Chapter 23 – Assessing Teacher Questions in Classrooms

Andrew M. Olney[1], Sean Kelly[2], Borhan Samei[1], Patrick Donnelly[3], and Sidney K. D'Mello[3]
University of Memphis[1], University of Pittsburgh[2], University of Notre Dame[3]

Introduction

Prompted first by the federal No Child Left Behind Act and subsequently the Race to the Top grant program, states have moved to adopt accountability systems that not only hold schools accountable for producing learning, but also individual teachers (Gamoran, 2013; Kelly, 2012). These efforts are consistent with research demonstrating the variability in teachers' capacity to improve student achievement growth (Hanushek & Rivkin, 2006, 2010; Kane et al., 2013; Nye, Konstantopoulos & Hedges 2004). In most cases, educator evaluations are based in part on student test scores, but also incorporate observational measures of instruction. Observations of classroom practice are valuable because they capture dimensions of schooling not captured by test scores, such as socialization outcomes in elementary school (Jennings & Corcoran, 2012). Classroom observations also enhance school principals' role in managing teachers' work (Harris, Ingle & Rutledge, 2014). Moreover, the presence of observational measures places an emphasis on the process of instruction itself and can be used to facilitate professional development quite apart from teacher evaluation (Goe, Biggers, Croft, 2012). Thus, many experts advocate balanced systems of accountability that include observational measures of instruction (Gates Foundation, 2013; Hamilton, 2012; Stein & Matsumura, 2009).

To date, several observational protocols have been developed, some for use across multiple classroom contexts (e.g., the Danielson Framework for Teaching [FFT], see Sartain, Stoelinga & Brown, 2011), and some targeted to instruction in specific subjects (e.g., the Protocol for Language Arts Teaching Observation [PLATO], see Grossman et al., 2013; and the Mathematical Quality of Instruction Instrument [MQI], see Hill et al., 2008). Systems of observational evaluation are currently in use in 47 states in the United States (American Institutes for Research, 2016). Yet, current methods are logistically complex, requiring observer training and are also an expensive allocation of administrator's time (Archer et al., 2016). For example, for use in evaluation, studies show that, typically, four class observations of each teacher are needed to provide a reliable sampling of teachers' instruction and afford an adequate opportunity to demonstrate excellence in multiple instructional domains (Kane & Staiger, 2012). Without a carefully managed classroom observation process, the observation results are open to criticisms of bias or arbitrariness.

To address the problems of cost, reliability, and bias inherent in traditional observational protocols, we have undertaken the development of a system called Classroom Language Assessment System (CLASS) 5.0 that automates the process of classroom observation through speech recognition and machine learning. The primary focus of our work is on teacher question-asking behavior, which is a common component across various well-known observation protocols. We first review some recent results in the classroom observation literature before describing our own work and making recommendations for future research.

Observing Effective Teaching

Efforts to evaluate teachers' performance are based on the logical assumption that the individual teacher's classroom is the most important site of student learning, and thus, closer evaluation and support for teachers' work constitutes a powerful lever of educational reform. Basic research on teacher effectiveness supports this perspective (Hanushek & Rivkin, 2010; Kane et al., 2013; Konstantopolous, 2014). For example, Nye et al. (2004) estimate that a 1 standard deviation increase in teacher effectiveness would increase student achievement by about 1/3 of a standard deviation. Other research finds somewhat smaller, but still important achievement gains attributable to teacher-to-teacher variability (Cantrell & Kain, 2013; see Hill et al., 2008 for a discussion of interpreting effect sizes in education research). In contrast, in much prior research, readily available indicators of teacher quality, such as years of experience, educational attainment, or certification status have generally explained a frustratingly small proportion of the variance in teacher effectiveness (Clotfelter, Ladd & Vigdor, 2006; Hanushek, 1986). Given these two sets of findings, directly assessing individual teachers' performance via test scores and/or teacher observations may offer the best insight into teaching quality. What does existing research demonstrate about the role of observation in assessing effective teaching?

A recent, multi-year study called Measures of Effective Teaching (MET) examined several different classroom observation measures, student perception surveys, and student achievement gains across approximately 3,000 teachers in seven states (Cantrell & Kane, 2013). Although previous research has shown a connection between various classroom observation measures and student achievement, the MET study is unusual in two respects. First, it used a randomized controlled trial design that, in year 1, collected teaching effectiveness data and built predictive models of teaching effectiveness, and in year 2, randomly assigned teachers to new classrooms to see if the predictive models from year 1 could account for changes in student outcomes in the randomly assigned classrooms. The purpose of the randomized controlled trial was to establish a causal, rather than correlational, correspondence between teaching quality and student outcomes, making MET the largest study of its kind to do so. Second, classroom observations were conducted using recorded video, allowing multiple observers and diverse observation measures for each video. This approach allowed for an in-depth examination of the reliability of the various classroom observation protocols across different types of observers.

