berlin} - \mathbf{v}_{Germany} + \mathbf{v}_{France}$. This simplicity is why we are particularly interested in whether this technique can similarly characterize students from their complicated collaborative learning behaviors, thus facilitating easy identification of targeted actions (e.g. pairing students that sum up to a "beacon"). In our implementation, student sequences are constructed according to their order of appearance in the raw clickstream events sorted by time. We construct separate sequences for each week because, with the weekly course design, procrastinators for Week 1 and early birds for Week 2, although chronologically adjacent, may not share common traits. Moreover, as different event types in the raw dataset may or may not represent distinct behavioral signals, we ex-




(a) Gateway page of the Asset Library

(b) Interface of Whiteboard collaboration

Figure 1: SuiteC components

periment with three approaches to grouping raw event types:

- Raw event type grouping: There are 38 unique values in the "event" column of the raw data set, depicting the action that a student takes (e.g. create asset comment). We construct separate weekly sequences for each of these values and feed all resulting sequences to the model.
- Instructor coding grouping: We ask the instructor to group the 38 events based on their perceived nature to more accurately capture the kind of participation represented by a specific event. This process produces 15 groups, where each event belongs to one group only. For example, when students are authoring a Whiteboard, 9 different events could be triggered as they add shapes, assets, and free-hand drawing elements to their canvas, so all of nine events are categorized as "Whiteboard Composing." We separately construct weekly sequences for each group and feed all sequences to the model.
- No grouping: We do not differentiate event types and simply construct weekly sequences from the entire dataset.

Table 1 gives a generalized example of our approach. In the "raw event" approach, the "event" column contains the original event name in the dataset. For "instructor coding", that column is the group that the raw event belongs to. The "no grouping" approach, however, treats the columns as if filling the same value for all entries in the table. Whenever a student appears two more or times consecutively in a sequence, we remove the duplicate occurrence(s). In the remainder of this paper, we refer to this representation approach as *student2vec*. As for the hyperparameters of the model, we search $[8, 32, 64]$ for the vector size and $[5, 20, 40]$ for the contextual window size and plot all the results in Section 4.2.

## 3.2 Predicting sociality and learning outcome measures

We are interested in how well the unsupervised *student2vec* representations capture signals of students' social learning

Table 1: Example of student tokenization for connectionist representation (*student2vec*)

(a) Raw clickstream data table

| Timestamp | Week | Event | Student ID |
|---|---|---|---|
| 2/22 23:19 | 3 | View asset | 101 |
| 2/23 13:12 | 3 | Create whiteboard | 104 |
| 2/25 21:23 | 3 | View asset | 102 |
| 2/26 12:10 | 3 | Create whiteboard | 102 |
| 2/27 14:27 | 3 | Create whiteboard | 104 |
| 2/27 15:03 | 3 | View asset | 103 |
| 2/28 13:08 | 4 | Create whiteboard | 102 |
| 3/1 15:27 | 4 | View asset | 103 |
| 3/2 16:04 | 4 | Create whiteboard | 101 |
| 3/2 21:21 | 4 | Create whiteboard | 104 |
| 3/3 15:23 | 4 | View asset | 101 |
| 3/5 12:13 | 4 | View asset | 102 |

(b) Student sequences as input to the *student2vec* model

| Event × week | Student ID sequence |
|---|---|
| View asset, Week 3 | 101, 103 |
| Create whiteboard, Week 3 | 104, 102, 104 |
| View asset, Week 4 | 103, 101, 102 |
| Create whiteboard, Week 4 | 102, 101, 104 |

behavior as they conceptually do. Thus, we test the ability of these student vectors to predict an array of human-engineered measures of learning. We use predictive modeling as a more formal alternative to qualitatively examining algebraic properties of these vectors or looking at whether they exhibit meaningful clusters with respect to the learning measures.

