Toward the Development of Learning Analytics: Student Speech as an Automatic and Natural Form of Assessment

Marcelo Worsley
Stanford University

Paulo Blikstein
Stanford University

Abstract
While many of the nation's educators and leaders are calling for students to develop 21st century competencies through student-centered, hands-on learning, most school systems continue to cling to traditional forms of instruction. This reliance on traditional forms of instruction is not without merit, however, assessment within open-ended learning environments remains difficult, and, often times, seemingly unsatisfying. This is further complicated by the large emphasis placed on students demonstrating their knowledge through standardized tests. As a way of addressing this discontinuity between practice and theory, we have worked to develop Learning Analytics—a set of multi-modal sensory inputs, that can be used to predict, understand and quantify student learning. Central to the efficacy of Learning Analytics is the belief that educators will be able to more easily adhere to learning recommendations when they are given the proper tools; in this case, tools for more accurately assessing student knowledge in open-ended learning tasks. Accordingly, this study presents finding related to one of the Learning Analytics modalities: speech. By leveraging the tools of text and speech analysis, we are able to identify domain independent markers of expertise. Some of the most prominent markers of expertise include: user certainty, the ability to describe things efficiently and a disinclination to use unnecessary descriptors or qualifiers. While many of these are things one would expect of an expert, some of them are also observed among novices. To explain this we report on learning theories that can reconcile these seemingly odd findings, and expound on how these markers can be useful for identifying student knowledge learning over the course of an intervention or classroom experience.

Introduction
The 2010 National Education Technology Plan was adamant in the desire to usher in transformational changes to the US education systems. Such changes would arise through a more widespread recognition of the 21st century competencies—critical thinking, complex problem solving, collaboration and multimedia communication—as paramount to being an expert learner and to being productive in the current world economy. Moreover, the plan calls for the proliferation of student-centered learning curricula that are flexible, and highly customized to student interests, needs and cultures. Such curricula would more heavily privilege experiential and project-based learning, and lie in direct opposition to traditional forms of classroom instruction. Anyone educated in the American public school system in the last century would likely agree that seeing such changes in the way students learn would be transformational. Nonetheless, the general motivation behind a more student-centered approach to learning is not new. Throughout the 20th century educators, humanitarians and psychologists alike were fervently making the case for a more personalized form of learning (Dewey, 1902; Freire, 1970; Freudenthal, 1973; Montessori, 1964, 1965; Papert, 1980; von Glasersfeld, 1984). While some of the learning theories associated with their works have gradually made their way into the

education system, others were met with rapid adoption, followed by a rapid demise and an eventual resurgence. Such was the case for Papert's constructionism, which, when implemented by other researchers, produced controversial results. (For an example of positive results see Clements and Nastassi (1988) and for an example of negative results, see Pea (1987)). These findings seemed to be a precursor to the work of Kafai et al (2009), who described the ambiguities associated with understanding and assessing how students learn in constructionist learning environments. Similarly, Bruckman (2000), reports difficulties with uneven student learning and challenges in the assessment of student portfolios. Many have used such findings to discredit project-based learning.

In contrast, we hypothesize that most of the project-based learning theories preceded the technological tools needed to reliably assess them. More specifically, until recently, the tools available for trying to understand learning in open-ended learning environments were limited to traditional approaches, largely because of issues of scalability. However, we have finally reached a time where the tools available for assessing open-ended learning are sufficiently advanced to truly measure the nature of student progress that takes place in these environments. Such tools allow educators to study learning processes in large scales, instead of simply looking and learning products (Blikstein, 2009, 2011). Furthermore, these tools shift the paradigm from periodic and intrusive assessments, to largely continuous and passive assessments. Beyond this, the types of assessment techniques that we will be discussing venture away from domain-specific representations of knowledge, and look to capture more generalizable markers to student learning that can help both students and teachers better recognize learning. In particular, we are examining the progression from novice to expert as an emergent phenomenon populated by a myriad of small, atomistic expertise micro events, or "markers of expertise" (Blikstein, 2011)

This study focuses on one such form of natural assessment, speech, with the hope of answering the following questions:

1. How can we use informal student speech to decipher meaningful *markers of expertise* (Blikstein, 2011) in an automated fashion?
2. What do these markers of expertise tell us about learning?

Prior Work

This research is situated at the intersection of language, automatic text and speech, and sentiment analysis. In the following sections, we will highlight salient research in these areas, paying particular attention to previous publications that address two or more of these areas.

Techniques for Performing Automatic Assessment of Student Text and Speech

There are three primary forms of text used by researchers and practitioners in this space. These include student essays, student responses to short answer questions, and transcribed student dialogue in intelligent tutor environments. In this section, we highlight the most common techniques for extracting and analyzing these different forms of text. We also identify different approaches that researchers use to facilitate both unsupervised and supervised learning of student learning models.