MET used both general and subject specific classroom observation protocols, such that various observational measures of effectiveness could be compared to each other and value-added estimates of student achievement growth (Mihaly et al., 2013). One of the major MET findings was that these classroom observation protocols were all positively correlated with student achievement gains and were highly correlated with each other at the summary score level when using dis-attenuated correlations to account for measurement error (Kane & Staiger, 2012). Considering the correspondence in teacher ratings across different observational protocols, FFT, Classroom Assessment Scoring System (CLASS), and PLATO had pairwise correlations above 0.86, and FFT, CLASS, and MQI had pairwise correlations above 0.67, ostensibly lower due to the specific mathematics focus of MQI. A principal component analysis on each protocol yielded three major components that accounted for approximately 90% of the variance in scores across teachers. The first two components were the same across protocols: overall quality and classroom management (Kane & Staiger, 2012). The third factor varied across protocols, but most often involved question asking behavior, the focus of this chapter. Considering the relationship between observational scores and achievement gains, the teachers rated most effective on observations were also effective in raising test scores. For example, correlations between FFT (Danielson, 2011) scores and the value-added achievement measures (state tests) ranged from 0.17 to 0.41 (Mihaly et al., 2013, Table 3).

Measurement of Dialogic Instruction

One limitation of the protocols used in the MET study is the relatively coarse-grained nature of the coding as CLASS, FFT, and PLATO were all coded on 15-minute intervals, i.e., every 15 minutes, while MQI was coded on 7.5-minute intervals (Kane, Kerr & Pianta, 2014). The time delay between a classroom event and the coding of it on these interval boundaries may have facilitated observer's use of holistic judgments of instructional quality, which would explain the lack of differentiation among the various dimensions of these protocols with respect to the first principal component (e.g., the FFT has 22 rating components and 76 smaller elements within them).

In contrast, the present study is based on Nystrand's CLASS, a real-time, or "live" coding system first developed by Martin Nystrand and colleagues in the mid-1980s (Nystrand, 1988). Nystrand's CLASS focuses on individual questions and their properties, in addition to the basic allocation of classroom time to various instructional activities. In this study, we use data from updated versions of the original CLASS program, which was used in coding both archival data from the Partnership for Literacy Study (see Kelly, 2008) and in newly collected data. We pair these human coded measures from the CLASS program with new automated codes, referring to the automated version of the system as CLASS 5.0. Note that the MET study used a separate system, also called CLASS, developed by Robert Pianta and colleagues (see Allen et al., 2013).

The micro (i.e., individual question events) rather than macro orientation of CLASS 5.0 is highly salient to adopting a machine learning approach to classroom observation, because it provides labeled data conducive to training classification models. Application of machine learning to the data coded in the MET study seems much less promising, because the long durations between class events and actual coding creates a credit assignment problem (Minsky, 1961) in which it is unclear what action or event led to a given code. In CLASS 5.0, the relationship is much more direct, though not perfectly so. Instructional segments (e.g., discussion, lecture) have clear timestamped boundaries and categories. Within these timestamps, classroom activities and language are strong markers of the segment category. Questions likewise have clear timestamps, as do some clear properties like speaker identity. However, some associated properties extend beyond the question per se and instead are more properly considered part of a question event. These question properties include whether there was a response, cognitive level, authenticity, and uptake. These last two properties are the hallmarks of dialogic instruction, in which questions do not have predefined responses (authenticity) and are part of an evolving discussion that incorporate ideas from the respondent (uptake). Compared to common initiation-response-evaluation (IRE) format of classroom instruction in which the teacher quizzes students by asking them "test" questions, dialogic instruction focuses on the open-ended discussion and the exchange of ideas (cf. Bakhtin, 1981). For example, "What was your reaction to the end of the story?" is an authentic question because there is no pre-scripted response, and a follow-up question "Why do you think that?" has uptake because "that" refers to the student's previous reply. As is clear in these examples, dialogic properties are contextualized by the discourse and not purely determined by the question alone. Thus, the question event is characterized by the antecedents and consequents of the question in addition to the question itself.

Currently, it is not clear from the MET results whether instructional processes surrounding question-asking behaviors (one of the principal components of effective instruction in the MET study) have a significant effect on student achievement. On the other hand, previous research on the CLASS 5.0 system has shown that authenticity and uptake are significant predictors of student achievement (Gamoran & Nystrand, 1991; Gamoran & Kelly, 2003; Nystrand & Gamoran, 1997). Moreover, teacher training can increase the prevalence of dialogic instruction (Caughlan, Juzwik, Borsheim-Black, Kelly & Fine, 2013). Thus, the rationale of our work is that dialogic instruction, by promoting student achievement and being responsive to professional development, might be a crucial factor to assess when it comes to using classroom observation for measuring teaching effectiveness. In addition, unlike principal components of overall quality derived from statistical analyses of covariance, it can be precisely defined. However, an obvious limitation is that question-asking behaviors constitute a narrower domain of classroom instruction than assessed in global ratings. For example, in English and language arts, many important instructional dimensions (e.g., goal clarity, challenge) pertain to writing activities rather than discourse. Overall though, we view the measurement of dialogic instruction as an appropriate target of classroom observation to improve teaching effectiveness via feedback and professional development.