First, we collaborate with the instructor[1] and construct four metrics of sociality (tendency to engage in interactive activities) for each student:

- median_asset_popularity: across all assets that a student (co-)creates throughout the semester, the median of their popularity values, where popularity of an asset

[1]The instructor is the third author on this paper

is defined as the unique number of non-author students who interact with it

- total_asset_popularity: across all assets that a student (co-)creates throughout the semester, the sum of their popularity values
- count_asset_authored: the total number of assets that a student (co-)creates throughout the semester
- count_peer_asset_visited: the total number of assets that a student interacts with of which she is not an author

The first two variables measure popularity, or "passive" processes of socialization, while the latter two capture "active" processes. All four variables are calculated from asset-related logged events, which is a tiny fraction ($\sim 5\%$) of all recorded activities.

We further look at course grades as reflected in formal assessments, including the following variables:

- final_score: the final grade in the gradebook, out of 100
- grade_gain: difference between the scores of the second (final_paper) and the first (midterm_paper) assessment

We then build models to predict these six measures using the learned student vectors. Because the number of data points is much smaller than in typical deep learning applications, we implement two simple models: linear regression and feed-forward neural network with a single layer of 8 neurons. Each dimension of the student vector serves as a feature in the model input. As the magnitude of these vectors might correlate with the number of occurrences of students and hence with sociality measures, we standardize them to unit length before feeding into the model. For each target measure, only students with valid values are included in the model training and testing processes. Four-fold cross-validations are performed for all the models and in each fold, 20% of the training data is used as the validation set during the training process to avoid overfitting.

# 4. RESULTS

## 4.1 Descriptive analysis

A summary of the basic statistics of six variables we use as prediction targets, plus the scores for two assessments is shown in Table 2. The inconsistent number of observations reflect missing values in some of the variables. The last two measures of students' activity have valid values for all 114 students appearing in the dataset. Among them, 15 students did not author any asset throughout the course and therefore have missing values for the two popularity measures. Moreover, only 79 students finished the course with grades.

All of the three course grades average 85-90 points with a standard deviation of around 6 points (out of 100). Also, the difference between median and mean is small for all three, suggesting relatively symmetric distributions. A student authored on average 4.8 assets a week (62.11 in total), which aligns with the weekly course requirements. Each of these 4.8 assets had around 2.4 peer visitors (150.36 in total). If we recalculate these measures only among students who received grades, the average number of authored assets goes up to 6.5, the popularity per asset remains similar with 2.1 peers, and the standard deviation of both measures shrinks substantially due to the removal of a large number of zero values (not reported here).

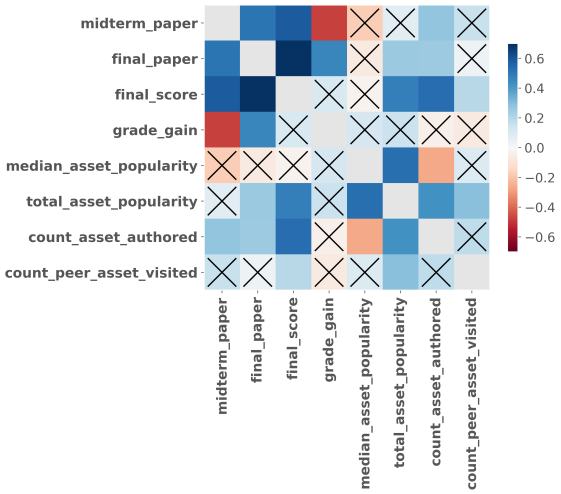



Figure 2: Rank correlation between learning outcome measures (first four rows/columns) and student interaction measures (last four rows/columns), with statistically insignificant correlations ($p > 0.1$) crossed out

## 4.2 Predictive analysis

We examine Spearman's rank correlations between course performance and sociality measures. Figure 2 depicts the correlation matrix in graphical terms. The three course grades are moderately to highly correlated with each other, all statistically significant at the 0.1 level (upper-left quadrant). The correlation between sociality and performance is more complicated (lower-left quadrant). In more cases the correlation is weak or insignificant, but two sociality measures (the number and the total popularity of assets authored) and two final outcomes (paper and course total) have moderate to high correlations. Lastly, the four sociality measures are mostly significantly correlated with each other, with low to moderate magnitudes (lower-right quadrant).