Word Tokenization

By far the most common and simplest analysis used for doing automatic grading of text, is through the use of the bag-of-words model to represent documents and/or utterances. Such an approach involves tokenizing the words in a document to create feature vectors. These feature vectors are then processed using a number of techniques to ascertain a certain document's similarity to an expert model or some other predefined set of classes. It is important to realize that the feature vectors treat each word independently, and to acknowledge that word order is not typically taken into account. Instead, the

assumption is that the words contained in the document, are, by themselves, sufficiently indicative of a student's proficiency. Later on, we will present some of the ways that researchers use for getting around this assumption.

Unsupervised Learning

Chen et al (2010) describe a technique for automatically scoring essays using cluster analysis. Under this approach, the system need only receive the number of clusters that the user would like. For example, if using a traditional grading scale of A-F, the user could specify 6 clusters, and the system would proceed to organize the different items based on their similarity. Again, this similarity is based on the occurrence and frequency of the words in each text. Using this system, Chen et al (2010) were able to achieve over 90% accuracy when comparing their results to adjacent matches, and 52% when comparing their results to exact matches.

Supervised Learning

Unlike Chen et al (2010), most models use a supervised learning approach to classify student text. Valenti et al (2003) describe a variety of systems that operate using this model, including Intelligent Essay Assessor(IES), Educational Testing Service WE, Electronic Essay Rater (E-Rater), Bayesian Essay Test Scoring sYstem(BETSY) Automark, . These systems varied in accuracy, producing scores that ranged from 80 % to 96%.

Improving the Word Tokenization

The simple process of fragmenting a piece of text into individual words, and eliminating punctuation can be an involved process. One must ensure that the method used is the same across all texts. For example, one must be careful of how she deals with contractions, capitalization, varying verb tenses and proper nouns. Furthermore, in building a specific language model (in the supervised case) one has to be sure to reserve some word probability for new words. To address these issues, Chen et al (2010), for example used a combination of word stemming and stopword extraction. Stemming involves the removal of word suffixes that would cause a system to interpret “scientist” differently than ”scientists” even though they refer to the same basic term. Removal of stop words is a process of eliminating all words that are extremely common in the English language (“the”, “of”, “he”, etc.). Litman et al (2009), also use stemming and stop word elimination to improve their model's performance., but were met with mixed success from stemming.

A final approach often used to improve word tokenization is presented by Rus et al (2009). Rus et al, was using short student essays to assess prior knowledge about the circulatory system. This was part of a pretest for the students' upcoming science unit. Like Litman et al (2009) and Chen et al (2010), Rus et al (2009) also used stemming to reconcile words of a similar root. However, in addition to stemming they also introduced bigrams into their model. The introduction of bigrams typically allows for the system to better capture the context of each utterance. For their analysis, across a variety of machine learning algorithms (Naïve Bayes, Bayesian Nets, SVM, J 48 Decision Trees and Logistic Regression) the use of bigrams consistently outperformed unigrams.

Dimensionality Reduction

Another difficulty with using all of the words in a document to represent that document's feature vector, is the sheer size of the data. In order to avoid this, one can use LSA (latent semantic analysis) (Valenti, 2003). LSA is a special form of singular value decomposition, in which a large, often sparse, matrix is approximated using a lower dimension matrix. This approximated matrix uses “concepts” as the matrix’s columns. As such, using LSA is often associated with performing concept mapping between the original word tokens and specific content domains/sub-domains (Manning and Schultz, 1999). IES

is an example of this type of a system.

Additional Text Based Features

Word tokenization and subsequent analysis represents a simple way of analyzing documents that can produce reasonable accuracy given sufficient training data, or a high level of certainty about the number of classes in the samples (in the case of cluster analysis). However, as noted above, there are a number of problems with simply using the bag-of-words model to analyze student learning. Not surprisingly, researchers have identified a host of other meaningful features to extract from text. We describe these features, and how they are extracted, in the following sections.

Content Word Extraction Models

A commonly used feature in student learning models is the number of content words. Whereas the Naïve Bayes approach looks at all of the words in a document, this approach is only concerned with the number and distribution of content words in a piece of text. In some since, having an understanding of the number of content words keeps track of how much of the student's text is on topic. However, without the use of additional information about the context in which each of these concept words is used, this approach would suffer from many of the same problems as the bag-of-words model. To address this concern, some researchers have looked at the distribution of concept words across a text, especially a dialogue. In their studies they found that having an unequal distribution of concept words, in transcribed dialogue, suggests that learning has taken place. In this case, one can safely assume that the learning would be likely to have occurred when the majority of the content words are used closer to the end of the transcription.

Chi et al (2010) also looked at the use of concept words in student dialogue. Instead of looking for an uneven distribution of these words, they looked at the number of utterances or sessions that included one of more concept word. Their findings suggest that the number for "student concept sessions" is positively correlated with learning.