Class 5.0

In our work, we have pursued the goal of measuring dialogic questions in classrooms from two perspectives, text based and speech based. Text-based work makes use of both human transcripts of questions, which are available as archival data from previous CLASS 5.0 research, and automatic speech recognition (ASR) transcripts derived from current data collection. Speech recognition in classrooms is quite challenging given the ambient noise and the impracticality of individual microphones for each student. As a result, we have primarily focused on high-quality teacher audio collection and lower-quality classroom audio collection, from which we can determine student speech activity (D'Mello et al., 2015). Although speech features like prosodic, spectral, and voice quality features do contribute to the accuracy of question detection and instructional segment classification, text-based features alone are very effective (Blanchard, D'Mello, Olney & Nystrand, 2015; Blanchard et al., 2016b, 2016a; Donnelly et al., 2016a, 2016b, in press). Thus, text-based features are central to assessing classroom discourse, so there is considerable need for transcripts from which to extract these features. Perhaps surprisingly given the recent advances in ASR and industry claims of word error rates at 10% or less (Ryan, 2016), word error rates in our classrooms are closer to 50%, even when using a high-quality microphone and available state-of-the-art deep neural network-based ASR systems (Donnelly et al., in press).

We are currently still exploring how speech recognition errors differentially affect our models at the feature level. In our previous work, based on archived human transcripts from 418 classes, we trained J48 decision trees (Quinlan, 1993) that were able to automatically detect dialogic question properties at approximately 64% accuracy, which rivals humans coding questions out of context (Samei et al., 2014). In contrast, using new data collection ASR transcripts from 77 classes, our models are only 54% accurate. There are two likely causes for this difference in performance between models trained on human transcripts and ASR transcripts. First, ASR errors could be corrupting key features needed to build successful models. Second, the new data collection ASR transcripts have only a fraction of the original amount of human transcript training data, so the difference could be that there is not enough data to build a successful model. Both likely causes for performance differences may be examined by subsampling the human transcripts, i.e., creating new data sets for training by sampling without replacement various percentages of the total data set. Figure 1 shows model performance on human transcripts at 1% to 100% increments of the total data set. As the amount of data increases, model performance (defined as percent correct) improves, but the growth of improvement slows as more data are added.

In terms data size, the amount that is available from new data collection with ASR transcripts is approximately 6% of the human transcript data set size, for which Figure 1 shows trained models are 57.6% percent correct. Based on this analysis, it seems likely that a significant part of the difference between human transcript and ASR transcript model performance is due to lack of data, because when there is a comparable lack of data in the human transcript data set, the results are only 3.6% better rather than 10% better as they initially appeared to be. Overall, this implies that ASR errors are only negatively impacting model accuracy by 3.6%, and that this deficit could be narrowed simply by collecting more ASR transcript data for training. Using the equation of the line of best fit in Figure 1, the amount of data required would be approximately 2.5 times the number of human transcripts currently available. However, this estimate should be treated with caution since there is no guarantee that ASR based models will exactly follow the curve for models trained on human transcripts.

