Automatically Measuring Question Authenticity in Real-World Classrooms

Sean Kelly

University of Pittsburgh

Andrew M. Olney

University of Memphis

Patrick Donnelly

California State University-Chico

Martin Nystrand

University of Wisconsin-Madison

Sidney K. D'Mello

University of Colorado Boulder

Please direct correspondence to Sean Kelly, 5527 Posvar Hall, 230 South Bouquet Street, Pittsburgh, PA 15260. Email: spkelly@pitt.edu. This research was supported by a grant from the Institute for Education Sciences (R305A130030). Amanda Godley, Ian Wilkinson, and Adam Gamoran generously provided comments on an earlier version of this manuscript. Any opinions, findings, and conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of IES.

Abstract

Analyzing the quality of classroom talk is central to educational research and improvement efforts. In particular, the presence of authentic teacher questions, where answers are not predetermined by the teacher, helps constitute and serves as a marker of productive classroom discourse. Further, authentic questions can be cultivated to improve teaching effectiveness and consequently student achievement. Unfortunately, current methods to measure question authenticity do not scale because they rely on human observations or coding of teacher discourse. To address this challenge, we set out to use automatic speech recognition, natural language processing, and machine learning to train computers to detect authentic questions in real-world classrooms automatically. Our methods were iteratively refined using classroom audio and human-coded observational data from two sources: (1) a large archival database of text transcripts of 451 observations from 112 classrooms, and (2) a newly collected sample of 132 high-quality audio recordings from 27 classrooms, obtained under technical constraints that anticipate large-scale automated data collection and analysis. Correlations between human-coded and computer-coded authenticity at the classroom level were sufficiently high (r = .602 for archival transcripts and .687 for audio recordings) to provide a valuable complement to human coding in research efforts.

Automatically Measuring Question Authenticity in Real-world Classrooms

Teacher observation has become increasingly central to educational research and school improvement efforts (American Institutes for Research, 2016; Hamilton, 2012; Stein & Matsumura, 2009). Observations of classroom practice are valuable because they identify specific domains of practice for improvement and can target dimensions of schooling not captured by test scores. When equipped with a well-developed, nuanced observational protocol, an observer should, in principle, be able to capture "the feel" of the classroom learning environment. Is this classroom a lively, engaging, and supportive learning environment?

We report the results of research efforts to develop fully automated, computerized methods to measure an important dimension of teachers' practice—question-asking behavior. Teacher questions are an essential pedagogical element which perform many functions, from monitoring student learning to creating a space for students' interests and ideas (Wilkinson & Son, 2009). It is not an exaggeration to describe much of teachers' work as facilitating "talking to learn," which can be achieved by asking good questions (Alexander, 2008; Britton, 1969; Juzwik, Borsheim-Black, Caughlan, & Heintz, 2013; Resnick & Schantz, 2015). As such, many well-known observational protocols place emphasis on teacher questioning practices (Measures of Effective Teaching Project, 2012). Observations of teacher practice can help teachers to improve classroom talk by—referencing Clarke and Hollingsworth's (2002) model of teacher learning—making salient important discourse features while giving teachers an external source of information on their practice.

Observational tools also play an increasing role in pre-service teacher education, especially in the student-teaching phase (Wei & Pecheone, 2010). Currently, transcripts and video from pre-service teachers' own classrooms provide data for them to use in analyzing their

classroom discourse and in investigating the effectiveness of experiments to improve their discourse. Observational tools also provide structure and focus as their supervisors assess teacher learning evident in instructional samples (Caughlan, Juzwik, Borsheim-Black, Kelly, & Fine, 2013; Kersting, Givvin, Thompson, Santagata, & Stigler, 2012; Kucan, Khasnabis, & Chang, 2009; Kucan, 2007; Rosemary, Freppon, & Kinnucan-Welsch, 2002; Roskos, Boehlen, & Walker, 2000). Teachers who participate in structured analysis of their own and others' practice experience growth in pedagogical reasoning; in particular, their analyses become more evidence-based and focused on student thinking and learning (Armstrong & Curran, 2006; Sherin & Han, 2004).

Yet, observational methods of supporting pre-service and in-service teacher learning are currently logistically complex, requiring observer training and an expensive allocation of peer, staff, administrator, or supervisor time (Archer et al., 2016). Alternatively, computational linguists and computer scientists have begun to apply automated observation and coding techniques in the instructional sciences. This research strives to capitalize on the ability of computers to sort, analyze, and summarize the vast amounts of fine-grained information entailed in interpersonal communication. Many advances have been made in developing conversational intelligent tutoring systems and other language-based learning technologies (Graesser, McNamara, & VanLehn, 2005; Roscoe & McNamara, 2013; Rosé & Ferschke, 2016; Rus, D'Mello, Hu, & Graesser, 2013), though most of these systems have been tested in the lab or in online learning contexts rather than in traditional classroom environments.

In research in natural classroom settings, Miller and colleagues used the Language Environment Analysis system (LENA; Ford, Baer, Xu, Yapanel, & Gray, 2008), a wearable audio recording device, to assess when teachers were speaking, students were speaking, speech

was overlapping, or there was silence (Wang, Miller, & Cortina, 2013; Wang, Pan, Miller, & Cortina, 2014). These data on teacher-student turn-taking dynamics were used to develop automated systems to discriminate among basic patterns of instructional time use. Despite the pioneering nature of this work, LENA is prohibitively expensive (> $10k), utilizes proprietary software, is mainly suited for young child/caregiver interactions, and does not provide audio of suitable fidelity to facilitate automated speech recognition, a requisite for measuring nuanced question properties. This necessitated a new approach for automated collection and analysis of classroom discourse.

Classroom Discourse and Authentic Questions

Classroom discourse and the types of talk teachers use to promote learning can take many forms, from teacher-directed lectures, communication of procedures and directions, and question and answer recitations, to open-ended discussions, where students and teachers exchange ideas collaboratively (Alexander, 2008; Juzwik et al., 2013). Although lecturing and other teacher-directed talk serves useful pedagogical functions, much research in classroom discourse has focused on interactive forms of talk encompassing genuine discussions, deliberations, etc., which are often collectively termed *dialogic*. A key feature of high-quality teacher discourse is that it takes students' ideas seriously (Gamoran & Nystrand, 1992), and is thought to elicit both increased cognition and engagement (Kelly, 2007; Nystrand, 1997; Resnick & Schantz, 2015). Nystrand (2006) argued that discourse is an extension of teachers' underlying pedagogical approach (see also Kelly et al., 2018), such that dialogically organized instruction is uncommon among transmission-oriented teachers who view their role as providing information to students and ensuring content coverage. Empirical studies show that genuine classroom discussions are indeed quite rare (Chinn, Anderson, & Waggoner, 2001; Nystrand & Gamoran, 1997), so

perhaps even teachers who embrace student-centered/constructivist pedagogies have difficulty eliciting discussion. As a result, many efforts have been made by teacher education and professional development experts to communicate to teachers specific strategies to enhance the quality of classroom discourse and learning outcomes (Adler & Rougle, 2005; Alexander, 2008; Beach & Myers, 2001; Boyd & Galda, 2011; Burke, 2010; Caughlan et al., 2013; Rymes, 2009; Thompson, 2008).

Authentic questions, those questions for which the answers are not presupposed by the teacher (e.g., "Do you think Abigail is going to tell the truth?"; Juzwik et al., 2013, p. 27), are a core feature of dialogically organized instruction. Authentic questions serve several functions in the classroom. First, they serve to convey some authority to students, giving students input into the flow of inquiry. This provides an opportunity for students to link classroom topics with their own experiences, thereby improving engagement, coherence, and retention (Nystrand, 1997). Second, authentic questions and other discourse moves serve a normative function in the classroom, setting the expectation that students will be active participants in the learning process (Nystrand, 1997; Gamoran & Nystrand, 1992). Finally, authentic questions may also serve to provoke thought and analysis, as opposed to a simple reporting of information or facts (Kelly, 2007).

