Assessing the Dialogic Properties of Classroom Discourse: Proportion Models for Imbalanced Classes

Andrew M. Olney
Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
aolney@memphis.edu

Borhan Samei
Institute for Intelligent Systems
University of Memphis
Memphis, TN 38152
bsamei@memphis.edu

Patrick J. Donnelly
Department of Computer Science
California State University
Chico, CA 95929
pjdonnelly@csuchico.edu

Sidney K. D'Mello
Departments of Psychology & Computer Science
Notre Dame University
Notre Dame, IN 46556
sdmello@nd.edu

ABSTRACT

Automatic assessment of dialogic properties of classroom discourse would benefit several widespread classroom observation protocols. However, in classrooms with low incidences of dialogic discourse, assessment can be highly biased against detecting dialogic properties. In this paper, we present an approach to addressing this imbalanced class problem. Rather than perform classifications at the utterance level, we aggregate feature vectors to classify proportions of dialogic properties at the class-session level and achieve a moderate correlation with actual proportions, $r(130) = .50,\ p < .001,\ CI_{95}[.36, .61]$. We show that this approach outperforms aggregating utterance level classifications, $r(130) = .27,\ p = .001,\ CI_{95}[.11, .43]$, is stable for both low and high dialogic classrooms, and is stable across both automatic speech recognition and human transcripts.

Keywords

dialogic instruction, questions, authenticity, machine learning, imbalanced classes

1. INTRODUCTION

Classroom observation for measuring teaching effectiveness is currently used in 47 states [1]. Simply stated, classroom observation involves a trained evaluator watching how a class is taught and using a rubric to score the teacher's performance. The widespread use of classroom observation is based on previous research which indicates that instructional quality has a greater impact on student achievement than class size, teacher experience, or teacher graduate education [16]. Beyond such research findings, classroom observation is also driven by the teacher accountability era coinciding with the passage of the federal No Child Left Behind Act, which mandated annual testing of students by all states. In this highly politicized environment, classroom observation is increasingly being used to determine teacher's salary and tenure.

Curiously, given the high stakes associated with classroom observation, the majority of research linking instructional quality to student achievement over the past several decades has been correlational only. However there has been an increasing interest in randomized controlled trials. One recent randomized trial is the multi-year Measures of Effective Teaching (MET), which tracked approximately 3,000 teachers in seven states [4]. In year 1, MET researchers built predictive models of teaching effectiveness, and in year 2, teachers were randomly assigned to new classrooms to test the predictive models from year 1. Major MET findings were that teaching effectiveness measured via classroom observation protocols correlated with achievement gains and that question asking behavior was a key component of variability in teaching quality [11].

Although instructional quality is linked to achievement, the current practice of assessing instructional quality through classroom observation is logistically complex and expensive, requiring observer rubrics, observer training, and continuous assessment to maintain a pool of qualified observers [2]. To address these practical challenges, our work has focused on the automated assessment of classroom discourse, with a particular emphasis on measuring dialogic questions in classrooms. Our approach is to automate an existing, fine grained classroom observation protocol that focuses on dialogic questions, known as the Classroom Language Assessment System[1] [13]. Unlike the classroom observation protocols used in the MET study, in which an observer makes rubric-based judgments approximately every 10 minutes, CLASS uses fine-grained coding at the question level, creating suitably detailed labeled data for machine learning purposes.

[1] CLASS denotes the CLASS created by Nystrand and colleagues, as opposed to the CLASS used in the MET study.

The dialogic instruction measured by CLASS is characterized by open-ended discussion and the exchange of ideas (cf. [3]), which in turn are characterized by questions that truly seek information (authentic questions) and which incorporate ideas from the student (questions with uptake). For example, “How did you feel by the end of the story?” is an authentic question because there is no pre-scripted response, and a follow-on question “Why do you think that is?” has uptake because “that” refers to the student’s previous reply. As is clear in these examples, dialogic properties are contextualized by the discourse such that the antecedents and consequents of the question shape whether a question is authentic or has uptake. Previous research using CLASS has shown that authenticity and uptake are significant predictors of student achievement [10, 9, 14].