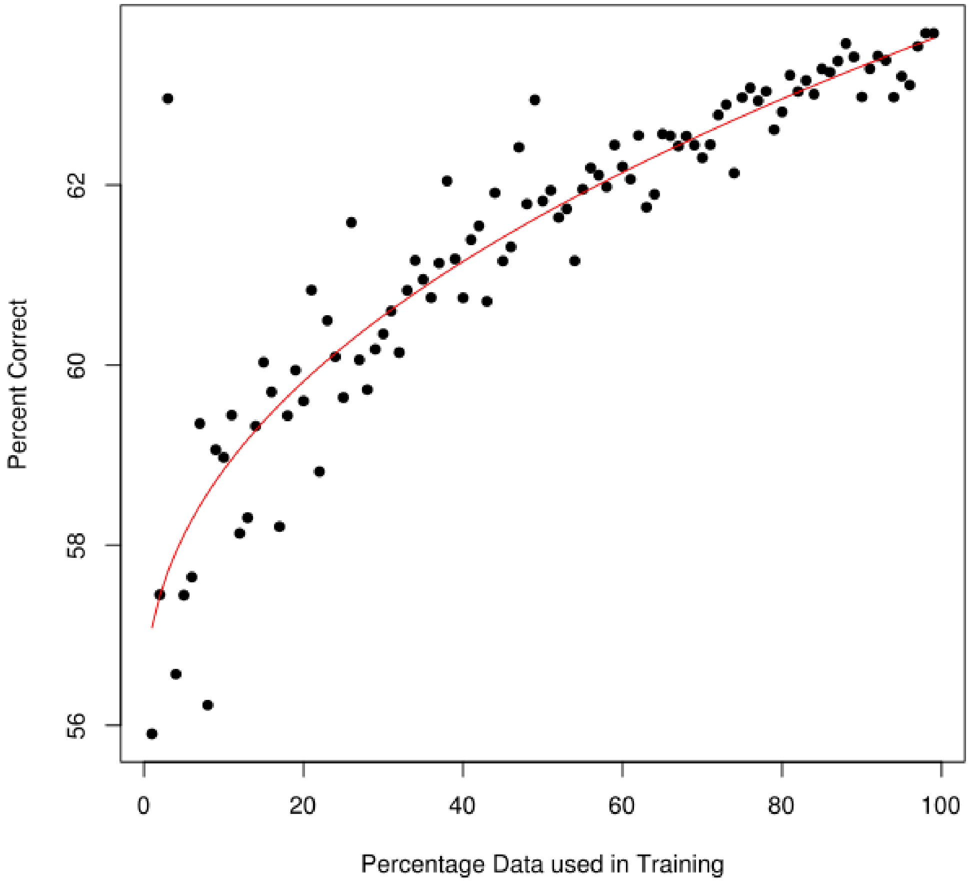

Figure 1. Model accuracy as a function of training data size for models trained on human transcripts.

In this chapter, we focus on our recent work training models for authenticity and uptake using human transcripts of questions collected in previous projects by Nystrand and colleagues.

Assessing Questions in Isolation

Our initial work on the assessment of dialogic questions focused on questions in isolation, meaning questions removed from the discourse context in which they occurred. The rationale for this line of inquiry was pragmatic, in terms of both the data available and building better machine learning models. The data available at the beginning of our project consisted entirely of archival data produced by previous versions of CLASS 5.0. These data contained human transcriptions only of coded questions and not of the surrounding speech. Non-instructional questions, such as procedural questions, were excluded from the coding scheme (Nystrand, 1988).

From the perspective of creating machine learning models, investigating isolated questions is also pragmatic because it raises the question of just how much information is needed to measure dialogic questions effectively. The human observers, situated in the classroom during live coding, have access to a considerable amount of contextual information, including spoken language, non-verbal communication, whether the students are paying attention, etc. Although we assume that the bulk of the coding decisions are based on spoken language, it may be the case that these other sources of information have some role to play.

Therefore, one of our first questions was whether a new set of trained human coders would code isolated questions as accurately as the original live coders. We randomly sampled 200 questions for authenticity and uptake (400 in all) from the approximately 25,000 questions in our archival data. The questions were evenly balanced such that half had the property in question, e.g., uptake, and the other half did not. Four trained human judges recoded these questions independently, with questions being presented in a random order. For comparison, though studies from CLASS 5.0 prehistory did not use chance-corrected agreement statistics like Cohen's kappa, inter-rater agreement defined as percent agreement has been reported as 81.7% for uptake and 78% for authenticity (See Nystrand & Gamoran, 1997, Chapter 2, Footnote 3).

Inter-rater agreement for the authenticity and uptake samples is shown in Tables 1 and 2, respectively, which are elaborated versions previously published in Samei et al. (2014). Three patterns are apparent in these results. First, kappa between new raters (R1-R4) on the isolated questions ranges between 0.18 and 0.46 for authenticity and between 0.31 and 0.51 for uptake. Although interrater reliability appears to be low, it corresponds to historical percent agreement on this task, given that 80% agreement on two evenly balanced classes would yield a kappa of about 0.35 (Bakeman, McArthur, Quera & Robinson, 1997). Thus, the agreement between new raters is reasonable except for the low 0.18 kappa between R3 and R4 for authenticity. Second, the kappa between the new coders and the original (O) live coders drops substantially and is equivalent to a 20–30% drop in percent agreement. It is noteworthy that while the original live coders presumably agreed with each other at the same level as the new coders, agreement between original and new coders is low. This indicates that different criteria are being used by original and new coders as the basis for their coding decisions. Third, the kappa between the J48 decision tree model (M) and the original live coders is substantially higher than the kappa between the original live and new coders. Indeed, the kappas between the model and the original live coders approaches what we would have expected to see between the original live coders themselves, if that data were available.

Table 1. Kappa for authenticity on isolated questions.

Rater	R1	R2	R3	R4	M
R2	0.44	-	-	-	-
R3	0.41	0.36	-	-	-
R4	0.46	0.55	0.18	-	-
O	0.13	0.17	0.25	0.10	0.34

Table 2. Kappa for uptake on isolated questions.

Rater	R1	R2	R3	R4	M
R2	0.45	-	-	-	-
R3	0.31	0.46	-	-	-
R4	0.51	0.47	0.36	-	-
O	0.22	0.25	0.30	0.23	0.46

As further discussed in Samei et al. (2014), these results are somewhat tempered by the finding that when considering the entire data set (i.e., approximately 25,000 questions), the percent agreement between the model and the original coders is 64% for authenticity and 62% for uptake. Thus, it appears that our models still need to account for 15–20% agreement to be on par with live coders, at least on this task where human transcripts of questions are classified without any context.