Authentic questions are related to student engagement (Kelly, 2007) and achievement growth (Gamoran & Kelly, 2003; Nystrand & Gamoran, 1997), and are central to many conceptual models of effective discourse practices. Instructional frameworks including *Quality Talk* (Wilkinson, Soter, & Murphy, 2010), *Accountable Talk* (Resnick, Michaels, and O'Connor, 2010), and *Questioning the Author* (McKeown & Beck, 2015) all emphasize authentic questions. More generally, by promoting substantive conversation, connections to the world beyond the

classroom, and higher-order thinking, authentic questions play an important role in *Authentic Pedagogy*, a set of instructional and assessment practices that balance active learning and intellectual quality and are linked with achievement growth (Newmann, Marks, & Gamoran, 1996).

Current Study: Developing a New Technology for Measuring Classroom Discourse

Nystrand and Gamoran's analyses of instructional time use and question properties provided a foundation for our present work (Gamoran & Nystrand, 1992; Nystrand & Gamoran, 1997; Gamoran & Kelly, 2003). This prior research utilized a computer-based observational system relying on expert human coders to identify discourse practices at the level of individual questions, and thus provided exceptionally precise measures of instructional practice, including teachers' use of authentic questions. We seek to reproduce the capability of Nystrand's CLASS[1] program using automated methods, and as a proof-of-concept of automated analysis of classroom discourse, present an analysis comparing measures of authentic questions using traditional human coding to an automated approach.

The development of this technology entailed iteratively developing three basic automated processes: (1) automatic speech recognition (ASR) to obtain text transcriptions from spoken audio; (2) natural language processing (NLP) to extract meaningful language features (representations) from the transcripts; and (3) machine learning to train computers how to classify the content and communicative intent of utterances. These methods are beginning to be used in educational research, including studies of text characteristics (Graesser, McNamara, & Kulikowich, 2011) and writing proficiency (Allen, Snow, & McNamara, 2016). In our work, we combine them to produce estimates of teachers' use of authentic questions in a scalable and generalizable manner.

At the outset, experienced classroom observers might be rightfully dubious of these efforts, all for good reason. First, many real-world classrooms are noisy, boisterous environments, and dialect variation and multiparty chatter remain non-trivial technical problems for ASR (Stolcke, 2011). Second, the automated approach would need to be effective in a wide array of diverse classrooms, not just addressing diverse dialects, but also providing an *unbiased* analysis of teacher practice, a concern of growing importance in artificial intelligence research (Hardt, Price, & Srebro, 2016). Third, classification of classroom discourse, where the base rates of many question properties are low (see Table 1 in Results), would seem to be especially susceptible to the challenges associated with imbalanced data, a major concern in machine learning (Yang & Xu, 2006). Fourth, visual cues, such as head nods, smiles, etc., which communicate valuable information, are not available when only speech is analyzed (as with our approach), potentially disadvantaging automated methods. Fifth, and perhaps most fundamentally, authenticity and other discourse properties are conceptualized most fully as "question events" or as being embedded within spells of interactive discourse, rather than properties of individual utterances (Nystrand, 1997). Thus, in Nystrand's coding scheme (which provided the "gold-standard" or criterion reference codes for authenticity in this study), the coding of each question is not done in isolation but is instead based on the sequence and flow of discourse around each question. Would it really be possible to automate the analysis of such fundamentally complex aspects of human interaction?

We address this question by developing methods and applying them to data from two sources, the Partnership for Literacy Study (Kelly, 2007) and the newly collected CLASS 5.0 data (see below). Together, these datasets offer complementary attributes that allow us to develop and test automated methods; and importantly, they both feature gold standard human-

coded criterion estimates of question authenticity. The archival Partnership for Literacy Study provided a large and diverse corpus of question transcripts (about 27,000 teacher questions), with pronounced variability in teachers' discourse practices (due to the study design) as well as detailed information on classroom socio-demographic and achievement composition. Thus, the Partnership data were useful for both initial development and testing of the automated methods and for testing the sensitivity of our methods to classroom context. However, due to the poor quality of Partnership archival audio recordings, we could not conduct fully automated analyses of these data; we relied instead on human transcriptions. To address this, we developed a cost-effective and scalable method to record audio of sufficient quality using newly collected data from 27 Midwestern classrooms (CLASS 5.0 sample), paving the way for a fully automated approach to authenticity measurement.

Data and Methods

Data Sources

Data for the present analyses come from two sources, the Partnership for Literacy Study and the CLASS 5.0 sample. The Partnership for Literacy Study included classroom observations and other related data collected in 7th and 8th grade English and language arts classrooms in Wisconsin and New York from 2001-2003. The Partnership data have three noteworthy features. First, the Partnership sample of 432 observations in 119 classrooms consisted of only regular or un-tracked classrooms (no honors, remedial, or otherwise tracked classes were included).[2] Second, because the Partnership entailed a professional development intervention, classroom observations revealed higher mean levels of dialogically organized discourse, and greater variability in discourse practices than is commonly found in naturalistic samples (see

Table 1). Third, the students participating in the Partnership study were of varying race/ethnicity and socioeconomic status; 36.5% of students listed a race/ethnicity other than white and about 61% of the students had at least one parent with a college degree. However, students were clustered at the school and class level especially in terms of race/ethnicity; for example, 18% of classes were 76-100% black.[3]

The CLASS 5.0 sample was comprised of 132 observations of 27 middle-school English and language arts classrooms, taught by 14 teachers in seven Midwestern schools from 2013-2016. The CLASS 5.0 sample includes data from a mostly homogenous, predominantly-white rural district in Year 1 (selected for expediency in order to begin immediate testing of alternative recording systems), and a more diverse district serving a large town in Years 2 and 3.[4] The selected classrooms represent a range of track-levels. Although prior research suggests a higher incidence rate of authentic questions and other dialogic discourse properties in high-track classrooms, these classrooms also have a more homogenous sociodemographic composition (Gamoran & Kelly, 2003).

The Gold-Standard Human Coding Approach

CLASS, a computer program where instructional processes are coded in real-time and can be subsequently revised in the laboratory using an audio or video record of the class, provided the gold-standard criterion reference codes for this study (Nystrand and Gamoran, 1997). CLASS provides a platform for coding specific attributes of teacher and student questions, including question authenticity. Two major coding processes occur during each classroom observation. First, time use in the classroom is characterized as one of 17 instructional activities (segments in CLASS's terminology), such that time summary statistics are generated

for varying instructional activities (e.g., the amount of time spent in lecture, seatwork, etc.). Second, within each question and answer segment, a common and key instructional activity, all instructional questions are recorded and coded by an expert based on both question transcripts and the audio/video recording. Procedural questions, such as asking about pages to be read as homework, often occur during procedures and directions segments and are not coded. In addition to the amount of time spent in discussion (which occurs only rarely in the typical classroom, see Nystrand & Gamoran, 1997), the specific attributes of teacher and student questions provide the primary insight into the nature of classroom discourse in each observation.[5] Reliability studies from double-coding of observations suggests that question authenticity has high inter-rater agreement for trained raters. Nystrand and Gamoran (1997) report a 78% rate of agreement for authenticity (r = .938 across observations), and in our most recent data, we find agreement on 82% of questions (r = .872 across observations).

Table 1 provides a summary of basic descriptive findings on the prevalence of several dialogic discourse features in a series of studies using the CLASS program. Across multiple studies involving more than a thousand classroom observations in all, teacher questions ranged from a low of 18% authentic to a high of 48% authentic (in the Partnership treatment group).[6, 7, 8]

The Automated Approach

Our automated approach is grounded in the linguistic theory of *speech acts* (Austin, 1962; Searle, 1975), which is particularly relevant to the study of interactive discourse. Speech act theory recognizes the dissociation between what is said (semantic meaning) and what is meant (pragmatic meaning). For example, the utterance “Can you pass the salt?” is not an information-seeking question about the listener’s capability for salt passing (in which case an

appropriate response might be “Yes”) but rather is an indirect request or directive for salt passing. Speech act theory informed later theories of question asking (Lehnert, 1978; Graesser, Person, & Huber, 1992) that analyze and categorize questions based on their effect on the listener (i.e. on how the listener responds). Determining the pragmatic intent of utterances, known as speech act classification, is a challenging problem in artificial intelligence (Jurafsky & Martin, 2000). Nevertheless, previous work has shown it is tractable for general dialogue (Stolcke et al., 2000). Olney et al. (2003) further showed that the pragmatic intent of questions is reflected in surface cues like part of speech, keywords, and their relative order. Amongst other question types, Olney et al.’s approach can identify questions that request the listener’s interpretation and judgment, making this approach prima facie suitable for measuring authentic questions.