As one can imagine, automatically extracting the number of concept words can take on a variety of forms. For many of the systems used (ie. intelligent tutoring systems) previous student dialogues and existing software organization produces the necessary training data. It is also possible to manually construct such a list of concept words, as was the case in Rus et al (2009). Unfortunately, manually building such a lexicon is extremely time consuming and laborious. Moreover, it is likely to exclude a number of important features unless it is constructed by a domain expert. For these reasons Purandare and Litman(2008) and Litman et al (2009) introduced the use of web-based resources as domain dictionaries. They used an online physics glossary to populate a list of concepts related to mechanics and electricity and magnetism. They also used a stop word list to eliminate useless words like, "of" in the "Law of Conservation," from their domain-specific lexicon. Using this approach proved to be quite successful in their work.

Discourse Analysis

So far, we have primarily discussed content of student text. A number of researchers also explore discursive aspects of text and speech utterances. Litman et al (2009), explore discourse in the context of when an utterance is made. Accordingly, utterances are tracked as happening before or after tutor generated questions, advancing to a deeper aspect of a particular concept, or moving forward to a different concept. As will be discussed in section 5, these discursive features, when combined with information about student affective state will prove to be a superior predictor of learning than either item on its own.

Dependency Parsing

Many of the discursive features utilized by Litman et al (2009) are specific to intelligent tutoring systems and cannot easily be extracted from written work. Semantic role labeling and dependency parsing, however, can be automatically extracted from student text. Semantic role labeling and dependency parsing are both based on the presence of ample training data. This can either come from previous correct utterances from an intelligent tutoring system, or from hand-coded rules. Assuming that one has a corpus to use for determining dependencies or semantic roles, these items can be compared to expert answer choices. Conceptual Rater (C-Rater), for example, uses a predefined set of relationships to compare with student responses for conceptual accuracy. While less common in automatic essay grading, automatic assessment of conceptual accuracy is a key component in all intelligent tutoring systems. This is because the goal of conceptual accuracy is to model student correctness, which has been shown to be a good predictor of learning by several studies (Forbes-Riley and Litman, 2010, Forbes-Riley et al, 2009).

Common Textual Features

It is also common to use a number of other features to describe a text. These include: the total number of words in a single text (can be a proxy for the duration of speech), the number of concept repetitions, the average number of words per turn, the number of student turns (in transcribed dialogue), the percentage of words from the student (also in intelligent tutors). Interestingly these features' correlation to transcribed text and written text is not always the same. For example, Purandare and Litman(2008) did a comparison between the two modalities and found that verbosity (ie the number of words in a text) is positively correlated with learning for written work, but negatively correlated with learning for spoken text. As such, one must be careful when looking to transfer results from one modality onto another modality.

Extracting Lexical, Prosodic and Spectral Features from Speech and Dialogues

While text mining can provide useful cues about student learning, analysis of student speech, when available, can also be of utmost importance. In the predictive learning space, the majority of the work has focused on detecting user uncertainty and user affect. Accordingly, we will review some of the literature on extracting lexical, prosodic and spectral features that identify some of these different emotions and their implications.

Lexical Features

Lexical features consist of a variety of linguistic cues that a user may be expressing an emotion. These features typically include a whole class of disfluencies: filled pauses, restarts, interruptions, repairs, edits, fragments, creaks and whispers. Forbes-Riley and Litman(2010) use lexical features in conjunction with prosodic features to train an uncertainty classifier. This classifier was built using a logistic regression over previous data from the ITSPOKE-WOZ corpus. This corpus consists of a hand annotated and manually transcribed spoken dialogue from the ITSPOKE tutoring system.

Prosodic Features

Prosodic features consist: of maximum, minimum and mean values in pitch and intensity; duration, accents and intonation. Researchers have been able to use these features to examine and predict a number of human emotion states. In the student learning literature the emotion states most frequently observed are those of certainty, frustration, boredom and anger.

Through the work of Liscombe et al, the ability to use prosodic features to detect certainty in student responses has accurately been solved. Liscombe et al (2005), demonstrated the ability to detect certainty in student responses with 76.42% accuracy using a combination of turn based and breathe

group based features. While Liscombe et al (2005) found that the increased granularity provided by breathe group features outperformed a classifier that was based only on turn based features, it was still the case that the combination of turn based features and breathe group based features produced the best result. This suggests that certainty is based on both local and global features of a turn.

When identifying the predictive power of uncertainty in reflecting student learning, the results tended to vary. Litman and Riley(FOAK) found that while uncertainty is not explicitly useful as a predictor or learning, "Feeling of Another Knowing" (FOAK) is meaningful. FOAK is a measurement that is characterized by the Harmann coefficient and Gamma coefficient. Both of these values consist of a ratio between uncertainty, incorrectness, certainty and correctness. In this way, being able to measure uncertainty is still quite useful. Contrarily, Forbes-Riley et al (2009) reports that uncertainty is a meaningful measure in and of itself. In this case, they attribute uncertainty as being an indicator of an opportunity for constructive learning. More specifically they report that demonstrating neutrality (the opposite of certainty, or uncertainty) indicates a lack of student engagement. Other researchers have found similar results (Forbes-Riley and Litman (2010).