Assessing Questions in Sequence

We repeated the recoding task in various forms, incrementally including information such as who was speaking (useful since student questions are more likely to be authentic than teacher questions). In our latest evaluation, we think we uncovered the simplest combination of factors that can be used in a machine learning model. These include speaker identity and question transcript for all questions in each question/answer segment (assuming instructional segment boundary detection and classification). Two raters each with over 10 years of experience in coding dialogic questions independently rated a sample of 102 questions. The questions were sampled at the segment level, meaning question/answer segments were first randomly sampled, and then all questions from these segments were extracted in temporal order. The raters were presented with lists of questions with corresponding speaker identity (either teacher or student), question transcript, and segment boundaries. There were 14 segments in all, ranging from 1 to 24 questions in length. The prevalence of authenticity and uptake were representative of the entire dataset, with 48% of questions being authentic and 30% of questions having uptake. Unlike previous work, these properties were coded simultaneously for each question rather than having separate question sets for each.

The inter-rater agreements in kappa are shown in Tables 3 and 4 for authenticity and uptake, respectively. It should be noted that one coder (R1) failed to code six questions coded by both R2 and the original live coder (O). Therefore, kappas involving R1 are only based on the 96 questions that were rated, but all other kappas are based on the full set of 102 questions. Given that the kappa between R1 and O, using only 96 questions, is only 0.03 higher than the kappa between R2 and O, using all 102 questions, the exclusion of six questions appears to be contributing a very small bias in agreement, if any. Unlike the results for isolated questions in Tables 1 and 2, agreement between the new coders and the original live coder was quite high; in one case (R1 to O) the kappa of 0.61 is quite high for authenticity. In terms of percent agreement, these results range from 72% to 81% for authenticity and 71% to 77% for uptake. Although different coders achieved the highest agreements for authenticity (R2) and uptake (R1), the fact that they were able to do so with this restricted set of information suggests that it is possible for a machine learning approach to do equally well given the same information. This is remarkable considering that dialogic discourse is defined by the antecedents and consequents of questions, as only this context reveals the function, or effect, of a given question on the discourse. However, it appears to be the case that speaker, question transcript, and segment identification convey the same information or sufficiently correlated information to perform the coding task as well as a live coder with full access to the classroom context. Accordingly, building models with only this information is a focus of our current work.

Table 3. Kappa for authenticity with identity and segment information.

Rater	R1	R2
R2	0.48	-
O	0.44	0.61

Table 4. Kappa for uptake with identity and segment information.

Rater	R1	R2
R2	0.31	-
O	0.45	0.41

Assessing Questions without Bias

Although accurate measurement of classroom discourse is of considerable importance, of equal importance is unbiased assessment with respect to various socio-economic factors. This is a growing concern in artificial intelligence research (Hardt, Price & Srebro, 2016) because data-driven models will naturally reflect the biases in that data. When the predictions of the model are heavily weighted in high-stakes decisions, such as teacher assessment for promotion or tenure, it is critically important to ensure that all teachers are treated fairly. To better understand how our models were affected by bias, or equivalently to demonstrate that they work equally well for various socio-economic groups, we undertook an analysis of our original work that measured the dialogic properties of question in isolation. Specifically, we subdivided the data for various groups, built models with those subsets, and tested those models against different subsets as well as the full data set (Samei et al., 2015).

The distribution of schools in different geographic regions is shown in Table 5. Because the amount of data in some of these schools was rather small, we grouped them into Urban (Mid-size and Large Central City) and Non-urban groups (everything else). Furthermore, we were able to divide the data into groups who had received professional development training on dialogic instruction (Post-training) and those who had not but would later (Pre-training).

Table 5. Distribution of schools by geographic area.

School Category	Schools	Schools (%)
Large central city	4	19
Mid-size central city	7	33
Urban fringe of mid-size city	7	33
Small town	1	5
Rural inside MSA*	1	5
Rural outside MSA*	1	5

* Metropolitan Statistical Area.

The new groups we defined had different levels of uptake and authenticity, as shown in Table 6. Non-urban and Post-training groups had higher levels of authenticity and uptake than Urban and Pre-training groups. Also, the difference between Pre- and Post-training levels of authenticity and uptake was relatively large compared to the difference between Urban and Non-Urban levels. These different levels of authenticity and uptake across groups suggest the potential for bias if data from one group were used to build a model for another group.

Table 6. Percentage of authenticity and uptake across groups.

Group	Authenticity (%)	Uptake (%)
Non-urban	54	23
Urban	47	20
Post-training	52	24
Pre-training	39	15
Full	50	21

Note. Adapted from Samei et al. (2015).

To investigate the possibility of bias, we built two kinds of models for each group. In tenfold cross validation models, we used each group for both training and testing data. The second set of models trained on a group and tested on its dual, i.e., Pre-training vs. Post-training and Urban vs. Non-urban. The accuracies of these fitted models are shown in Table 7. For comparison, Table 7 also shows tenfold cross validation on the full model.

Table 7. Accuracy of models for authenticity and uptake when trained and tested on different groups.

