Exploratory Analyses of Data from a Collaborative Problem Solving Assessment

Junyan Yao(jy2269@nyu.edu), Kaushik Mohan(km4326@nyu.edu), Yoav Bergner(yoav.bergner@nyu.edu), and Peter Halpin(peter.halpin@nyu.edu)

Abstract

Technology-based assessments that involve collaboration among students offer many sources of process data, although it remains unclear which aspects of these data are most meaningful for making inferences about students' collaborative skills. Identification of informative features is a crucial step in helping students improve their collaborative skills. Recent research has focused largely on theory-based rubrics for coding of process data (e.g., text from chat dialogues, click-stream data), but many reliability and validity issues arise in the application of such rubrics. In this research, we take a more data-driven approach to the problem. Drawing on real data in which dyads collaborate via online chat to answer twelfth-grade mathematics items, we focus on features of chat and click-stream that can be extracted automatically

Data

Respondents were solicited using Amazon Mechanical Turk (AMT). The sampling frame for the study was comprised of AMT workers who self-reported to live in the United States and to speak English as their first language. The median age was 32 years, with an interquartile range of [27, 40]. The majority of the sampling frame (71%) self-identified as being of "White" ethnicity, 51% reported being female, and 88% reported having at least one year of post-secondary education. Two independent samples were taken from the sampling frame. A calibration sample ($N = 528$) was used to estimate item parameters of the 2PL model for a pool of $I = 60$ twelfth-grade mathematics items obtained from previous administrations of the National Assessment of Educational Progress (NAEP). Items had parameter estimates in the following ranges: $\hat{\beta}_i \in [-3.80, 2.62]$ and $\hat{\alpha}_i \in [0.65, 2.86]$. In the research sample ($N = 322$), all respondents were assessed under both individual and group testing conditions. The content of the individual and group tests were calibrated to have comparable difficulty and reliability and were counterbalanced. A demo of the assessment platform is available at `collaborative-assessment.org`

Methods

We demostrate regression method to test hypotheses about how these features in the students chat affect student's outcome score in terms of their collaboration.The dyadic performance outcome was computed using each respondent's average performance on the individual pre-test ($\bar{X}_i$) and their average performance on the conjunctively-score group test ($\bar{Y}$):

$$\text{Score} = \frac{\bar{Y} - \bar{X}_1\,\bar{X}_2}{\bar{X}_1 + \bar{X}_2 - 2\,\bar{X}_1\,\bar{X}_2}$$

The two test forms were calibrated to have comparable difficulty and reliability. The scoring formula is motivated by the SC-IRT model presented in Halpin & Bergner (in press). We considered the following features in our model:

- **Number of Words**: Calculated by using *space* as seperator
- **Topic Content Percentage**:Use the content from items and ideas from Information Retreival (TF-IDF). The formula for tf-idf is given as below:

$$\texttt{tf} - \texttt{idf}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} * \log \frac{N}{n_t}$$

  where $f_{t,d}$ is the frequency for a term t in document d, n_t is the number of documents where term t appears and N is the total number of documents. TF-IDF is used to identify the topic words in each document as it weights the word based on frequencey of appearance in a document while adjusting for the fact that some words appear more frequently in general. We use this statistics to emperically determine the probability of a word being a "topic word" by computing cumulative probability of it's TF-IDF from the distribution of TF-IDF in that document. Using this notion of the probability of topic content for each word in the item pool, we map them to the bag of words used in the chat by individuals working on the respective item. By averaging together these values for all the words used in the chat by a group, we arrive at the Topic-Content%.
- **Math Content Percentage**: We look for the use of mathematical symbols in the chat and classify them as Math Content, and calculate the math content percentage as the number of words tagged as Math-Content over the total number of words used in the chat.
- **Affect Score**: Use the existing lexicon packages such as *lexicon* and *tm*, to count chat content scores for each member in the team.
- **Group Synchronicity**: We capture synchronicity by observing the submission times for each item by each member of a group using clickstream data. The synchronicity for an item i for group g is given as follows: $ItemSync(g, i) = \frac{OverlappingTime(g,i)}{MaxTime(g,i)}$, where is the number of items worked on by members of group g. The overall synchronicity $Sync(g)$ for group g is the average of the item synchronicity over all the of items given as follows:$Sync(g) = \frac{1}{N_{i,g}} \sum_{i \in I} ItemSync(g, i)$, where $N_{i,g}$ is the number of items worked on by members of group g. Overall synchronicity is calculated as an equally-weighted average over the items and the time spent on each of the items is not considered
- **Average individual ability and individual ability difference**