Our work builds on the linguistic analysis of Olney et al. (2003) where question-relevant properties of language (called *features*) were deductively structured by the research team into patterns or rules. However, in contrast to this prior work, we use machine learning to extract patterns using Olney et al.’s work as an initial hypothesis. The advantage of applying machine learning is that the patterns used to measure authenticity can be automatically learned from language features rather than requiring them to be pre-specified. To avoid a purely data-driven approach, we restrict ourselves to a set of theoretically-grounded language features that span multiple aspects of discourse. This is in contrast to purely open-vocabulary approaches, common in atheoretical natural language processing, where many thousands of words and/or phrases are used as features (e.g., Joachims, 1998).

Our automated methods differ in the two datasets. In the Partnership data, our methods are more accurately “semi-automated;” we classified authenticity from the human-transcripts of

teacher questions. In the CLASS 5.0 sample, we carried out a fully automated measurement, beginning with raw audio. Here, we provide a high-level overview of our technical approach, in each case, for a broad readership. Additional details are discussed in previously published reports that we cite.

Semi-automated Authenticity Measurement from Archival (Partnership) data

Our semi-automated approach relied on human coders to: (1) segment the raw classroom audio into teacher utterances, (2) identify which utterances are questions, (3) discriminate instructional questions from non-instructional (e.g., procedural) questions; and (4) provide approximate transcriptions of the instructional questions. Then, syntactic and discourse parsing (Manning et al., 2014; Surdeanu et al., 2015) was used to compute sentence (a teacher utterance) and multi-sentence discourse features. At each level of structure (word, sentence, discourse), our approach measures various properties of language to arrive at a set of 244 features (for a full explanation of features, see Olney et al., 2017b). Example features include parts-of-speech tags (e.g., noun, verb), named entity type (e.g., PERSON; LOCATION), question stems (e.g., "what"), word position and order (e.g., whether the named entity PERSON feature occurred at the first word or the last word), and referential chains (connections across utterances via pronouns and their referents).

Why are so many features needed to detect a single discourse construct? Consider the following exchange:

Teacher: is advantage a good thing or bad thing?

Student: good thing.

Authentic version: Teacher: good thing. Ok. How would you describe advantage?

Non-authentic version: Teacher: advantage is a good thing

Taken in isolation, the teacher's initial disjunctive question does not have any strong indicators of authenticity. The teacher's second utterance for both authentic and non-authentic versions contains an equivalent reference chain connecting “good thing” and “advantage” to the first question. However, the authentic version has additional markers (i.e. “would,” “you,” and “describe”) that indicate authenticity when considered in conjunction with the first question. The machine learning method (see below) learns conditional configurations of features so that the authenticity signal of any one feature is highly contingent on the other features present, an approach that necessitates a comprehensive set of features.

In the course of developing the fully automated models (see next section), we shifted from measuring authenticity at the question level (Samei et al., 2014; Samei et al., 2015) to estimating the proportion of authentic questions in the observation (Olney et al., 2017b). Therefore, we first aggregated features (sum, mean, and standard deviation of feature counts) at the observation level (each observation consisted of an entire class period) to produce a final set of 732 (244 × 3) features. This array of features was incorporated in an M5P regression tree model (Frank et al., 1998) trained to predict the proportion of authentic questions per observation. A regression tree is a decision tree that contains regression models at the leaves; starting at the root of the tree, decisions to follow a left or right branch are based on the value of a particular feature until a leaf with the appropriate regression model is reached. The structure of the tree is automatically learned from the data using machine learning, and the combination of models under the tree is analogous to a piecewise linear function. To avoid overfitting to the data, we subdivided the data by school, an M5P model was trained on observations from k-1

schools, and was then used to generate predictions on the held-out school. The process was repeated until each school was designated the held-out school once.

Fully automated Authenticity Measurement from Classroom (CLASS 5.0) Data

The end-goal of the fully automated approach is to provide a measure of authenticity from a recording of classroom audio without any human involvement (other than wearing the recording device and submitting the recording to the computer). Whereas human auditory and language understanding systems have evolved to recognize human speech under noisy conditions, computers require relatively high-fidelity audio. Thus, the low-to-medium quality audio traditionally obtained in classroom recordings would not suffice for automated analyses; this was empirically confirmed in D'Mello et al. (2015). Thus, we experimented with several different recording designs to balance a set of technical requirements and constraints (see full discussion in D'Mello et al., 2015). For example, cameras could not be used due to privacy concerns and students could not be individually mic'd due to scalability concerns.[9]

Figure 1 provides a high-level overview of the fully automated approach. In our selected design, teachers wore a wireless Samson AirLine 77 vocal headset system, which costs $299. The mic wirelessly transmits audio to a receiver connected to a laptop, which was recorded with the open-source software program Audacity. We used this setup to record a total of 7,663 minutes of audio (i.e. the CLASS 5.0 data), comprising 132, 30-90-minute class sessions from 27 classrooms taught by 14 teachers in 7 schools over five semesters.[10]

Using a voice activity detection method we developed (D'Mello et al., 2015), we obtained a total of 45,044 teacher utterances, with a median length of 2.26 seconds. The spoken utterances were converted into text transcripts using Microsoft Bing Speech, which yielded a

speech recognition accuracy of 53.3% (Bing) when word order was considered and 61.5% when it was ignored (D'Mello et al., 2015).[11] Though imperfect, this result is encouraging given the noisy nature of classroom discourse, which includes conversational speech, multiparty chatter, background noise, and so on.

Before we could build the models, we first needed to align all detected teacher utterances with gold-standard authenticity codes, because human coders only provided codes for teacher instructional questions (a subset of all automatically detected utterances). This was done based on overlap between timestamps, a step not required for the Partnership data, which required humans to segment the audio.[12] To illustrate the difference between the two datasets:

(Partnership data) Proportion instructional questions that are authentic =

(Instructional questions $\cap$ Authentic questions) / Instructional questions;

(CLASS 5.0 data) Proportion utterances aligned with authentic questions =

(All Utterances $\cap$ Authentic questions) / All Utterances

Because the teacher utterances are often short phrases, more than one teacher utterance may align with a single coded question. Further, there were far fewer instructional questions (M = 36.53, SD = 24.14) compared to teacher utterances (341.24, 146.29), and because the data did not entail an instructional intervention, only 3.6% (SD = .034) of *utterances* were aligned with authentic questions in the CLASS 5.0 data compared to 51.9% (SD = .210) *instructional questions* deemed authentic in the Partnership data.[13]

The low incidence rate resulted in an *imbalanced classes problem*, which has been proposed as one of data mining's top 10 challenging problems (Yang & Wu, 2006). For this

reason, we shifted from measuring authenticity at the question level to estimating the proportion of authentic questions in a class session as noted above (Olney et al., 2017b).

The subsequent steps were similar to the ones applied to the Partnership data with two exceptions. First, language features were computed from speech recognition transcripts of all teacher utterances rather than manual transcripts of only instructional questions. Finally, because there were only seven schools in the CLASS 5.0 sample, the M5P models were validated using a leave-one-teacher-out approach compared to the previous leave-one-school-out method.

Results

We assess three aspects of the model performance. First, how well do the computer-estimates match the central tendency and variability in the human-codes; that is, how well do the distributions overlap? Second, at the level of individual observations, and when aggregated to the class-level, how well do the computer-estimates correspond to the human-coded estimates for each specific observation and class; are the computer-estimates reliable? Aggregating estimates to the class-level anticipates use in research or reform of class-level phenomena (e.g., curriculum tracking, instructional interventions), or where multiple observations of individual sessions are used to build a profile of teacher practice (see e.g. Measures of Effective Teaching Project, 2012). Third, how sensitive are the computer-estimates to classroom context? We present results for the much larger but only semi-automated Partnership for Literacy Study data, as well as the smaller, fully automated CLASS 5.0 data.