Training Data	Test Data	Authenticity Accuracy (%)	Uptake Accuracy (%)
Non-urban	Non-urban	61	59
Urban	Non-urban	62	62
Non-urban	Urban	60	63
Urban	Urban	62	60
Post-training	Post-training	63	61
Pre-training	Post-training	59	62
Post-training	Pre-training	60	64
Pre-training	Pre-training	64	61
Full	Full	64	62

Note. Adapted from Samei et al. (2015).

As shown in Table 7, training with Urban gives better results for authenticity than does training with Non-urban regardless of which group is used as test data, while training with Non-urban gives better results for uptake in the case of testing with Urban only. Pre- and Post-training give best results for authenticity when trained and tested against themselves (using tenfold cross validation), but give best results for uptake when tested against the other. These inconsistent results suggest that using any one group to train will create bias of some kind when testing using another group.

To investigate these inconsistent results, we analyzed the individual features used in each model using the Correlation-Based Feature Subset (CFS) algorithm. We found that different subgroups used different kinds of language to mark authenticity and uptake. For example, Urban groups used first and second person "be" verbs and judgmental words like "think," "find," and "thought" in authentic questions, but Non-urban groups did not. Likewise, Urban groups used second person pronouns and negation in questions with uptake, but Non-urban groups did not. Post-training groups had a greater prevalence of "be" verbs for authenticity, and a greater use of modal verbs like "would," "can," and "could," than did Pre-training groups.

For authenticity, training on the full data set led to the better or equal performance than training on any subset. We speculate this is because all subgroup-specific features were represented in the model and prevented it from being biased toward or against dialogic classifications because of the absence of a diagnostic feature. For uptake, training with the full data set is slightly worse than training with Non-urban or Post-training and testing on their duals. However, when tested on themselves (tenfold cross validation), both Non-urban and Post-training are worse than training and testing on the full data set.

While the best possible scenario would be for there to be no difference between groups, the current result can be considered the second best: there are differences, but they can be modeled without explicitly defining the groups in the model. A worse scenario would be if different groups used language in opposite ways, i.e., a marker for authenticity in Urban subset was a marker for non-authenticity in Non-urban subset. If this were the case, then it would be necessary to infer which group was being measured to "select" the right markers for measurement. Fortunately, it appears that group identity can be ignored at the modeling stage if the training data are sufficiently diverse to represent all groups. Diversity in the training data is critical for unbiased assessment in our models.

Conclusions and Recommendations for Future Research

We reviewed current work on classroom observation in the study of teacher effectiveness as well as our own work on measuring the dialogic properties of questions in classrooms. Previous research, including the MET project, has shown that though the year-to-year effects of high instructional quality relative to lower quality instruction are sometimes small, the cumulative effects across a student's K–12 career can be considerable. Moreover, classroom observations can be used to reliably identify effective instructional practice, and form the basis of professional development efforts and other approaches to school reform. Our work shows that automation holds much promise in scaling up classroom observations in a reliable and fair way.

However, several questions raised by our work present a challenge to future researchers. Currently, we have compared automated coding to relatively fine-grained, question-level human coding, but much existing educational improvement efforts use even courser-grained, global rubrics. Thus, further research is needed comparing automated portraits of effective instruction to a greater array of human-coded approaches. Might the MET data constitute a promising existing resource for such analyses? One concern is that the quality of the audio in these recordings is challenging, and there is also a practical barrier because the MET videos can only be accessed via a remote interface that includes a video viewer, but nothing else. Nevertheless, the potential is great given the large numbers of videos and overall diversity of the MET sample.

A second question for automation in future research concerns the connection between the dimensions of instruction that can be observed *by discourse alone* and achievement growth. In MET, the principal component analysis did identify a component associated with discourse. Yet, further research is needed, building on the MET design, to understand the robustness and malleability of discourse effects in contrast to more generic domains of practice (e.g., that include writing assignments).

Third, the archival data used in this study date to the early days of NCLB and, and it is possible that the prevalence of dialogic discourse and teacher-to-teacher variability in approaches to discourse have changed. On the one-hand, increased teacher accountability and other standards-based reforms may focus attention on test preparation activities and away from dialogic approaches. On the other hand, effective discourse practices, including dialogic practices, are an explicit component of common observational protocols used to evaluate teachers. Are current trends in education promoting teaching practices that are consistent with what research deems effective? Analyses of the MET and other new data may shed light on these questions.

Another question, closely following our own work, is how to build models based on the finding that speaker identification, question transcript, and segment identification seem to be all that is needed to reach live human-coder levels of agreement for dialogic question properties. A related question pertains to whether model-coded dialogic question properties predict achievement gains as well as human codes. If successful, the CLASS 5.0 system can be used a tool for accurately coding classroom discourse, thereby providing valuable information to researchers, teachers, teacher educators, and professional development personnel.