Our project, which we call CLASS 5, seeks to fully automate classroom observations under the CLASS protocol. In our work, we have used archival data collected in previous CLASS projects, containing human transcripts of dialogic questions, as well as new data using automatic speech recognition (ASR) of teacher speech. Models built with archival human transcript data are as effective at classifying authenticity and uptake as humans on isolated questions [18]. However, as we began to analyze the new CLASS 5 data, we realized that there were two serious limitations undermining our existing models. First, the archival data used in previous work [18, 17] contained only transcripts of questions, and even these did not represent all questions but a subset of questions that were *instructional*, and so excluded rhetorical questions, procedural questions, and discourse management questions [13]. In the archival data, approximately 50% of the questions were coded as authentic questions. In contrast, the new CLASS 5 data included all questions and non-questions, i.e. all utterances, from which authentic questions must be detected. Secondly, in the CLASS 5 data, the base rates for dialogic properties were dramatically lower than in previous samples. For example, authentic questions in our new data collection constituted about 30% of instructional questions compared to approximately 50% of instructional questions in the archival data; moreover, authentic questions in our new data constituted only about 3% of all utterances. Therefore to be robust in detecting dialogic properties across samples, our models must be able to deal adequately with imbalanced classes.

The so-called “class imbalance problem” is well known in the data mining community, and has been proposed as one of data mining’s top 10 challenging problems [20]. The essence of the problem is that a classifier can maximize accuracy by always selecting the majority class and that this strategy, typically considered as a baseline for performance, becomes increasingly hard to beat as the majority class distribution approaches 100%. A review of the class imbalance problem describes three major approaches for addressing it [8]. First, algorithmic approaches may be used to bias learning towards the minority class. Secondly, preprocessing methods may change the class distribution before learning occurs, either by undersampling the majority class or oversampling the minority class. Thirdly, cost-sensitive approaches may be used to assign higher costs, or weights, to minority class errors, such that the learning algorithm tries to minimize the total cost.

In this paper, we present another method for addressing the class imbalance problem, which is to transform the problem into a different problem that is easier to handle. Specifically, we explore the consequences of shifting from classifiers that classify utterances as authentic questions to classifiers that classify the proportion of authentic questions in a class session. As will be shown in the remainder of the paper, this problem transformation outperforms aggregating utterance level classifications, is stable for both low and high dialogic classrooms, and is stable across both automatic speech recognition and human transcripts.

2. METHOD

2.1 Data sets

CLASS 5 data. New data for the CLASS 5 project were collected between January 2014 and May 2016 at seven schools in rural Wisconsin. Observations for 132 class sessions taught by 14 different teachers were manually coded using the CLASS system, and audio was simultaneously recorded. Both teacher and school identifiers were preserved with the data. Given the logistical constraints of individual microphones for each student, the recording instrumentation instead focused on high quality teacher audio suitable for ASR that was recorded using a wireless microphone headset. Classroom audio, which included both teacher and student speech, was recorded from a stationary boundary microphone, and was not of sufficient quality to be used for ASR; however, it is useful for marking when students speak. The teacher audio was later automatically segmented into utterances and then submitted to a speech recognition service [6]. Thus this dataset differs from the archival data (see below) in that the transcripts are provided by ASR with its accompanying errors, only teacher speech is transcribed, and the transcripts contain all utterances rather than just instructional questions. The data contained 45,044 utterances, of which 1282 were authentic questions (3% of utterances; 30% of instructional questions) and 290 were questions with uptake (.01% of utterances; .07% of instructional questions). Authenticity and uptake are even more highly related in this data set than in the archival data since only 5 questions have uptake without authenticity. Given the small number of observations of uptake and the finding that virtually all questions with uptake are also authentic, we primarily focused on detecting authenticity.