Semi-automated approach on the Partnership for Literacy Data

Table 2 reports statistics summarizing the correspondence between the human-coded and computer-estimates of the prevalence of authentic teacher questions at both the observation (N =

428) and class-level (N = 116). For the observation-level analysis, two sets of findings are presented: the full N = 428 sample, and with observations restricted to those with 300 seconds or more of time spent in question and answer sessions (N = 374). Here, we discuss the later findings, under the assumption that if these methods were to be employed in instructional improvement efforts, observation-level estimates of discourse properties would only be reported when there is enough Q&A discourse to generate reliable estimates beyond noting that little time was spent in interactive discourse. The class-level analysis includes all observations, frequency weighted by the number of teacher instructional questions with a student response.

Figure 2 provides a graphical representation of distribution overlap and correspondence. At the observation-level, the computer estimates provided a central tendency for authentic questions that closely approximates the mean and median of the human codes (an Adjusted Wald test, correcting for clustering, revealed a nonsignificant ($p > .624$) difference in means between the human and computer codes – see Table 2). However, the computer estimates are more heavily clustered and peaky about the mean, and thus more normally distributed (kurtosis = 2.44 compared to 3.0 for a normal distribution), with a somewhat lower standard deviation than the human-codes (SD of .205 vs. .276). On average, the difference (absolute value) between the computer- and human-coded proportion of authentic teacher questions was .194. Although this discrepancy may seem quite large, it corresponds to a correlation of .506. Further, at the observation level, we performed a decomposition of variance (regression of the absolute deviation between human and computer codes on the set of class ids) to identify the possible contribution of class/teacher level influences on measurement. The R^2 from this model was .313, suggesting that as an upper-bound estimate; about 30% of the variance in model performance may be due to contextual features of the classroom.

At the class-level, the distribution of the computer-estimates of the prevalence of authentic teacher questions more closely matched the human-coded estimates, but with a slightly higher mean value and reduced variation (SD of .142 vs. .218). The difference (absolute value) between the two estimates fell to .141, with the correlation increasing to .602. To identify possible effects of classroom and teacher context, we regressed the observed measurement discrepancy (mean absolute deviation) on a set of eight class/teacher covariates: teacher gender, teacher race/ethnicity, teacher years' experience, class proportion Hispanic, class proportion Black, class-mean socioeconomic status, class-mean fall achievement, and a variable capturing the Partnership Study treatment effects (See Kelly, 2007 for further description of variables). The effect of measured context was quite modest in these data ($R^2 = .145$).

Fully automated Approach on CLASS 5.0 Data

Informed by the performance of the semi-automated approach on the Partnership data transcripts, we were initially concerned that ASR errors might make fully automated authenticity measurement in the CLASS 5.0 sample substantially less precise. However, the semi-automated measure was limited by the quality of the original Partnership question transcripts, which were never intended for this purpose. Rather, they were often shorthand transcriptions; because all of the human coding was done with video/audio data running, there was no need for high-quality transcripts. Thus, the measures of correspondence reported in Table 2 likely underestimate the potential of our approach. Further, although ASR error rates are non-trivial, we found in supplementary analyses that the decline in information due to ASR errors has a very modest effect on performance (Olney et al., 2017a, p. 264). Overall, there was reason to be optimistic about the potential performance of the fully automated codes.

Table 3 compares the observed and fully automated estimates of the prevalence of authentic teacher questions in the CLASS 5.0 sample using specifications tailored to automated methods. For example, the automated methods include utterances occurring in discussion and other segments not captured in the human-codes. As before, we report results for the full sample (N = 132) and for a restricted sample (n = 113) obtained using a 0.20 probability threshold to omit sessions with a low probability of Q&A—also automatically estimated (see Donnelly et al., 2016). This yields a 14.4% reduction in cases, comparable to the 12.6% reduction in the Partnership.

The CLASS 5.0 sample data does differ from the Partnership in having a peaky mass of values close to zero, with a skewness of 1.64 and kurtosis of 4.61 in the human codes. However, the central tendency and dispersion of the computer-estimates match the human codes quite well for both levels of analyses (Adjusted Wald tests for the difference in means between the human and computer codes yielded p values > .127 – see Table 3). Importantly, the computer estimates in the CLASS 5.0 data, which are based on ASR transcripts, are even more precise than those in the Partnership data (based on human transcripts), with a correlation of .671 at the observation level, and .687 at the class level.

Unlike the Partnership data, the precision of the fully automated estimates is clearly influenced by the specific classroom/teacher being analyzed (R^2 of .580 in the decomposition of variance). In the Partnership data, the effects of classes/teachers was restricted to contextual effects on syntactic features of words and sentences. In contrast, in the CLASS 5.0 data, in addition to these contextual effects, a whole host of speech-related features (e.g., background noise level, prevalence of interruptions, rate of speech, etc.) contribute to class id effects, some of which are not readily overcome. Relatedly, in the CLASS 5.0 sample, we see a more limited

improvement in precision when the observation level data are aggregated to the class level, compared to the improvement for the Partnership data. This feature might be explained in part by the fact that more of the variance in authenticity occurs at the between-class level (in the gold-standard as well as computer codes) in the CLASS 5.0 sample than in the Partnership sample.[14] Beyond that, the limited improvement at the class level may be evidence of the substantial impact of class/teacher level speech-related features that deserve attention in future research.

Discussion

This study combined speech recognition, natural language processing, and machine learning to measure the prevalence of one of the most important discourse features associated with a dialogic stance and constructivist teaching practice—question authenticity. Classroom discourse features are important not only for their immediate effect within a lesson or on the student who responds, but because they may even shape the overall learning environment by creating new norms of interaction that provoke thought and analysis and reduce the risk of negative evaluation (Morine-Dershimer, 1985; Kelly, 2007; Turner et al., 2002). A focus on question authenticity, in particular, is part of a long-standing emphasis in educational research and improvement on thought-provoking classroom discourse (Groisser, 1964). While the large-scale studies of Nystrand and Gamoran were conducted primarily in English and language arts classrooms, authentic questions and related discourse practices are theorized to be valuable in the sciences and other subject areas (Greenleaf et al., 2011; Osborne, Simon, Christodoulou, Howell-Richardson, & Richardson, 2013). As a result, the authentic question construct has been incorporated into widely used teacher observation protocols. For example, the Danielson Framework for Teaching includes “teacher uses open-ended questions, inviting students to think

and/or offer multiple possible answers" as an indicator of proficient questioning and discussion techniques (Danielson, 2011).

Yet, automating the measurement of classroom discourse is an uncertain endeavor given the many challenges of measuring discourse in real-world classrooms: noise and chatter, dialect diversity, data imbalance, and the fundamental complexity of the construct itself. Were we able to overcome these challenges in the present study?

The noise and chatter encountered in these data from middle schools had a relatively modest impact on automated measurement, ostensibly due to the high-fidelity mic used and also because when data is collected over an entire class period, and a multitude of features are used in model-building, the speech recognition errors likely only had a small effect on model performance. Although continued research is needed, the Partnership data included much dialect diversity, and the automated measurement was not highly sensitive to this variation. We did find model sensitivity to the overall variability in discourse features in the CLASS 5.0 sample, but we believe these features can be identified and accounted for in subsequent research. Although question authenticity is certainly a complex discourse phenomenon, by drawing on a rich set of information from sequenced utterances, our final best efforts at a fully automated coding process yielded a reliability ($r = .687$) sufficient, we believe, to provide a valuable complement to human coding in large-scale research efforts. With additional research, the automated approach we applied to authenticity can be readily extended to related aspects of classroom discourse (e.g., cognitive demand of questions).

However, imbalanced data clearly remain a challenge. In this project, this problem was so severe that it meant that authenticity could not be detected at the utterance level and automated methods for many discourse constructs (e.g., uptake), simply could not be fully

validated. We anticipate we can overcome this problem in further research, identifying models that produce reliable estimates across sufficiently aggregated data (not necessarily at the observation level). When applied to observations of individual class sessions, for example, to provide teacher feedback, automated systems must incorporate the inherent error caused by imbalanced data; teachers should not be given unreliable performance feedback on low-incidence rate measures. In other words, automated measurement that is sufficient for research purposes (e.g., comparing instructional practices across multiple teacher education curricula, etc.), may not be suitable for providing class-session-level feedback in certain use cases (e.g., teacher feedback). Relatedly, these methods should never be utilized for accountability purposes.