This work is relevant to GIFT in at least two ways. First, dialogic questions could be incorporated into intelligent tutoring systems (ITSs). The work reviewed in this chapter provides a foundation for the generation of such questions computationally. In some respects, this is trivial: asking the user what they think about a topic without some pre-scripted answer or weaving what the user says back into the conversation are common strategies used by chatbots. However, the simplicity of generating such questions is offset by the complexity in keeping the conversation going once they have been asked – in essence, understanding the student's response well enough to generate new dialogue on the fly. This ability, currently lacking in chatbots, is considered by some to be the ultimate proof of artificial intelligence, the so-called "Turing Test".

Thus, incorporating dialogic questions into ITSs may be beyond the current state of the art, but the existing work in GIFT supporting user personalization provides a starting point from which to build. For example, Sinatra (2015) proposes using a dialogue template approach where the user's log in name is stored as a variable and then inserted into dialogue templates to create dialogue like, "Welcome to the tutorial, [name]" and also proposes a survey to elicit user interests so that they can likewise populate templates for instructional dialogue. These proposals are similar to what currently is done in chatbots using a variable/template approach, but they differ in terms of how the variables are assigned. In the case of GIFT, Sinatra's proposal is to assign these variables outside of the dialogue. GIFT could benefit from variable assignment both outside and inside the dialogue to support dialogic instruction, as internal variable assignment taking place in dialogue would make it easier for an intelligent tutor to weave the student's responses back into the conversation.

The second way in which the work described in this chapter is relevant to GIFT is for hybrid instruction in which the human instructor and GIFT are members of the same instructional team but with different roles. For example, the human instructor may lead a face to face session with students and then pass the students off to GIFT for self-directed practice. In such a scenario, it is important for GIFT to understand what has taken place during the human-led portion of the instruction. Automated assessment of classroom discourse provides such a model of understanding. By using the techniques described in this chapter as well as the techniques by D'Mello and colleagues on instructional segment classification, GIFT could understand both coarse-grained classroom activities and fine-grained discussion, and use this knowledge to tailor its own instructional activities. For example, if the human instructor spent 20 minutes lecturing on a topic to provide an overview, GIFT could trim or eliminate that portion of its instruction. Similarly, highly dialogic discussion during the human-led portion could be modeled and used by GIFT, e.g., in the user personalization scheme described previously, to keep students engaged and motivated. In summary, for GIFT to function in hybrid human/artificial intelligence instructional teams, an understanding of the human-led portion of the instruction is essential.

Acknowledgements

This research was supported by the US Army Research Laboratory (W911NF-12-20030), the Institute of Education Sciences (R305A130030), and the Office of Naval Research (N00014-12-C-0643). Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the author and do not represent the views of these sponsoring agencies.

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B. & Pianta, R. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system-secondary. *School Psychology Review*, *42*(1), 76–98.

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.

Bakeman, R., McArthur, D., Quera, V. & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, *2*(4), 357–370.

Bakhtin, M. M. (1981). *The dialogic imagination: Four essays*. University of Texas Press.

Blanchard, N., D'Mello, S., Olney, A. M. & Nystrand, M. (2015). Automatic classification of question & answer discourse segments from teacher's speech in classrooms. In O. C. Santos et al. (Eds.), *Proceedings of the 8th international conference on educational data mining* (pp. 282–288). International Educational Data Mining Society.

Blanchard, N., Donnelly, P., Olney, A. M., Samei, B., Ward, B., Sun, X., ... D'Mello, S. K. (2016a, September). Identifying teacher questions using automatic speech recognition in classrooms. In *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue* (pp. 191–201). Los Angeles: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W16-3623.

Blanchard, N., Donnelly, P. J., Olney, A. M., Samei, B., Ward, B., Sun, X., ... D'Mello, S. K. (2016b). Semi-automatic detection of teacher questions from human-transcripts of audio in live classrooms. In *The 9th international conference on educational data mining* (p. 288–291).

Cantrell, S. & Kane, T. J. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Retrieved 2016-12-27, from http://metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf.

Caughlan, S., Juzwik, M. M., Borsheim-Black, C., Kelly, S. & Fine, J. G. (2013). English teacher candidates developing dialogically organized instructional practices. *Research in the Teaching of English*, *47*(3), 212–246.

Clotfelter, C. T., Ladd, H. F. & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of Human Resources, 41*(4), 778–820.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching.* ASCD.

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. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (pp. 557–566). New York, NY, USA: ACM. Retrieved from http://doi.ACM.org/10.1145/2818346.2830602 doi: 10.1145/2818346.2830602.

Donnelly, P. J., Blanchard, N., Olney, A. M., D'Mello, S. K., Nystrand, M. & D'Mello, S. K. (in press). Words matter: Automatic detection of questions in classroom discourse using linguistics, paralinguistics, and context. In *Lak '17: Proceedings of the seventh international conference on learning analytics & knowledge.* ACM.

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., ... D'Mello, S. K. (2016a). Automatic teacher modeling from live classroom audio. In *Proceedings of the 2016 conference on user modeling adaptation and personalization* (pp. 45–53). New York, NY, USA: ACM. Retrieved from http://doi.ACM.org/10.1145/2930238.2930250 doi: 10.1145/2930238.2930250.