While not yet developed to the extent of certainty, researchers have also looked at the correlations between frustration, anger, boredom and annoyance with learning. Computational linguists have been able to utilize prosodic features to build classifiers for many of these emotions, but much of the research in student learning still tends to use manually annotated data. The exception to this is in frustration detection. Forbes-Riley et al (2009) demonstrate that, like uncertainty, frustration can be a key component in predicting student learning. Again, they attribute frustration to being an indicator that a student is engaged in the learning process, whereas, non-frustration can signal disengagement.

Spectral Feature

Spectral features refer to the spectrum of frequencies that is observed when analyzing the formants of a spoken utterance. This form of analysis did not appear to be significantly common within the student learning community. Nonetheless, recent work in smile, laughter and sarcasm detection may prove to be useful in advancing the extraction of student affective state from spoken utterances, as these emotions may indicate student engagement.

Combining Textual and Acoustic Features

As one would anticipate, combining features can offer a richer understanding of the complex interactions that take place in learning. Litman et al (2009) describe impasses as a situation in which a student is wrong, or right, and expresses uncertainty. In the intelligent tutoring paradigm this represents an important learning opportunity in which knowing both correctness and student affect can substantially alter the system's response. Accordingly, Litman et al (2009) recognize impasse as being positively correlated with learning.

Similarly, as noted in section 3.1, Forbes-Riley and Litman (2010) use a combination of textual and speech features to train an uncertainty classifier. This ostensibly produces improved accuracy above simply training on text based features or speech-derived features.

Another result from Litman and Forbes-Riley (2010) uses both text based correctness and speech based uncertainty to calculate "Feeling of Another's Knowing". This term, which was briefly presented earlier, allows them to determine the Harmann coefficient and Gamma coefficient, which are said to identify FOAK. In their study, they found that the Harmann coefficient provided greater predictive utility than the Gamma coefficient. Moreover, FOAK when combined with correctness, was more predictive of learning that just student correctness. Furthermore, FOAK, was in and of itself a better

predictor of student learning than student correctness.

Exploring Student Sentiment

Using Video to Determine Affective State

Using the Facial Action Coding System (FACS), researchers have been able to develop a method for recognizing student affective state by simply observing their facial expressions. In the case of Craig et al.'s (2008) study, researchers were able to perceive boredom, stress and confusion by applying machine learning to the data produced. Data was collected while students interacted with AutoTutor, an intelligent tutoring system, in the context of learning principles of science. The technique that Craig et al. validated is a highly non-invasive mechanism for realizing student sentiment, and can be coupled with computer vision technology to enable machines to automatically detect changes in emotional state or cognitive-affect.

Determining Affective State through Dialogue

Researchers have also used conversational cues to realize student emotional state. Similar to the FACS study, D'Mello et al. (2008) designed an application that could use spoken dialogue to recognize the states: boredom, frustration, flow, confusion and neutral. The researchers were able to resolve the validity of their findings through comparison to emote-aloud (a derivative of talk-aloud where participants describe their emotions as they feel them) activities while students interacted with AutoTutor. Though both studies had shortcomings, the research captures the potential for empowering educators through student sentiment awareness. These findings give researchers additional resources for developing generalizable student models, while also allowing educators to easily track student progress using various forms of expression.

More recent work in this space by Conati (2009) is able to accurately predict the affective state, and the source of the change in affective state for users as they interact with a computer based tutoring system. In particular, the system was able to effectively predict when students experienced joy, distress and admiration.

Process Mining and Personalization

With a large portion of intelligent tutoring systems designed for creating effective distance learning solutions, one of the principle points of flexibility is dynamic a curriculum. By identifying particular learning goals, past experience and areas of interest, students are able to receive a highly customized curriculum. From an artificial intelligence standpoint these systems operate by matching the preferences identified by the current student to student models—created by collecting data from previous students (Brusilovsky, 1999).

Some researchers working in this space use collaborative filtering to probabilistically determine what content to present. Others, like the work of Ameshi and Conati (2007), use unsupervised clustering techniques to realize existing behavioral patterns in previously collected data. Once salient clusters have been identified the model is used to pinpoint the cluster that is most similar to the student that is currently using the system (a process sometimes referred to as process mining).

Knowledge Tracing

Related to curriculum personalization is knowledge tracing and process mining. Knowledge tracing involves using student responses and actions as a way for determining how well a student understands a particular concept. Several researchers make use of knowledge tracing for helping the system understand when a student is ready to move on to the next topic(Beck and Sison, 2006).