In interpreting the achieved reliability of .687, and our conclusion that this holds much promise for further development, it is useful to compare our results with contemporary observational methods in widespread use. Because we produce continuous, interval-scale estimates in contrast to the qualitative ordinal-scale (e.g., basic, proficient, etc.) estimates generated with other protocols, we cannot directly compare reliabilities using correlation coefficients. Nevertheless, rough calculations suggest our automated method compares favorably. For example, in the MET study (ICPSR, 2014), inter-rater reliabilities of .09, .14, and .20 (in the kappa metric using simple, linear weighted, and quadratic weighting respectively) were reported for the Framework for Teaching's Using Questioning and Discussion Techniques domain.[15] In the CLASS 5.0 sample, the kappa agreement between the computer and human codes at the observational level was .13, .40, and .54 (for simple, linear, and quadratic weighted kappas) when the continuous proportions were collapsed to an ordinal 1–9 scale.

Beyond our particular study, much has been learned about teacher effectiveness and instructional practice from observational research since the major process-product studies of

teaching in the 1960s and 1970s (Brophy & Good, 1986). Insights on instruction have been incorporated into educational reforms from teacher education, to comprehensive school reforms, to efforts to standardize and coordinate curriculum at the state level and beyond. Yet, the labor-intensive nature of classroom observation has meant that the bulk of teachers' work is either examined unsystematically or infrequently. Indeed, some policy approaches favor an emphasis on effective teachers (as identified by test score gains), rather than effective teaching, sidestepping the difficult process of measuring classroom practice (Gamoran, 2012). Research that pairs established theories of teaching and learning with technological innovations of the kind used in this study may soon lead to a new era in research and school improvement with a renewed emphasis on measuring classroom instruction.

NOTES

[1] Nystrand's CLASS is a research tool, not to be confused with the trademarked CLASS program developed by Robert Pianta and colleagues.

[2] The Partnership for Literacy Statistics in Table 1 are calculated on the maximum sample size with available observational data from the original study, the analyses in Table 2 are based on slightly fewer observations/classes with available audio data.

[3] Kelly (2008) provides further detail on the Partnership data.

[4] In Year 3, when student background surveys were administered, approximately 15% of students reported American-Indian ethnicity (consistent with a substantial tribal population in the area), 19% were black, 84% were white, and 3% selected Asian/American or other race categories (following federal OMB guidelines, race/ethnic categories are not mutually exclusive). Forty percent of students identified as Hispanic, with 23% reporting speaking a language other than English at home. There was substantial diversity in family socioeconomic background as well; 25% of students reported parents' highest educational attainment as a high school diploma or less, 23% reported some college, while 52% reported a college degree or higher.

[5] Further, questions that occur during discussion segments, defined as a free exchange of information among students and/or between at least three participants that last longer than 30 seconds, are not recorded. The decision to not record discussion questions was made in order to streamline the data collection; since the quantity of time spent in discussion is itself a central indicator of dialogically organized instruction, there was no need to additionally code that discourse. However, to support the development of automated methods, future research should code instructional questions during discussions.

[6]The instantiation of these constructs in Table 1 are specific to the Nystrand framework and may differ from use in other research. In the Nystrand framework, uptake entails the incorporation of a previous answer into a subsequent question, which goes beyond mere revoicing such that the [student's] response [is allowed to] redirects the flow of questioning. High cognitive level is a binary distinction between low-level questions requiring only a reporting/recitation of facts (e.g., what happened), and high-level questions requiring a generalization or analysis (e.g., comparisons, analyses, predictions, and explanations, etc.). In the most recent data (CLASS 5.0 sample), only about 11% of teacher questions were of high cognitive level or involved uptake, an average of about 3.7 per observation. Given these low prevalence rates, the present study focused only on automating the measurement of authentic questions.

[7] In the Partnership data, the interquartile range for question authenticity was 5.5% to 42.4% for the control group and 26.2% to 67.7% for the treatment group at the observation level. At the class level, the corresponding IQRs were 12.2% to 38.5% and

31.2% to 59.9% respectively. In the CLASS 5.0 data, question authenticity had an IQR of 18.7% to 52.5% at the observation level and 21.3% to 51.2% at the class level.

[8] Of interest to observational research designed to characterize teachers' long-run practice, the studies reported in Table 1 generally collected four observations per teacher. Ancillary analyses conducted during this study of the reliability (estimated by the Spearman-Brown prediction formula in STATAs 'loneway' command) of teacher-mean discourse practices show that indeed four observations is approximately a point of diminishing returns. For example, the reliability of the proportion of authentic teacher questions estimated with four teacher observations is approximately the same as that with three, while the reliability of uptake and cognitive level estimates improve by a mere 13% and 6% respectively with the fourth observation.

[9] We also experimented with microphone arrays and sound source localization using Microsoft Kinect devices, but this approach did not scale to a large number of students in a classroom.

[10] We also recorded general classroom audio using an additional microphone, but these data are not analyzed here. See D'Mello et al. (2015) for details.

[11] We experimented with both Google and Bing ASR, but only report results for authenticity measurement for Bing because it was superior to Google (despite having slightly higher ASR errors)

[12] As with the Partnership data, supervised learning approaches require labeled data in the form of language features aligned with "gold-standard" authenticity codes for instructional questions. This alignment requires filtering instructional questions from all other teacher utterances, which was done by the human coders for the Partnership data, a luxury not available to the fully automated approach. As a result, the data contained speech that would otherwise be excluded from CLASS coding (e.g., statements, procedural questions). Thus, rather than identifying instructional questions, we created our labeled dataset by aligning all teacher utterances with authenticity codes.

[13] Observation-level means and standard deviations (variability between observations) reported. N = 428 audio-available sample from Table 2 for Partnership reported.

[14] Importantly though, this increased variance does not correspond to an increase in the range of discourse practices; the variance in the CLASS 5.0 sample was driven by differences between very low prevalence rates in some teacher's classrooms and average prevalence rates in others, whereas the Partnership data included instructional observations with high incidence rates of authenticity.

[15] These values are specific to the rating process used in the MET study, and should not be assumed to apply to more general use of the FFT.

References

Adler, M., & Rougle, E. (2005). *Building literacy through classroom discussion: Research-based strategies for developing critical readers and thoughtful writers in middle school*. New York: Scholastic.

Alexander, R. (2008). *Towards dialogic teaching: Rethinking classroom talk* (4th edition). York, UK: Dialogos.

Allen, L., Snow, E., & McNamara, D. (2016). The narrative waltz: The role of flexibility on writing performance. *Journal of Educational Psychology, 108,* 911–924.

American Institutes for Research. (2016). *Databases on state teacher and principal evaluation policies.* Retrieved 2016-12-27, from http://resource.tqsource.org/stateevaldb/Compare50States.aspx

Archer, J., Cantrell, S., Holtzman, S. L., Joe, J. N., Tocci, C. M., & Wood, J. (2016). *Better feedback for better teaching: A practical guide to improving classroom observations*. Jossey-Bass. Retrieved 2016-12-27, from http://k12education.gatesfoundation.org/wp-content/uploads/2016/05/BetterF

Armstrong, V., & Curran, S. (2006). Developing a collaborative model of research using digital video. *Computers & Education, 46,* 336–337.

Austin, J. L. (1962). *How to do things with words*. Oxford: Oxford University Press.

Beach, R., & Myers, J. (2001). *Inquiry-based English instruction: Engaging students in life and literature*. New York: Teachers College Press.

Boyd, M., & Galda, L. (2011). *Real talk in elementary classrooms: Effective oral language practice*. New York: Guilford.

Britton, J. (1969). Talking to learn. In D. Barnes, J. Britton, & H. Rosen (Eds.), *Language, the learner, and the school* (pp. 79–115). Harmondsworth, UK: Penguin.

Brophy, J., & Good, T. (1986). Teacher behavior and student achievement. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 328–375). New York: Macmillan.

Burke, J. (2010). *What's the big idea? Question-driven units to motivate reading, writing, and thinking*. Portsmouth, NH: Heinemann.

Caughlan, S., Juzwik, M. M., Borsheim-Black, C., Kelly, S., & Fine, J. (2013). English teacher candidates developing dialogically organized instructional practices. *Research in the Teaching of English, 47,* 212–246.

Chinn, C. A., Anderson, R. C., & Waggoner, M. A. (2001). Patterns of discourse in two kinds of literature discussion. *Reading Research Quarterly, 36,* 378–411.

Clarke, D., & Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teaching and Teacher Education, 18*, 947–967.