Archival data. The archival data was collected during the Partnership for Literacy Study (Partnership), a study of professional development, instruction, and literacy outcomes in middle school English and language arts classrooms. The Partnership collected data from 7th- and 8th-grade English and language arts teachers in Wisconsin and New York from 2001 to 2003. Over that two-year period, 119 classes in 21 schools were observed twice in the fall and twice in the spring. Teacher identifiers were not embedded in the CLASS data files, and out of 119 teachers only 70 could be unequivocally matched to data files. However, school identifiers were directly embedded in data files. Classroom observations for Partnership were also conducted using the CLASS annotation system [13]. During this process instructional questions were transcribed, and the transcriptions were mostly accurate but not verbatim. Reliability studies using CLASS indicate that raters agree on question properties approximately 80% of the time, with observation-level inter-rater correlations averaging approximately .95 [14]. After removing questions with partially incomplete annotations, 25,711 instructional questions remained for use in our analyses, of which 12,862 were authentic questions (50%) and 5,489 were questions with uptake (22%). Authenticity and uptake were highly related: only 593 (2%) questions had uptake without authenticity.

2.2 Features

In early work, we established that word and part-of-speech features that are useful for classifying types of questions [15] were also useful for predicting dialogic question properties like authenticity and uptake [18, 17]. In the present work we have extended these 36 predictive features to include features obtained through syntactic and discourse parsing [12, 19]. At the word level, these new features include 45 part-of-speech tags as well as named entity type, which subdivides real world objects described by proper nouns into 13 classes including PERSON, LOCATION, and DATE. At the sentence level, the features include 47 syntactic dependencies like subject, agent, direct object, or indirect object. And at the discourse level, the features include 18 discourse relations including contrast, elaboration, and topic-change, as well as features for joint, nucleus, and satellite elementary discourse units. Because the discourse parse returns a tree of elementary discourse units, the discourse features were mapped to the sentence level by summing the discourse relations, satellite, joint, and nucleus features that occur in each elementary discourse unit composing the sentence. Anaphora resolution was converted into four features including the number of coreference chains in an utterance extending into future sentences, the sum of those chain's lengths, and the same features in the backwards direction. In other words, the anaphora features capture how well a sentence was connected to other sentences in both directions. While all features were encoded at the sentence/utterance level (i.e. a count of the feature in the utterance), the 36 question features used in previous work were additionally encoded as occurring at either the first token or after the first token. For example, if a *definition keyword* feature occurred in the first token, then that would be recorded as a single count in the corresponding overall feature and the first token feature, but not in the corresponding after the first token feature. Additionally, the named entity PERSON feature was encoded with first token and last token variants based on the observation that questions addressed to students typically use the name at the beginning or end of an utterance if at all. With the positional variants, there were 242 linguistic features in our models that span word, sentence, and discourse levels.

To generate these features we used the CLU processor, which contains syntactic and discourse parsers [19]. Because discourse parsing requires a discourse context, utterances for each classroom observation were grouped into separate files before parsing. The parsers were configured with a maximum sentence length of 120 words, which was empirically determined by observing the lengths of a subsample of utterances. Parses for each class-level file were converted into utterance level features and aggregated into a 242-dimension feature vector where the value at each position was the frequency count of a particular feature in that utterance. Models built at the question level for archival data or utterance level for new data used these 242-dimension feature vectors. Models built at the class-session level used these features but summed them over all questions (Partnership) or utterances (CLASS 5) in a given class. Models at the class-session level additionally added the means and standard deviations of these summed feature vectors, for a total of 726 features.

2.3 Model training

Cross validation. We used cross validation such that a given teacher would not appear in both the training and testing folds, in order to study generalizability to new teachers. For the CLASS 5 data, this was achieved using leave-one-teacher-out cross validation. For the archival Partnership data, the mapping between teachers and data files was incomplete and so the mapping between schools and data files was used instead. This leave-one-school-out cross validation makes the assumption that a teacher did not transfer between schools during the study (a likely assumption) and in a sense is even more conservative than leave-one-teacher-out validation because it controls for similarities shared by teachers at the same school. Ideally the same cross validation technique would be used for both data sets, but for CLASS 5 data there aren't enough schools (2) and for the Partnership data the teacher identifiers are incomplete.