Collective Benefits of Artificial Intelligence

Though artificial intelligence technologies tend to favor enacting improved individual learning, they also have significant relevance for adapting learning environments at the classroom level. Because of the intrinsic power of machine learning for processing large quantities of data, intelligent technologies can be effective at recognizing classroom level patterns that may not be readily apparent to educators. A good example of this in the work of Plummer, Cox and Dale (2009), who used machine learning to recognize shortcoming in the Grade Grinder, an automated assessment tool. Through the classification of problem difficulty and correction difficulty the group was able to establish a model for understanding student learning complications.

Similarly, Anaya (2009) was able to leverage machine learning for uncovering previously unperceivable information about student collaboration in a web based forum. Through this study both the educator and students were able to better manage the collaboration process. Finally, Dringus (2005) was also able to leverage the results from a textual analysis of a web-based discussion forum to formulate a coherent description of group knowledge formation. This description added significant value to the instructional team as it revealed to them order and structure in a learning space that they previously perceived to be devoid of any such coherence.

While the design for this study was informed by the aforementioned works, the current study ventures to explore student learning in a highly unstructured and open-ended learning environment. Many of the previously mentioned research studies were tied to intelligent tutoring systems that had constrained language models, scripted questions, and a limited, and known, set of possible solutions.

Data

The data for this study comes from interviews with 15 students from a tier-1 research university. Subjects were asked to draw and think aloud about how to build various electronic and mechanical devices. Below are the two prompts given to students.

> *Question 1*:
> Imagine that you want to build a system to maintain the temperature of your room at 80 degrees F. You have a temperature sensor, a fan, a heater, and a temperature controller with a sensor and switch for controlling the fan and cooler. How would you do it?
>
> *Question 2*:
> You have been asked to design a system to automatically separate glass, paper, metal and plastic. The pieces would arrive one by one, and won't overlap with one another. Each kind of materials can come in any size and shape. How would you design a system to accomplish this task?

The above questions were posed in a semi-structured clinical interview format. Question 1 was used as a control question, whereas question 2 was used for the eventual analysis. Student speech was transcribed by graduate and undergraduate students. Additionally, prior to the interviews, the subjects were labeled as being experts, intermediates or novices in engineering and robotics. This classification was based on previous formal technical training either through a degree program or through a lab course on physical computing. The classification was also influenced by previous face-to-face interactions between the raters and the research subjects. As such, we will describe the expertise classification as a "perceived" classification that may be somewhat noisy. The data consisted of audio files, transcriptions of the interviews, and digitized drawings that the students produced during the interview. There is also a video feed of the drawings as they were being constructed, in addition to

video of the student gestures, and facial expressions. This paper focuses on the pedagogically relevant features that were extracted from the audio files and transcripts as analyzed by humans and by artificial intelligence.

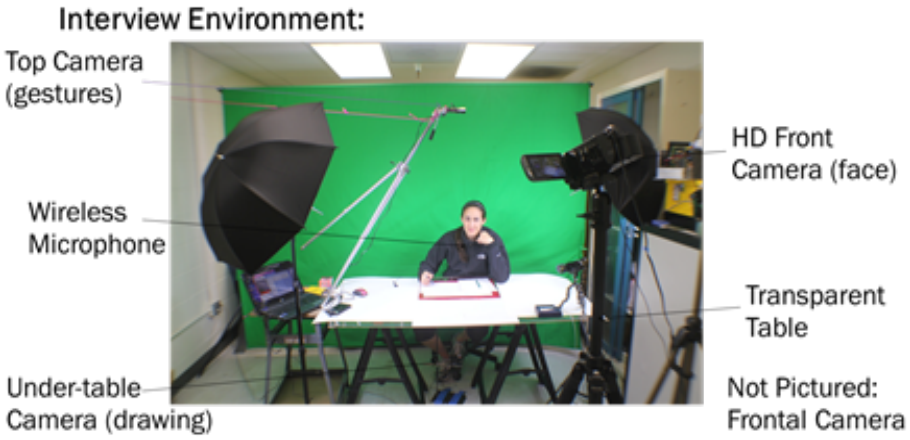

Figure 1- Picture of the Interview Environment

Research Subject Demographics

Of the 15 students, 8 were women, 7 were men; 7 were from technical majors, 3 were undergraduates, and 12 graduate students. It is also useful to note that the classes were not balanced. There were 3 novices, 9 intermediates, and 3 experts.

Table 1 - This table describes the research subjects. Expertise, Gender, Level of Education, Extent of Technical experience and participation in the a lab course on physical computing are presented

Subject	Expertise	Gender	Undergraduate	Technical	Physical Computing Course Participant
1	Intermediate	Female	1	0	1
2	Intermediate	Female	0	1	1
3	Expert	Male	0	1	0
4	Intermediate	Male	0	1	1
5	Intermediate	Male	0	0	1
6	Intermediate	Male	0	0	1
7	Intermediate	Female	1	0	1
8	Expert	Female	0	1	0
9	Novice	Female	0	0	1
10	Intermediate	Male	0	0	0
11	Intermediate	Female	0	1	1

12	Novice	Female	1	0	1
13	Novice	Female	0	0	1
14	Intermedi-ate	Male	0	0	1
15	Expert	Male	0	1	0

Methods

This study uses a mixed method approach of assessing qualitative data using quantitative techniques. Many of the techniques are borrowed from machine learning and artificial intelligence. We will briefly describe the techniques used for transforming the text and raw audio into data that can be quantitatively analyzed, and then move into the specifics of data analysis.