Danielson, C. (2011). *The Framework for Teaching evaluation instrument* (1st edition). Princeton, NJ: The Danielson Group.

D'Mello, S. K., Graesser, A. C., & King, B. (2010). Toward spoken human-computer tutorial dialogues. *Human Computer Interaction, 25*(4), 289–323.

D'Mello, S., Olney, A. M., & Person, N. (2010). Mining collaborative patterns in tutorial dialogues. *Journal of Educational Data Mining*, *2*, 1–37.

D'Mello, S. K., Olney, A. M., Blanchard, N., Samei, B., Sun, X., Ward, B., & Kelly, S. (2015). Multimodal capture of teacher-student interactions for automated dialogic analysis in live classrooms. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI 2015)* (pp. 557–566). New York: ACM.

Donnelly, P., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., Kelly, S., Nystrand, M., & D'Mello S. K. (2016). Automatic teacher modeling from live classroom audio. In L. Aroyo, S. K. D'Mello, J., Vassileva, & J. Blustein (Eds.) *Proceedings of the 24th ACM International Conference on User Modeling, Adaptation, and Personalization* (UMAP 2016) (pp. 45–53). New York: ACM.

Follmer, D. J., Sperling, R. A., & Suen, H. K. (2017). The role of MTurk in education research: Advantages, issues, and future directions. *Educational Researcher*, *46*, 329–334.

Ford, M., Baer, C., Xu, D., Yapanel, U., & Gray, S. (2008). The LENA language environment analysis system. Boulder, CO: LENA Foundation Technical Report LTR-03-02.

Frank, E., Wang, Y., Inglis, S., Holmes, G., & Witten, I. H. (1998). Using model trees for classification. *Machine Learning, 32*(1), 63–76. https://doi.org/10.1023/A:1007421302149

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). 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.

Gamoran, A. (2012). Improving teacher quality: Incentives are not enough. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 201–214). New York: Teachers College Press.

Gamoran, A., & Kelly, S. (2003). Tracking, instruction, and unequal literacy in secondary school English. In M. T. Hallinan, A. Gamoran, W. Kubitschek, & T. Loveless (Eds.), *Stability and change in American education: Structure, process, and outcomes* (pp. 109–126).

Clinton Corners, NY: Eliot Werner Publications Incorporated.

Gamoran, A., & Nystrand, M. (1992). Taking students seriously. In F. M. Newmann (Ed.), *Student engagement and achievement in american secondary schools* (pp. 40–61). New York: Teachers College Press.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix Providing Multilevel Analyses of Text Characteristics. *Educational Researcher, 40*(5), 223–234.

Graesser, A. C., McNamara, D., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist, 40*(4), 225–234.

Graesser, A. C., Person, N. K., & Huber, J. D. (1992). Mechanisms that generate questions. In T. E. Lauer, E. Peacock, & A. C. Graesser (Eds.), *Questions and information systems* (pp. 167–187). Hillsdale, NJ: Lawrence Erlbaum Associates.

Graesser, A. C., & Person, N. (1994). Question asking during tutoring. *American Educational Research Journal, 31*, 104–137.

Greenleaf, C., Litman, C., Hanson. T. L., Rosen, R., Boscardin, C. K., Herman, J.,….Jones, B. (2011). Integrating literacy and science in biology: Teaching and learning impacts of reading apprenticeship professional development. *American Educational Research Journal, 48*, 647–717.

Grossier, P. (1964). *How to use the fine art of questioning*. New York: Teachers' Practical Press.

Hamilton, L. (2012). Measuring teaching quality using student achievement tests: Lessons from educators' response to No Child Left Behind. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 49–76). New York, NY: Teachers College Press.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323). NIPS.

ISCPSR. (2014). *Measures of effective teaching: Observation measures report*. Ann Arbor, MI: Inter-University Consortium for Political and Social Research. (ICPSR 34771).

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In C. Nédellec & C. Rouveirol (Eds.), *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings* (pp. 137–142). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0026683

Jurafsky, D., & Martin, J. H. (2000). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.

Juzwik, M. M., Borsheim-Black, C., Caughlan, S., & Heintz, A. (2013). *Inspiring dialogue: Talking to learn in the English classroom*. New York: Teachers College Press.

Kelly, S. (2007). Classroom discourse and the distribution of student engagement. *Social Psychology of Education, 10*(3), 331–352.

Kelly, S. (2008). Race, social class, and student engagement in middle school English classrooms. *Social Science Research, 37*, 434–448.

Kelly, S., Zhang, Y., Northrop, L., VanDerHeide, J., Dunn., M., & Caughlan, S. (2018). English and language arts teachers' perspectives on schooling: Initial exposure to a teacher education curriculum. *Teacher Education Quarterly, 45*, 57–85.

Kelly, S., & Caughlan, S. (2011). The Hollywood teachers' perspective on authority. *Pedagogies, 6*, 46–65.

Kersting, N. B., Givvin, K. B., Thompson, B. J., Santagata, R., & Stigler, J. W. (2012). Measuring usable knowledge: Teachers' analyses of mathematics classroom videos predict teaching quality and student learning. *American Educational Research Journal, 49*, 568–589.

Kucan, L. (2007). Insights from teachers who analyzed transcripts of their own classrom discussions. *The Reading Teacher, 61*, 228–236.

Kucan, L., Palincsar, A. S., Khasnabis, D., & Chang, C-I. (2009). The video-viewing task: A source of information for assessing and addressing teacher understanding of text-based discussion. *Teaching and Teacher Education, 25*, 415–423.

Lehnert, W. (1978). *The Process of Question Answering*. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

McKeown, M. G., & Beck, I. L. (2015). Effective classroom talk *is* reading comprehension instruction. In L. B. Resnik, C. S. C. Asterhan, & S. N .Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 51–62).Washington, DC: American Educational Research Association.

Measures of Effective Teaching Project. (2014). *Observation measures report*. ICPSR 34771.

Measures of Effective Teaching Project. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill and Melinda Gates Foundation.

Morine-Dershimer,G. (1985). *Talking, listening and learning in elementary classrooms*. NewYork: Longman

Murphy, P. K., Wilkinson, I. A., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' high-level comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*, 740–764.

Newmann, F. M., Marks, H. M., & Gamoran, A. (1996). Authentic pedagogy and student performance. *American Journal of Education, 104*, 280–312

Nystrand, M. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom.* New York, NY: Teachers College Press.

Nystrand, M. (2006). Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English, 40*, 392–412.

Nystrand, M., & Gamoran, A. (1997). The big picture: Language and learning in hundreds of English lessons. In M. Nystrand, *Opening dialogue: Understanding the dynamics of language and learning in the English classroom* (pp. 30–74). New York: Teachers College Press.

Olney, A. M., Louwerse, M., Mathews, E., Marineau, J., Hite-Mitchell, H., & Graesser, A. C. (2003). Utterance Classification in AutoTutor. In *Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing* (pp. 1–8). Philadelphia: Association for Computational Linguistics.

Olney, A. M., Graesser, A. C., & Person, N. K. (2010). Tutorial dialog in natural language. In R. Nkambou, J. Bourdeau, & R. Mizoguchi (Eds.), *Advances in Intelligent Tutoring Systems, Studies in Computational Intelligence* (Vol. 308, pp. 181–206). Berlin: Springer-Verlag.

Olney, A. M., Kelly, S., Samei, B., Donnelly, P., & D'Mello, S. K. (2017a). Assessing teacher questions in classrooms. In R. Sottilaire, A. Graesser, X. Hu, & G. Goodwin (Eds.), *Design recommendations for intelligent tutoring systems: Assessment (Volume 5)* (pp. 261–274). Orlando, FL: U.S. Army Research Laboratory.

Olney, A. M., Samei, B., Donnelly, P. J., & D'Mello, S. K. (2017b). Assessing the dialogic properties of classroom discourse: Proportion models for imbalanced classes. In X. Hu, T. Barnes, A. Hershkovitz, & L. Paquette (Eds.), Proceedings of the 10th International Conference on Educational Data Mining (pp. 162–167).

Osbourne, J., Simon, S., Christodoulou, A., Howell-Richardson, C., & Richardson, K. (2013). Learning to argue: A study of four schools and their attempts to develop the use of argumentation as a common instructional practice and its impact on students. *Journal of Research in Science Teaching, 50,* 315–347.