Models. Different models were used depending on the nature of the task and the class imbalance. For question-level authenticity prediction in the archival Partnership data, where classes are balanced, a J48 decision tree was used. J48 models were chosen because of their previous performance on this task and data set [18]. For utterance-level authenticity prediction in the new CLASS 5 data, where classes are highly imbalanced, SMOTEBoost was selected [5]. SMOTEBoost combines oversampling of the minority class by synthesizing new exemplars (SMOTE) with boosting, which builds a serial ensemble of models such that each successive model increases the weight, or focus, to instances misclassified in the previous model. SMOTEBoost applies SMOTE in each of these successive models in order to improve accuracy over the minority class, and evidence suggests it is one of the best all-purpose algorithms for imbalanced problems, though not necessarily the fastest [8]. Several other algorithms were evaluated on this task, including k-nearest neighbors, random forests, various cost-sensitive classifiers, and various ensembles, but SMOTEBoost had the best utterance-level performance. For class-level authenticity prediction (for both Partnership and CLASS 5 data), M5P model trees, which are decision trees with regression functions at the leaves [7], were used to predict the proportion of authentic questions in the class period. As a comparison to the class-level models, we aggregated over the question- and utterance-level classifications to calculate a proportion score at the class level.

3. RESULTS & DISCUSSION

3.1 Proportion models for imbalanced data

Our first comparison was between class-session level proportion models and aggregated utterance level classifications for the new CLASS 5 data where authenticity was very rare. A M5P model trained to predict the proportion of authentic questions per class made predictions that had a significant correlation with the actual proportions, $r(130) = .50$, $p < .001$, $CI_{95}[.36, .61]$. A SMOTEBoost

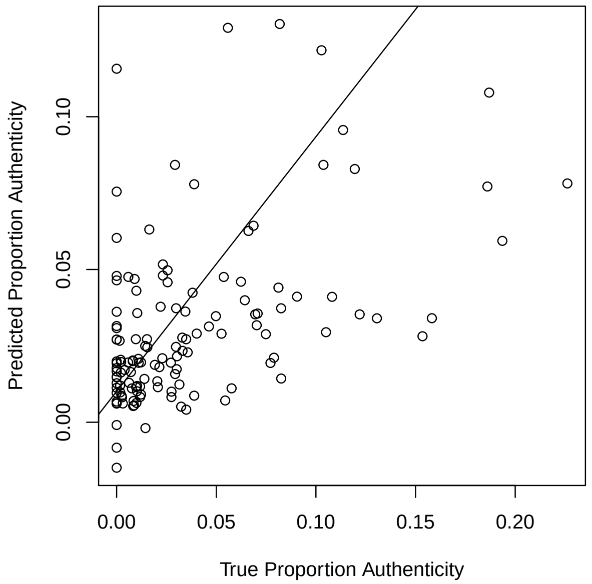

Figure 1: M5P session-level proportion predictions on the CLASS 5 data set.

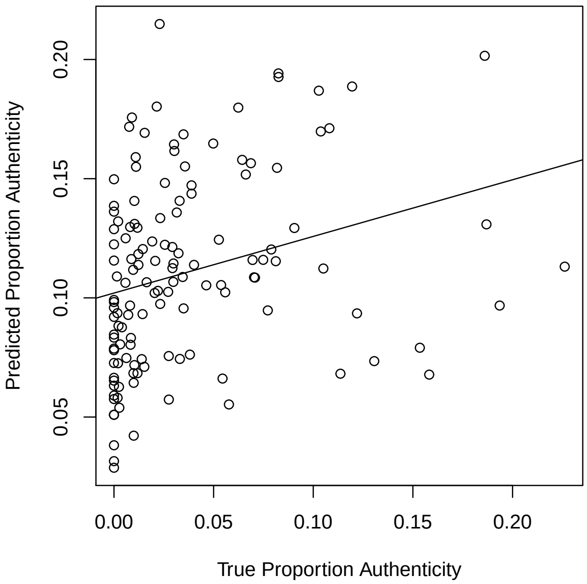

Figure 2: SMOTEBoost utterance-level predictions aggregated to session-level proportions on the CLASS 5 data set.

model trained to predict the authenticity of utterances and whose predictions were aggregated to obtain class-session level proportions made predictions that had a significant size correlation with actual proportions, $r(130) = .27,\ p = .001,\ CI_{95}[.11, .43]$. However, these two correlations were significantly different, $t(258) = 2.42,\ p = .017$. These results suggest that class-session level proportion predictions are more accurate than aggregating predictions from utterance level models.