Human Transcript Rating:

Crowd-sourcing was used to determine the human perceived level of expertise depicted in each transcript. A minimum of 5 readers rated each transcript as coming from a novice, intermediate or expert student. These ratings were normalized and then used to determine the human-perceived label for each research subject.

Feature Extraction Techniques

Prosodic Feature Extraction

For the pitch, intensity and duration analysis, sound files were fragmented into interviewee turns, and individually analyzed for the mean, minimum and maximum values. After extracting this information across each turn, the data was aggregated to produce average, minimum and maximum values across all of the turns for a given user. Range values were also calculated for pitch and intensity based on the average minimum and maximum values across turns. In extracting the pitch from the sound the minimum and maximum were 75 Hz and 300 Hz, respectively. The prosodic features also included the ranges for speech rate (words/second).

Spectral Feature Extraction

Spectral analysis was also completed using the Praat software package. The first three formants were extracted from all audio files, based on turn, with maximum frequency values differing for men and women (5100 Hz and 5500 Hz, respectively). These values were also used to determine the range between F1 and F2, and F2 and F3, and these values have previously been shown to be significant for a number of different sentiments.

Linguistic Feature Extraction

Linguistic features were mined from the transcribed audio using the Python Natural Language Toolkit (NLTK) module for tokenization. The transcripts had been systematically labeled for: sentence restarts; filled pauses (“ums,” “ahs,” etc.); pauses; other filler words (“like,” “I mean”). The count for each of these values was normalized by the total number of words that a subject uttered, and reported as a percentage of that individual's words. Individual counts, and aggregate counts across all disfluency types were included in the model. This text-based parsing also afforded the extraction of questions, since the transcripts included punctuation. This allowed for us to include this as an additional feature for each user.

Sentiment Feature Extraction

For user sentiment we leveraged the Linguistic Inquiry and Word Count (LIWC) and the Harvard

Inquirer. Again, utilizing NLTK, we ran the transcribed audio files against all of the classes in LIWC and the Harvard Inquirer. For Harvard Inquirer, a given word was mapped to a sentiment if any of its variants were associated with that sentiment. Thus, no contextual disambiguation was used. Lastly, only sentiment classes exhibited by two or more research subjects were included in the final analysis. For a list of the sentiment classes that were taken into consideration, please visit the LIWC website and Harvard Inquirer Website. Both LIWC and Harvard Inquirer are hand-constructed lexicons that have been validated using other testing sets.

Content Word Extraction
We also compared student text to individual content word lexicons from chemistry, mathematics, computer science, material science and general science. These word lexicons were mined from domain specific websites. An individual transcription was credited with a certain word only if the utterance was an exact match to how it appeared in the lexicon. Stemming was not used.

Word Dependencies Extraction
Dependency parsing was completed using the Stanford Parser (Klein and Manning, 2003). The parser creates a syntactic parse of each sentence and enumerates the different word dependencies. It also provides a specification about the type of relationship that exists between each word. For our data we used two sets of dependency data. One set was the normalized number of parses and average depth of each user's speech. The second set only looked at the parses that contained a content word. This was done to approximate the frequency with which individuals said something meaningful, with regard to specific domain content, as opposed to merely using a content word out of context. In both cases, the word pair relationships of interest were the noun-subject relations, direct object relations, adverb-modifier relations and general dependencies that do not fall into any of the other predefined categories.

Extracting N-grams
Finally, n-grams were extracted from the speech transcriptions as a way of capturing contextualized domain content. Unigrams through tri-grams were included in the model, as long as they appeared at least twice within the corpus. The n-grams were tokenized by the word_punctuation method in NLTK, and then passed into the Weka n-gram tokenizer (from previous experience this proves to be a consistent way of doing the tokenization, whereas simply doing it in Weka can produce peculiar results). Stop words were removed from the n-grams, and the total number of n-grams was capped at 3000. These n-grams were all converted to lower case, for accurate comparison across occurrences.

Data Analysis Techniques

Expectation-Maximization Analysis
Expectation maximization (EM) is an iterative algorithm that involves two steps to determine the maximum likelihood estimates of parameters based on unobserved latent variables. The expectation step computes the log-likelihood that the estimated latent variables belong to their currently assigned clusters. The maximization step estimates the distribution of parameters from the cluster probabilities and stores them as instance weights. These newly calculated estimates of parameters are recycled through the expectation-maximization sequence until no further gains are made in cluster parameters. Because EM uses equal weighting of each feature, all values were standardized to have zero mean and unit variance. Additionally, EM seems like an appropriate tool for analysis because it is entirely unsupervised, which is to say that it has no knowledge of the labels that we assigned to each research subject. Instead, it simply uses the data to try to find the most closely related items.