Resnick, L., Michaels, S., & O'Connor, C. (2010). How (well structured) talk builds the mind. In D. Preiss, & R. J. Sternberg (Eds.), *Innovations in educational psychology, Perspectives on learning, teaching, and human development* (pp. 163–194). New York: Springer.

Resnick, L. B., & Schantz, F. (2015). Talking to learn: The promise and challenge of dialogic teaching. In L. B. Resenick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 441–450). Washington, DC: American Educational Research Association.

Roscoe, R. D., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*(4), 1010–1025.

Rosemary, C. A., Freppon, P., & Kinnucan-Welsch, K. (2002). Improving literacy teaching through structured collaborative inquiry in classrom and university clinical settings. In D. Schallert, D., C. M. Fairbanks, J. Worthy, B. Maloch, & J. V. Hoffman, (Eds.), *51st Yearbook of the National Reading Conference* (pp. 368–382). Oak Creek, WI: National Reading Conference.

Rosé, C. P., & Ferschke, O. (2016). Technology support for discussion based learning: From computer supported collaborative learning to the future of massive open online courses. *International Journal of Artificial Intelligence In Education, 26*(2), 660–678.

Roskos, K., Boehlen, S., & Walker, B. J. (2000). Learning the art of instructional conversation: The influence of self-assessment on teachers' instructional discourse in a reading clinic. *The Elementary School Journal, 100,* 229–252.

Rus, V., D'Mello, S. K., Hu, X., & Graesser, A. (2013). Recent advances in intelligent tutoring systems with conversational dialogue. *AI Magazine, 34*(3), 42–54.

Rymes, B. (2009). *Classroom discourse analysis: A tool for critical reflection*. New York: Hampton Press.

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

Samei, B., Olney, A., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., & Graesser, A. (2015). Modeling classroom discourse: Do models that predict dialogic instruction properties generalize across populations? In C. Romero, M. Pechenizkiy, J. Boticario & O. Santos (Eds.), *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)* (pp. 444–447). International Educational Data Mining Society.

Searle, J. (1969). *Speech acts*. Cambridge, UK: Cambridge University Press.

Sherin, M. G., & Han, S. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education, 20*, 163–183.

Stein, M. K. & Matsumura, L. C., (2009). Measuring instruction for teacher learning. In D.H. Gitomer (Ed.), *Measurement issues and assessment for teacher quality* (pp. 179–205). Thousand Oaks, CA: Sage Publications.

Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., … Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, *26*(3), 339–373. http://dx.doi.org/10.1162/089120100561737

Stolcke, A. (2011). Making the most from multiple microphones in meeting recognition. 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 4992–4995).

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

Thompson, P. (2008). Learning through extended talk. *Language and Education, 22*, 241–256.

Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology*, *94*, 88–106.

Wang, Z., Miller, K., & Cortina, K. (2013). Using the LENA in Teacher Training: Promoting Student Involement through automated feedback. *Unterrichtswissenschaft, 4*, 290–305.

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Automatic classification of activities in classroom discourse. *Computers & Education, 78*(1), 115–123.

Wei, R. C., & Pecheone, R. L. (2010). Assessment for learning in preservice teacher education: Performance-based assessments. In M. Kennedy (Ed.), *Teacher assessment and the quest for teacher quality* (pp. 691–32). San Francisco, CA: Jossey-Bass.

Wilkinson, I. A. G., & Son, E. H. (2009). Questioning. In E.M. Anderman and L.H. Anderman (Eds.), *Psychology of classroom learning: An encyclopedia* (pp. 723–728). Detroit, MI: Gale/Cengage.

Wilkinson, I. A. G., Soter, A. O., & Murphy, P. K. (2010). Developing a model of Quality Talk about literary text. In M. G. McKeown & L. Kucan (Eds.), *Bringing reading research to life* (pp. 142–169). New York: Guilford.

Yang, Q., & Wu, X. (2006). Challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, *5(4)*, 597–604.

Table 1. Question property statistics from four classroom observation studies employing Nystrand's CLASS program. Revision of table first reported in Kelly and Caughlan (2011).

	Opening Dialogue Sample[a]	CELA's National Study[b]	CELA's Partnership for Literacy Study[c]		CLASS 5.0 Study
			Control Group	Treatment Group	
Sample: # of observations (classrooms)	451 (112)	287 (72)	88 (26)	344 (93)	132 (27)
Sample Characteristics[d]	8th & 9th grade English, 1987-1989	7th, 8th, 10th, and 11th grade English, 1999-2001	7th & 8th grade English, 2001-2003		6-8th grade English, 2013-2016
Question Properties[e]					
Asked by Teacher	92%	84%	92%	90%	94%
Authentic Teacher Questions	**18%**	**25%**	**27%**	**48%**	**35%**
Teacher Questions with Uptake[f]	18%	32%	7%	19%	7%
High Cognitive Level Teacher Questions	41%	18%	10%	26%	4%
Teacher Test Questions[g]	Not reported	32%	40%	18%	64%

[a] Results are a combined average for Grade 8 and Grade 9, weighted by number of coded questions. See Tables 2.5 and 2.6 in Nystrand and Gamoran (1997)

[b] Statistics reported here are from the raw data, and depart slightly from those reported in Gamoran and Kelly (2003), which used an analytic sample of 64 of the classes.

[c] See Kelly (2007) for further details on the Partnership for Literacy Study.

[d] The *Opening Dialogue* study took place in eight Midwestern communities. CELA's National Study is also called the "5 State Study" because it took place in California, Florida, New York, Texas, and Wisconsin, and in some reports, "The Validation Study." The Partnership for Literacy Study took place in Wisconsin and New York, and included only regular-track or untracked classrooms; all other studies used an approximately balanced sample of low-, middle-, and high-track classrooms. The CLASS 5.0 Study took place in the Midwest.

[e] In the Partnership for Literacy Study and the CLASS 5.0 Study, question properties are calculated for teacher questions using observation level data. In the National Study, the data was collected to include student questions for uptake and high cognitive level, but only teacher questions for authenticity. In the *Opening Dialogue* sample, only authentic questions are explicitly labeled as including only teacher questions, and the text is not entirely clear on whether the proportions included student questions, but student questions constitute a very small percentage of all recorded questions.

[f] Beginning with the Partnership study, a more restrictive definition of uptake was used ("authentic uptake").

[g] Test questions are those without uptake or authenticity and are of low cognitive level, i.e. teacher questions asking for a report of information.

Table 2: Correspondence between observed (human coded) and semi-automated estimates of the prevalence of authentic teacher questions in the Partnership for Literacy Study data.

	Observation-level (N = 428)	**Observation-level, restricted (N = 374)[a]**	**Class-level (N = 116)[b]**
Measures of central tendency and dispersion			
Mean (SD)			
Human-coded	.511 (.288)	.520 (.276)	.505 (.218)
Computer-estimated	.519 (.210)	.519 (.205)	.514 (.142)
Adjusted Wald test, diff != 0	p < .624	p < .940	p < .710
Median			
Human-coded	.536	.544	.511
Computer-estimated	.530	.530	.507
Skewness			
Human-coded	−.135	−.161	−.035
Computer-estimated	−.115	−.091	.175
Kurtosis			
Human-coded	1.94	1.94	2.46
Computer-estimated	2.44[c]	2.28	2.52
Measures of correspondence			
Mean |Diff|	.213	.194	.141
IQR of |Diff|	.081, .297	.076, .278	.066, .196
Correlation Coeff	.424	.506	.602
Effect of class/teacher variables on correspondence			
R^2, from class ids model[d]	.258	.313	
R^2 from covariate model			.145

[a] Analysis restricted to observations with 300 seconds or more of time spent in question and answer segments.

[b] All observations, frequency-weighted (by # of teacher instructional questions with response) class means.

[c] Standardized kurtosis; normal distribution has kurtosis of 3.0.

[d] Output from STATAs 'loneway' command.

Table 3: Correspondence between observed (human coded) and fully automated estimates of the prevalence of authentic teacher questions in the CLASS 5.0 sample.