We illustrate the prediction performance by target variables in Figure 3. In each model configuration, rooted mean squared error (RMSE) is used as the evaluation metric for testing results. We define a naive baseline where the mean value of the training sample is used as the predicted value in each fold. To evaluate the performance of a model in relation to this baseline, we calculate the percentage of improvement from baseline:

$$\%\Delta RMSE = \frac{RMSE_{baseline} - RMSE_{model}}{RMSE_{baseline}} \quad (2)$$

Each histogram in Figure 3 depicts the $\%\Delta RMSE$ across different combinations of hyperparameters of *student2vec*,

Table 2: Descriptive statistics of main variables

| Variable | N | mean | std | min | median | max |
|---|---|---|---|---|---|---|
| midterm_paper | 79 | 88.21 | 6.18 | 72 | 88 | 100 |
| final_paper | 79 | 85.94 | 6.10 | 65 | 86 | 98.67 |
| final_score | 79 | 87.82 | 6.35 | 69.27 | 89.06 | 98.86 |
| grade_gain | 79 | -2.27 | 6.14 | -16 | -2.33 | 15 |
| median_asset_popularity | 99 | 1.52 | 1.48 | 0 | 1 | 10.5 |
| total_asset_popularity | 99 | 150.36 | 91.92 | 0 | 148 | 502 |
| count_asset_authored | 114 | 62.11 | 39.96 | 0 | 74.5 | 148 |
| count_peer_asset_visited | 114 | 154.46 | 134.74 | 0 | 132 | 534 |

Table 3: Summary of prediction error (RMSE) using the best-performing *student2vec* representation (vector size: 8; context window size: 20; event grouping: instructor's coding)

| Target | Baseline | Neural net (% improved) | Regression (% improved) |
|---|---|---|---|
| median_asset_popularity | 1.48 | 1.45 (2.26) | 1.50 (-1.08) |
| total_asset_popularity | 91.51 | 78.40 (14.33) | 80.27 (12.28) |
| count_asset_authored | 34.68 | 27.33 (21.19) | 27.57 (20.50) |
| count_peer_asset_visited | 129.90 | 119.36 (8.11) | 113.97 (12.26) |
| final_score | 6.39 | 6.39 (0.06) | 8.01 (-25.40) |
| grade_gain | 6.12 | 6.41 (-4.80) | 6.30 (-2.94) |

including vector size, contextual window size and event grouping (each combination referred to as a "case"). This approach to presenting results allows for a high-level view of the predictive power of this student representation approach. Figure 3a suggests that *student2vec* has moderate predictive power on sociality measures, especially the total amount of popularity a student gains and the number of assets a student authors where it can beat the baseline by 12% on average. By contrast, Figure 3b sees a complete failure of these student representations to predict learning outcomes: in most cases the prediction performance is outweighed by a naive baseline. These results suggest that the connectionist representation can, at least, extract low-level behavioral signals that relate to social processes but not those that contribute to performance.