Results

Because this study involved an exploration into a large number of analysis types, we will focus our attention on the results that seem most pertinent to understanding how to assess learning in open-ended learning environments. We will present human-derived predictions of expertise and conclude with highlights from the automated assessment of the student speech.

Human Rating

Table 2 - Confusion Matrix of Human Rating against the true class labels

Classified as → True class ↓	Novice	Intermedi-ate	Expert
Novice	0	3	0
Intermediate	2	5	2
Expert	1	0	2
Total	3	8	4

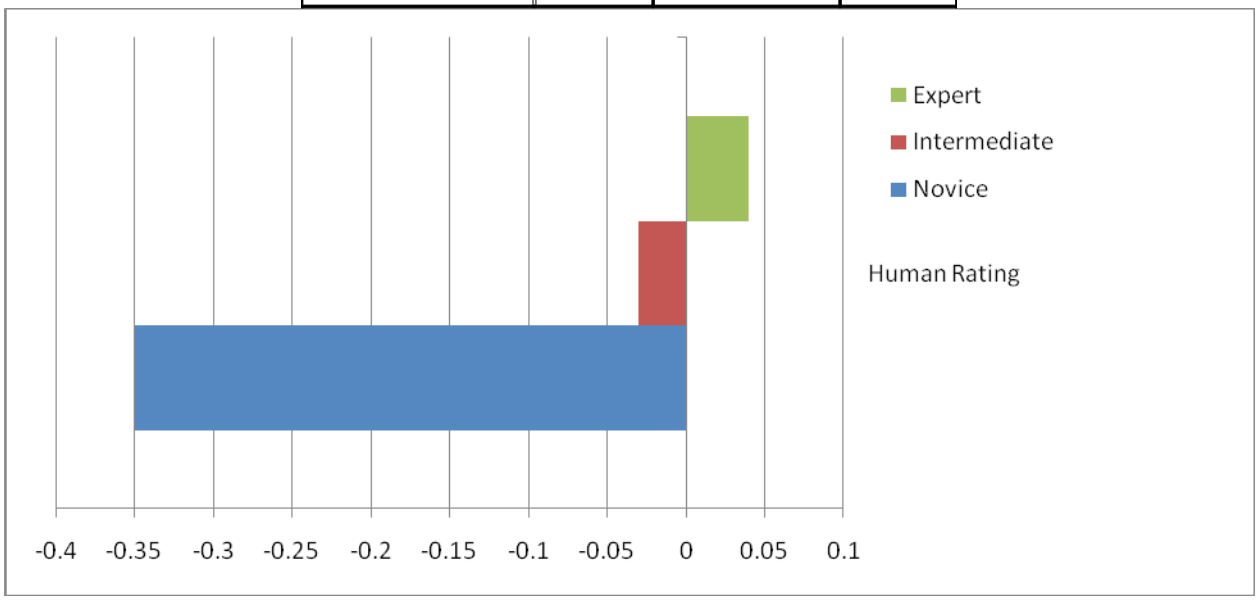

Figure 2 - Average Human Rating Based on Class Labels. The mean scores (standardized) were -0.35, -0.03 and 0.04 for Novice, Intermediates and Experts, respectively