	Observation-level (N = 132)	**Observation-level, restricted (N = 113)[a]**	**Class-level (N = 27)[b]**
Measures of central tendency and dispersion			
Mean (SD)			
Human-coded	.042 (.051)	.047 (.053)	.039 (.034)
Computer-estimated	.036 (.034)	.037 (.031)	.036 (.025)
Adjusted Wald test, diff != 0	p < .334	p < .127	p < .713
Median			
Human-coded	.023	.029	.029
Computer-estimated	.026	.027	.031
Skewness			
Human-coded	1.77	1.64	1.13
Computer-estimated	1.59	1.33	1.04
Kurtosis			
Human-coded	6.08	5.58	3.21
Computer-estimated	5.72	4.61	3.14
Measures of correspondence			
Mean |Diff|	.027	.028	.017
IQR of |Diff|	.007, .034	.008, .036	.006, .021
Correlation Coeff	.613	.671	.687
Effect of class/teacher variables on correspondence			
R^2, from class ids model[c]	.568	.580	
R^2 from covariate model			-

[a] Analysis restricted to observations where automated estimates of the probability of question & answer segments exceed 0.2.
[b] All observations, frequency-weighted by class mean number of utterances automatically estimated by the computer.
[c] Output from STATAs 'loneway' command.

Figure 1. Illustration of the fully automated approach

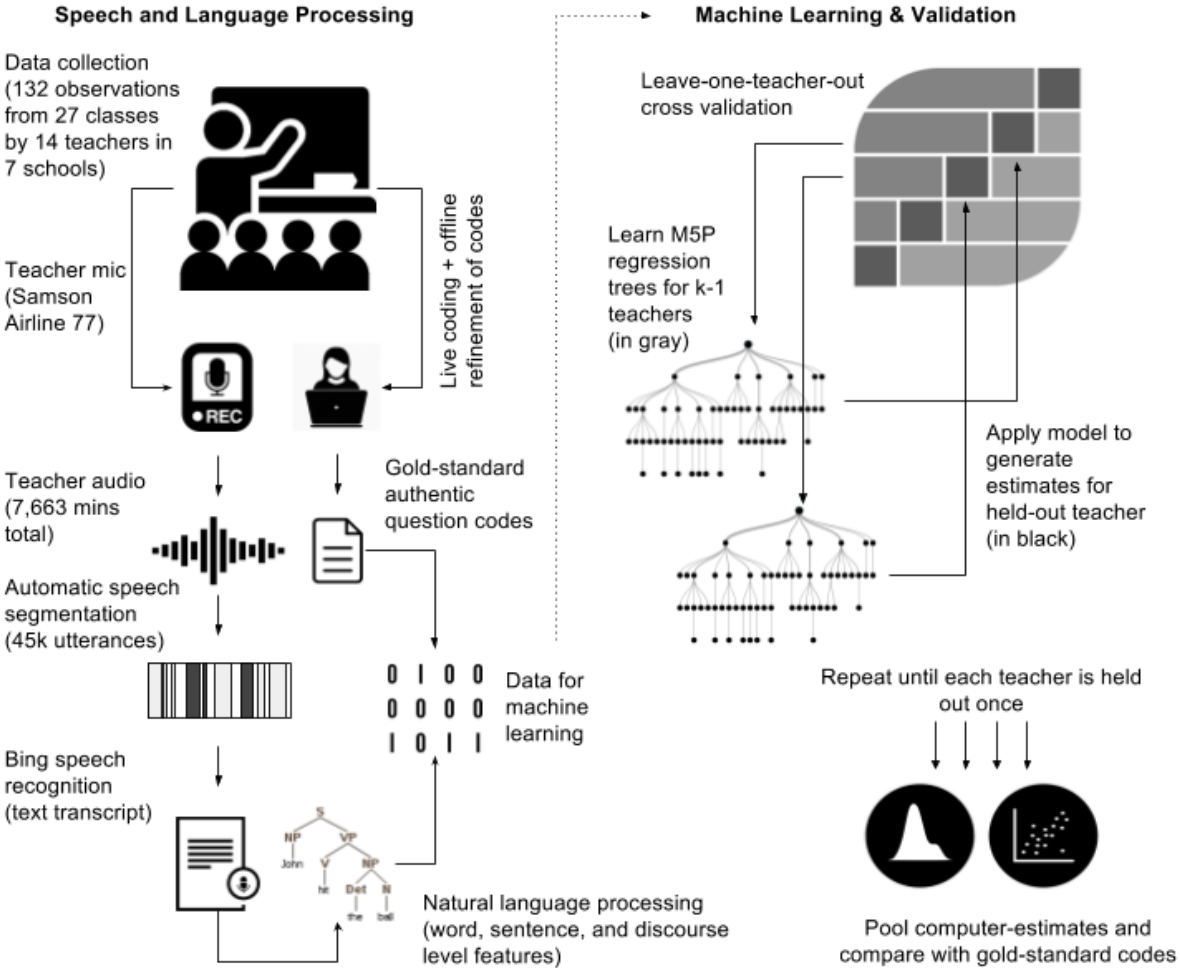

Figure 2. Distribution overlays and scatter plots of human- and computer-coded estimates of the prevalence of authentic teacher questions at the observation and class level: Semi-automated approach on Partnership for Literacy Study data.

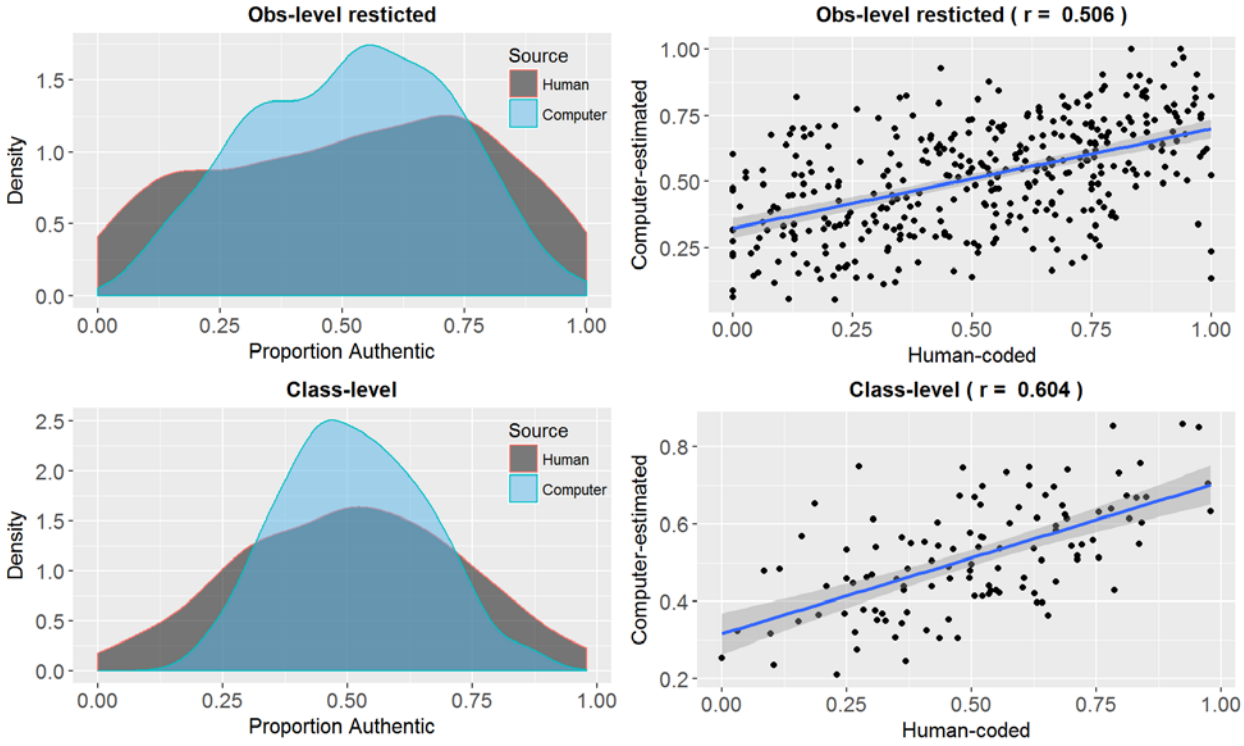

Figure 3. Distribution overlays and scatter plots of human- and computer-coded estimates of the prevalence of authentic teacher questions at the observation and class level: Fully automated approach on CLASS 5.0 sample.