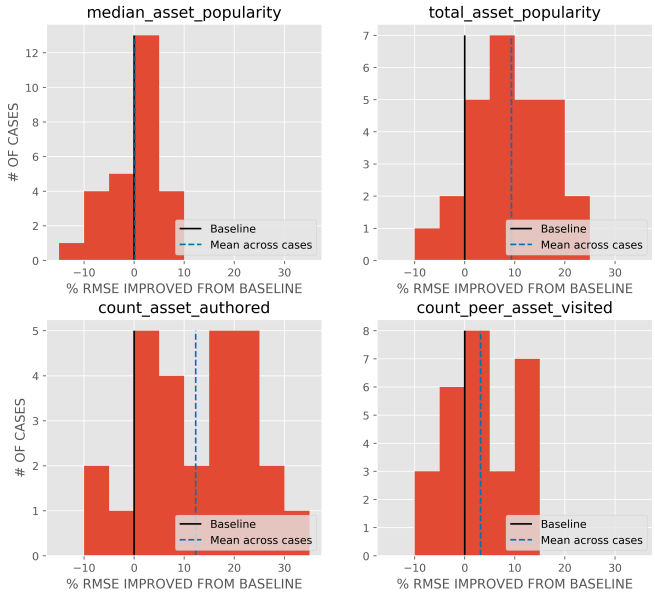
Finally, we qualitatively compare the performance of different cases. Across the three event grouping approaches, instructor coding produces similar performance to raw event, while both perform better in general than no grouping. To give an example of the best results, we select the vector size of 8, context window size of 20 coupled with instructor's coding, and report the detailed performance metrics associated with different prediction targets in Table 3. With regard to sociality measures, *student2vec* can improve the baseline RMSE by 8-21%, except for median_asset_popularity. In predicting course outcomes, however, this student representation performs 0.06% better than the baseline at best (6.39 vs. 6.39 for final_score).

# 5. DISCUSSIONS AND CONCLUSIONS

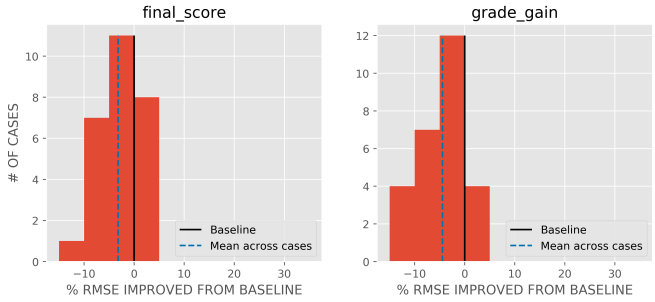
Granular learning process data in online learning environments afford the possibilities of real-time personalized learner support by way of detecting behavioral signals of unsuccessful learning. However, the correspondence between low-level student actions and their performance on assessments, outside of social pedagogies, has been a tenuous one, challenging this possibility in the wild. In the context of an edX MOOC, it was found that the addition of video viewing and other passive learning activity information did not improve prediction of future assessment performance beyond what past assessment performance alone achieved [10]. This result re-emerged in a college-level chemistry tutor setting, where past assessment performance alone predicted future assessment performance as well or better than if mixed with detailed eye-tracking telemetry [12].

The analyses presented in this paper reveal similar challenges yet some opportunity for using student clickstream from a mostly collaborative course to predict learning outcomes. We found that our representations of students, summarized from their low-level behaviors of sharing, creating, and socializing around artifacts, did correspond to human-engineered sociality measures, but not to assessed performance in the course as much as a naive baseline. Given our relatively low magnitude of data, an exceptionally high prediction accuracy was not expected, and the results may be seen as the lower bound of these representations' predictive power. However, their null relationship with summative assessment results still serves as another data point suggesting difficulty in linking raw behavior, absent of prior grade information, with assessment performance.

On the other hand, the model was able to predict measures of students' interactivity above baselines, and these manually engineered measures do not consistently predict course performance. These suggest that vector representations in general might not be the culprit. A similar methodology for representing undergraduate students also predicted on-time graduation with over 90% accuracy [7], an improvement over their baseline. These mixed results nudge us to reflect on the roles of data-driven behavioral representations and theory-based feature engineering [5, 9, 17] in building useful predictive models of student learning (and thus, support systems) in the context of collaborative learning. It is perhaps not enough to learn representations of students based on behavior without a more careful dissection of the nature of the behavior. This takeaway parallels the observation in




(a) Sociality prediction targets




(b) Learning outcome prediction targets

Figure 3: Histograms of prediction results for different target variables using *student2vec* representations. Each graph illustrates the performances of predicting the variable in its title across different combinations of model hyperparameters (i.e., "cases" on the y-axis).