Scatterplots of the actual vs. predicted proportion of authentic questions in the new CLASS 5 data are shown in Figures 1 and 2. Perhaps the major difference between these two scatterplots is the relationship between predicted and authentic proportions for values near zero. For the aggregated utterance-level predictions generated by SMOTEBoost, the scatterplot in Figure 2 shows a large vertical column of predictions above zero, indicating that for values near zero the classifier is overestimating the true occurrence of authentic questions. Conversely in Figure 1, predictions at zero are more tightly clustered.

Based on these results, it appears that session-level proportion models like M5P are more forgiving of the imbalanced classes than are utterance-level models like SMOTEBoost. There are two plausible explanations for why this might be. First, the session-level models are predicting a continuous number between 0 and 1 rather than making crisp binary judgments as in the case for the utterance-level models. Continuous predictions more closely match the model's internal probability, as opposed to a binary judgment where the binary prediction is the same irrespective of how far the model's internal probability is from the threshold, so long as it is on the same side of the threshold. Secondly, utterance-level models do not take advantage of the probability of a previous utterance's authenticity in determining the current utterance's authenticity, whereas the session-level models are accumulating all of this weak evidence before rending a proportion authenticity prediction. Based on this reasoning, an additional comparison of interest would be to take the utterance-level prediction probabilities and aggregate over them instead of the binary classifications. Unfortunately in the case of SMOTEBoost, these probabilities are within 10^{-6} of zero and one, so the results are no different than aggregating over class predictions.

3.2 Proportion model stability

To demonstrate model stability we undertook two comparisons. First, predictions of a M5P model for the Partnership data trained to predict the proportion of authentic questions per class session were significantly correlated with the actual proportions, $r(426) = .42,\ p < .001,\ CI_{95}[.34, .50]$. This correlation is remarkably similar to the 0.5 correlation obtained for the new CLASS 5 data. The similarity in correlations is particularly noteworthy given the differences between data sets: for CLASS 5, the classifier is operating over ASR transcribed utterances where authentic questions are 3% of the total data, but for the Partnership data, the classifier is operating over human transcribed instructional questions where authentic questions are 50% of the total data.

Secondly, a J48 model for the Partnership data trained to predict the authenticity of utterances and whose predictions were aggregated to class-session level proportions made predictions that were correlated with actual proportions, $r(426) = .44, p < .001,\ CI_{95}[.36, .51]$. These two correlations were not significantly different, $t(870) = .37$, $p = .71$. Scatterplots of the actual vs. predicted proportion authentic questions in the Partnership data in Figures 3 and 4 further

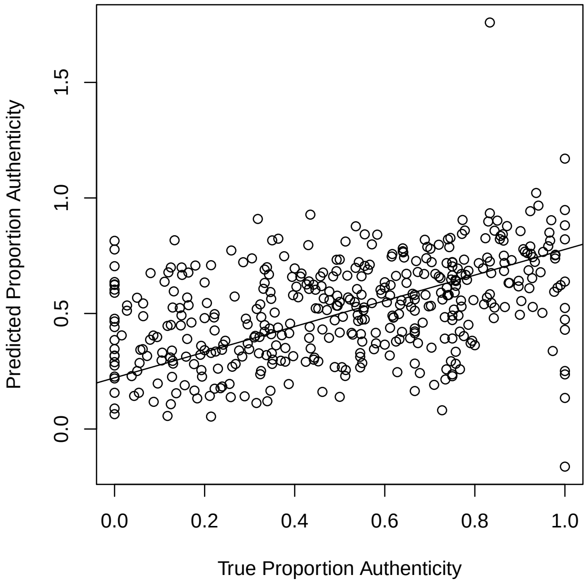

Figure 3: M5P session-level proportion predictions on the Partnership data set.