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., ... D'Mello, S. K. (2016b). Multi-sensor modeling of teacher instructional segments in live classrooms. In *Proceedings of the 18th ACM international conference on multimodal interaction* (pp. 177–184). New York, NY, USA: ACM. Retrieved from http://doi.ACM.org/10.1145/2993148.2993158 doi: 10.1145/2993148.2993158.

Gamoran, A. (2013). *Educational inequality in the wake of No Child Left Behind.* Spencer Foundation Lecture to the Association for Public Policy and Management, Washington, DC. Retrieved from: http://www.appam.org/awards/spencer-foundation-lectureship.

Gamoran, A. & Kelly, S. (2003). Tracking, instruction, and unequal literacy in secondary school English. In R. Dreeben & M. T. Hallinan (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. (1991). Background and instructional effects on achievement in eighth-grade English and social studies. *Journal of Research on Adolescence*, *1*(3), 277–300.

Gates Foundation. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study*. Bill and Melinda Gates Foundation. January.

Goe, L., Biggers, K. & Croft, A. (2012). *Linking Teacher Evaluation to Professional Development: Focusing on Improving Teaching and Learning.* Research & Policy Brief. National Comprehensive Center for Teacher Quality.

Grossman, P., Loeb, S., Cohen, J. & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores. *American Journal of Education*, *119*, 445–470.

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

Hanushek, E. A. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature, 24*(3), 1141–1177.

Hanushek, E. A. & Rivkin, S. G. (2006). Teacher quality. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the Economics of Education* (Vol. 2, pp. 1051–1078). Amsterdam: North Holland.

Hanushek, E. A. & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, *100*(2), 267–271.

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

Harris, D. N., Ingle, W. K. & Rutledge, S. A. (2014). How Teacher Evaluation Methods Matter for Accountability: A Comparative Analysis of Teacher Effectiveness Ratings by Principals and Teacher Value-Added Measures. *American Educational Research Journal, 51*(1), 73–112.

Hill, C. J., Bloom, H. S., Black, A. R. & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*, 172–177.

Hill, H. C., Blunk, M. L., Charalambos, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L. & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, *26*, 430–511.

Jennings, J. L. & Corcoran, S. P. (2012). Beyond high stakes tests: Teacher effects on other educational outcomes. In S. Kelly (Ed.), *Assessing teacher quality: Understanding teacher effects on instruction and achievement* (pp. 77–96). New York: Teachers College Press.

Kane, T. J., Kerr, K. A. & Pianta, R. C. (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project*. Jossey-Bass.

Kane, T. J. & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation. Retrieved 2016-12-27, from http://k12education.gatesfoundation.org/wp-content/uploads/2016/06/MET_-Gathering_Feedback_for_Teaching_Summary1.pdf.

Kane, T. J., McCaffrey, D. F., Miller, T. & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* Bill & Melinda Gates Foundation.

Kelly, S. (2012). Understanding teacher effects: Market versus process models of educational improvement. In S. Kelly (Ed.), *Assessing Teacher Quality: Understanding Teacher Effects on Instruction and Achievement* (pp. 7–32). NY: Teachers College Press.

Konstantopolous, S. (2014). Teacher effects, value-added models, and accountability. *Teachers College Record*, *116* (1).

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: Sage Publications. Retrieved from http://d-scholarship.pitt.edu/26219/.

Mihaly, K., McCaffrey, D., Sass, T. R. & Lockwood, J. R. (2013). Where You Come From or Where You Go? Distinguishing Between School Quality and the Effectiveness of Teacher Preparation Program Graduates. *Education Finance and Policy, 8*(4), 459–493. https://doi.org/10.1162/EDFP_a_00110

Minsky, M. (1961, Jan). Steps toward artificial intelligence. *Proceedings of the IRE, 49*(1), 8-30. doi: 10.1109/JRPROC.1961.287775.

Nye, B., Konstantopoulos, S. & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*(3), 237–257.

Nystrand, M. (1988). *CLASS (classroom language assessment system) 2.0: A windows laptop computer system for the in-class analysis of classroom discourse*. Wisconsin Center for Education Research.

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

Quinlan, R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

Ryan, K. J. (2016). *Who's smartest: Alexa, Siri, and or Google Now?* Retrieved 2016-12-29, from http://www.inc.com/kevin-j-ryan/internet-trends-7-most-accurate-word-recog.

Sartain, L., Stoelinga, S. R. & Brown, E. R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Consortium on Chicago School Research.

Samei, B., Olney, A. M., Kelly, S., Nystrand, M., D'Mello, S., Blanchard, N., ... 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. M., 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 O. C. Santos et al. (Eds.), *Proceedings of the 8th international conference on educational data mining* (pp. 444–447). International Educational Data Mining Society.

Sinatra, A. M. (2015). A Personalized GIFT: Recommendations for Authoring Personalization in the Generalized Intelligent Framework for Tutoring. In *Foundations of Augmented Cognition* (pp. 675–682). Springer, Cham. https://doi.org/10.1007/978-3-319-20816-9_64.