EDM that refined knowledge component modeling is often necessary to accurately estimate cognitive mastery. Nevertheless, it was a natural expectation, in our data-driven approach, that similar students, in terms of when and what they do, would also be similar in their course outcomes. Although this turned out not to be the case in the instance we examined, it remains an open question for learning science researchers to consider if this is merely an anomaly or part of greater lesson to be learned on effective ways to fit behavior into the learner process tracing picture. For our research, a combination of interpretable activity representation and the current embedding approach may be tested in the future to gain some insights into the mechanism of interaction between the two in the learning process.

## 6. REFERENCES


[1] H. Cho, G. Gay, B. Davidson, and A. Ingraffea. Social networks, communication styles, and learning performance in a CSCL community. *Computers & Education*, 49(2):309–329, 2007.

[2] R. Ferguson and S. B. Shum. Social Learning Analytics: Five Approaches. In *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, pages 23–33, Vancouver, BC, Canada, 2012. ACM Press.

[3] N. Gillani and R. Eynon. Communication patterns in massively open online courses. *Internet and Higher Education*, 23:18–26, 2014.

[4] S. M. Jayaprakash, J. M. Scott, and P. Kerschen. Connectivist Learning Using SuiteC - Create, Connect, Collaborate, Compete! In *Practitioner Track Proceedings of the 7th International Learning Analytics & Knowledge Conference*, pages 69–76, Vancouver, BC, Canada, 2017.

[5] S. Joksimović, A. Manataki, D. Gašević, S. Dawson, V. Kovanović, and I. F. de Kereki. Translating network position into performance: Importance of centrality in different network configurations. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 314–323, Edinburgh, United Kingdom, 2016. ACM.

[6] S. B. Kellogg, S. Booth, and K. M. Oliver. A Social Network Perspective on Peer Support Learning in MOOCs for Educators. *International Review of Research in Open and Distance Learning*, 15(5):263–289, 2014.

[7] Y. Luo and Z. A. Pardos. Diagnosing University Student Subject Proficiency and Predicting Degree Completion in Vector Space. In *Proceedings of the Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, New Orleans, LA, USA, 2018.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pages 3111–3119. Curran Associates Inc., 2013.

[9] J. K. Olsen, V. Aleven, and N. Rummel. Predicting Student Performance In a Collaborative Learning Environment. In *Proceedings of the 8th International Conference on Educational Data Mining (EDM)*, 2015.

[10] Z. A. Pardos, Y. Bergner, D. T. Seaton, and D. E. Pritchard. Adapting Bayesian Knowledge Tracing to a Massive Open Online Course in edX. In *Proceedings of the 6th International Conference on Educational Data Mining (EDM)*, pages 137–144, Memphis, TN, 2013.

[11] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep Knowledge Tracing. In *Advances in Neural Information Processing Systems*, pages 505–513, 2015.

[12] M. A. Rau and Z. Pardos. Adding eye-tracking AOI data to models of representation skills does not improve prediction accuracy. In *Proceedings of the 9th International Conference on Educational Data Mining*, pages 622–623, 2016.

[13] G. Siemens. Connectivism : A Learning Theory for the Digital Age. *International Journal of Instructional Technology and Distance Learning*, 2(1):1–7, 2005.

[14] M. Teruel and L. A. Alemany. Co-embeddings for Student Modeling in Virtual Learning Environments. In *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, pages

73–80, Singapore, Singapore, 2018. ACM Press.

[15] L. S. Vygotsky. Interaction between Learning and Development. In *Mind in Society: Development of Higher Psychological Processes*, pages 71–91. Harvard University Press, Cambridge, MA, USA, 1978.

[16] M. Wilson, P. Gochyyev, and K. Scalise. Modeling Data From Collaborative Assessments: Learning in Digital Interactive Social Networks. *Journal of Educational Measurement*, 54(1):85–102, feb 2017.

[17] D. Yang, M. Wen, and C. Rose. Weakly Supervised Role Identification in Teamwork Interactions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1671–1680, Stroudsburg, PA, USA, 2015.