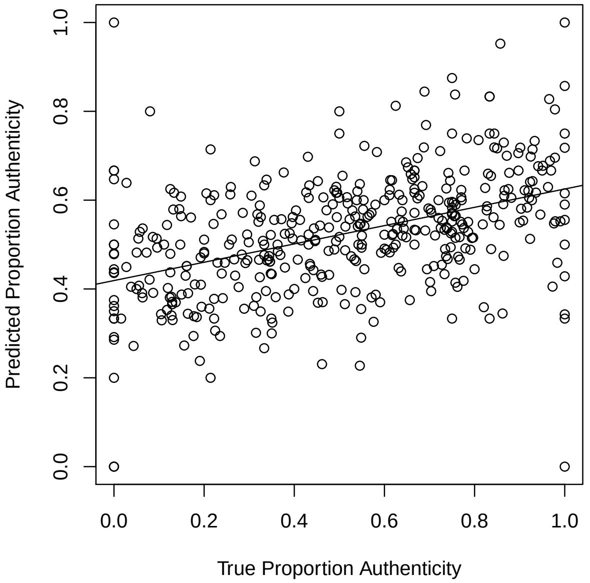

Figure 4: J48 utterance-level predictions aggregated to session-level proportions on the Partnership data set.

illustrate the similarities of these predictions. The equivalence between utterance- and session-level models for the Partnership data (shown in in Figures 3 and 4) and lack of equivalence between utterance- and session-level models for the new CLASS 5 data (shown in Figures 1 and 2) serves to further illustrate the enhancement to predictive stability that comes from using session-level models for this task. When the classes are relatively balanced, as in the case of the Partnership data, there is no difference between aggregating utterance-level predictions and session-level predictions. However, when the classes are imbalanced, as in the case of the new CLASS 5 data, the differences are significant and favor the session-level model.

4. DISCUSSION

We have presented and validated a method for assessing classroom instructional quality based on authentic questions that is effective even when such questions are rare. Our approach transforms the problem of utterance-level authentic question classification into the problem of session-level regression predicting the proportion of authentic questions. This problem transformation outperforms aggregating utterance-level classifications when classes are imbalanced, is stable for both low and high dialogic classrooms, and is stable across both automatic speech recognition and human transcripts. As such it is more appropriate for use in assessing classroom instructional quality across a wide range of dialogic discourse, complementing previous work that has investigated model generalization in different discourse communities [17]. Because question asking behavior of this type is a common component of the major classroom observation protocols in use today (e.g., those used in the MET study [11]), this research may potentially be used to help automate various protocols in addition to the target protocol here, CLASS.

Because many major classroom observation protocols call for judgments of quality approximately every 10 minutes, session-level proportion predictions are not too dissimilar from current practice. A useful point for future research would be to obtain data coded with these protocols in addition to the speech data we used, subdivide the data into 10 minute bins, and then calculate accuracy. On the other hand, the CLASS protocol is much more fine grained, and the current approach sacrifices the utterance-level resolution CLASS specifies for robustness. From a teacher professional development perspective, fine grained annotations are more useful because they can be replayed to the teacher to highlight particularly effective portions of the class. Our session-level approach in its present form appears to be less useful for professional development.

An avenue for future work would be to combine session-level and utterance-level models. For example, a session-level model could first be applied to the data, generating a session-level prediction variable, and then that variable could be used as a feature in an utterance-level model. Presumably this would be used by the model as an intercept to adjust the baseline probability of authenticity for all utterances in that session. Of course the session- and utterance-level processes could also be jointly modeled, e.g. using a hierarchical Bayesian approach.

Finally, we raise the question of why authentic questions were rarer in our new CLASS 5 data collected from 2014-2016 compared to the archival Partnership data collected from 2001-2003. The question is whether the low rate of authentic questions in our new sample is something that can reasonably be expected to reoccur, or whether it is the product of a relative small homogeneous sample. Indeed we find that some of the first studies with CLASS found levels of authenticity between 10% and 30% [14], suggesting that the rate of authentic questions in our new sample is in the normal range. The fact that rates as low as 10% have been observed serve as a warning and challenge to future research. In our new CLASS 5 data, authenticity rates of 30% for instructional questions translated to 3% of utterances being authentic. Presumably a 10% authenticity rate for instructional questions would mean that only 1% of utterances are authentic.