Hypotheses

We hypothesize that:

1. The process features will account for significant variation in group performance, after controlling for chat word count.
   a. Topic content is positively related to group performance.
   b. Math content is positively related to group performance.
   c. Synchronization is positively related to group performance.
   d. Affect mirroring is positively related to group performance.
   e. Average affect is positively related to group performance.

Results

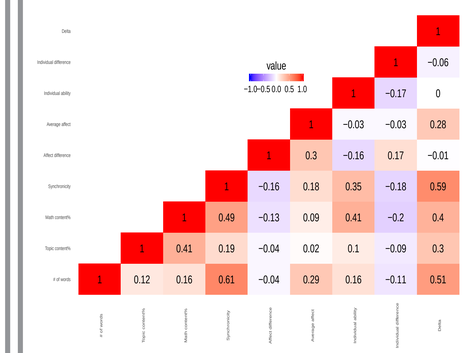
Hypothesis 1: The addition of process features accounted for a significant increment proportion of the variation explained (about 22%), with the total R-square equal to 51.63%. **Hypotheses a - e**: The only non-significant process feature was average affect. The raw correlations show that these measures were highly correlated with word count. The affect score distributions were examined in more detail for actual versus random pairs, showing little difference. The individual average ability is significant. This could result from the multicollinearity with math content% and Synchronization.

Model	Predictors	R_square
M1	Average individual ability + Individual ability difference	0.35%
M2	M1+ Number of words	30.36%
M3	M2+ Math-Content%, + Topic Content%+ Avg affect+ affect difference+ Synchronicity	51.76%

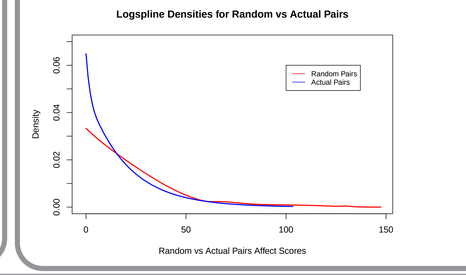
Model Output After Stepwise AIC Selection:

	Estimate	Std..Error	t.value	Pr...t..
Intercept	-1.2265	0.3677	-3.3359	0.0011
Individual average scores	-1.9462	0.5340	-3.6448	0.0004
Individual score difference	0.3370	0.4210	0.8007	0.4250
# of words	0.0006	0.0002	2.7208	0.0075
Topic content%	6.7162	3.5834	1.8743	0.0635
Math content%	4.7459	1.8315	2.5913	0.0108
Average affect score	6.1576	3.7313	1.6502	0.1016
Synchronicity	1.3580	0.3365	4.0352	0.0001

Correlation Matrix



Affect Score Distribution



Conclusion

This exploratory research showed that process features explain substantial variation team performance when dyads collaborate over online chat to answer NAEP math items. The results indicate that better team performance is predicted by when chat messages (a) include the content of the problems (b) use mathematical notation. We also found that synchronicity (working on the same problem at the same time) positively predicted performance. While affective chat content was associated with better team performance, this was highly corrected with raw word count. Future research should explore other automated methods of feature extraction and link these with research on teaching and learning.

Acknowledgement

This research was supported by a seed grant from the NYU Center for Data Science.