5. ACKNOWLEDGMENTS

This research was supported by the Institute of Education Sciences (IES; R305A130030). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of the IES. We also thank Marci Glaus, Xiaoyi Sun, and Brooke Ward of the University of Wisconsin-Madison for their help with the collection and annotation of the classroom data.

6. REFERENCES

[1] American Institutes for Research. Databases on state teacher and principal evaluation policies, 2016.

[2] J. Archer, S. Cantrell, S. L. Holtzman, J. N. Joe, C. M. Tocci, and J. Wood. *Better Feedback for Better Teaching: A Practical Guide to Improving Classroom Observations*. Jossey-Bass, 2016.

[3] M. M. Bakhtin. *The dialogic imagination: Four essays*. University of Texas Press, 1981.

[4] S. Cantrell and T. J. Kane. Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study. resreport, Bill & Melinda Gates Foundation, 2013.

[5] N. V. Chawla. Data mining for imbalanced datasets: An overview. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 853–867. Springer US, Boston, MA, 2005.

[6] S. K. D'Mello, A. M. Olney, N. Blanchard, B. Samei, X. Sun, B. Ward, and S. Kelly. Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 557–566, New York, NY, USA, 2015. ACM.

[7] E. Frank, Y. Wang, S. Inglis, G. Holmes, and I. H. Witten. Using model trees for classification. *Machine Learning*, 32(1):63–76, 1998.

[8] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2012.

[9] A. Gamoran and S. Kelly. Tracking, instruction, and unequal literacy in secondary school English. In R. Dreeben and M. T. Hallinan, editors, *Stability and change in American education: Structure, process, and outcomes*, pages 109–126. Eliot Werner Publications Incorporated, Clinton Corners, NY, 2003.

[10] A. Gamoran and M. Nystrand. Background and instructional effects on achievement in eighth-grade English and social studies. *Journal of Research on Adolescence*, 1(3):277–300, 1991.

[11] T. J. Kane and D. O. Staiger. Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. resreport, Bill & Melinda Gates Foundation, 2012.

[12] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

[13] M. Nystrand. *CLASS (Classroom language assessment system) 2.0: A Windows laptop computer system for the in-class analysis of classroom discourse.*, 1988.

[14] M. Nystrand, editor. *Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom.* Language and Literacy Series. Teachers College Press, New York, 1997.

[15] A. M. Olney, M. Louwerse, E. Mathews, J. Marineau, H. Hite-Mitchell, and A. C. Graesser. Utterance classification in AutoTutor. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing*, pages 1–8, Philadelphia, 2003. Association for Computational Linguistics.

[16] S. G. Rivkin, E. A. Hanushek, and J. F. Kain. Teachers, schools, and academic achievement. *Econometrica*, 73(2):417–458, 2005.

[17] B. Samei, A. M. Olney, S. Kelly, M. Nystrand, S. D'Mello, N. Blanchard, and A. Graesser. Modeling classroom discourse: Do models that predict dialogic instruction properties generalize across populations? In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, , and M. Desmarais, editors, *Proceedings of the 8th International Conference on Educational Data Mining*, pages 444–447. International Educational Data Mining Society, 2015.

[18] B. Samei, A. M. Olney, S. Kelly, M. Nystrand, S. D'Mello, N. Blanchard, X. Sun, M. Glaus, and A. Graesser. Domain independent assessment of dialogic properties of classroom discourse. In J. Stamper, Z. Pardos, M. Mavrikis, and B. McLaren, editors, *Proceedings of the 7th International Conference on Educational Data Mining*, pages 233–236, 2014.

[19] M. Surdeanu, T. Hicks, and M. A. Valenzuela-Escarcega. Two practical rhetorical structure theory parsers. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5, Denver, Colorado, June 2015. Association for Computational Linguistics.

[20] Q. Yang and X. Wu. 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(4):597–604, 2006.