Speech Analysis

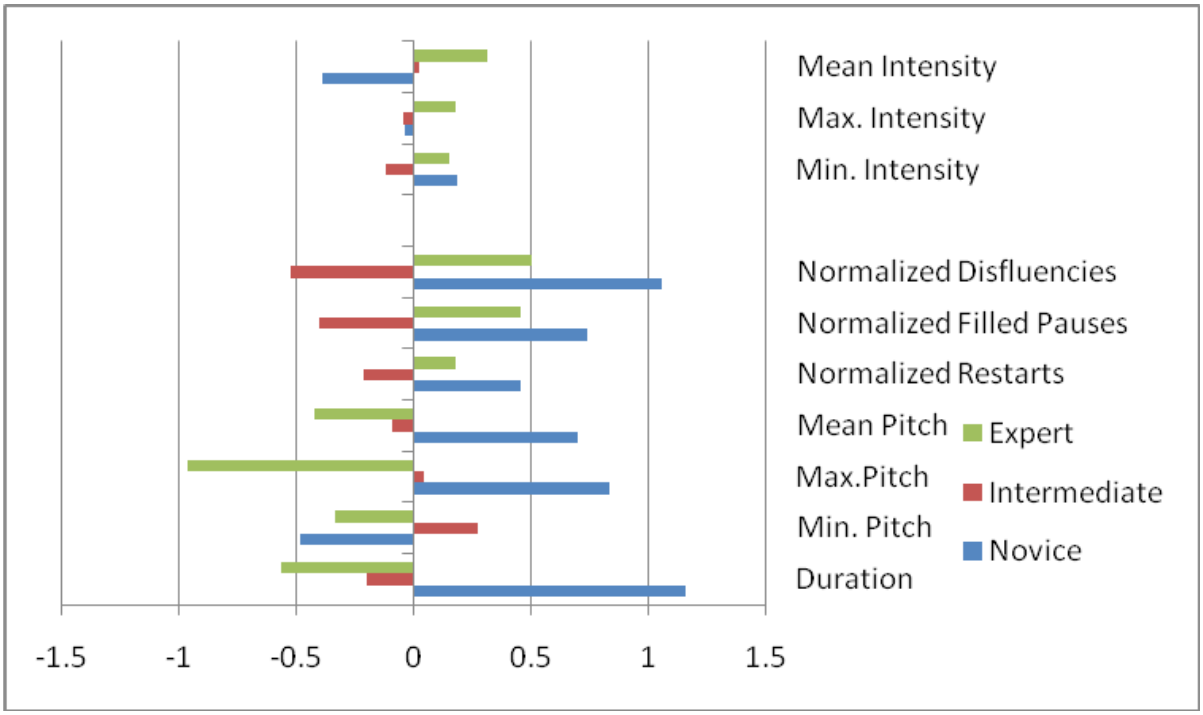

Figure 3- Summary values for the various speech features. These speech features include linguistic cues and prosody.

Sentiment Analysis

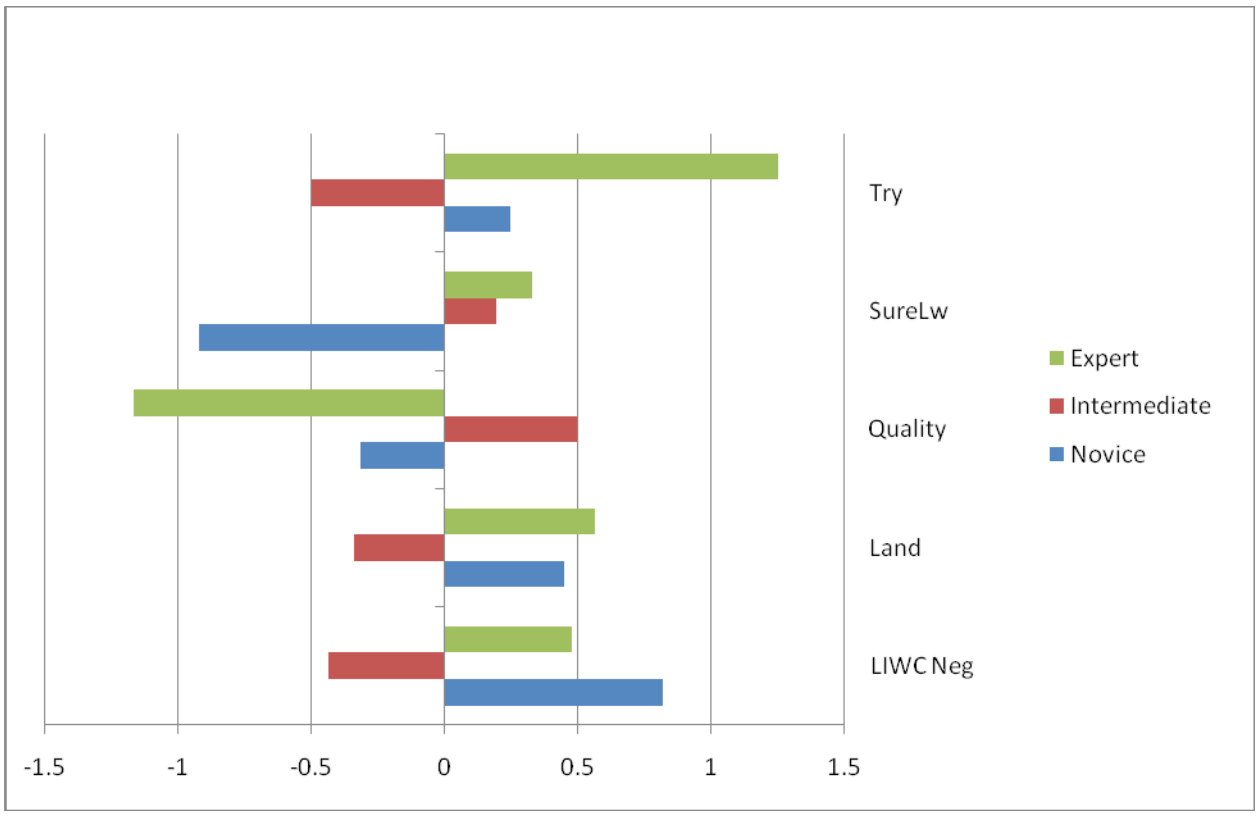

Figure 4- A Summary of meaningful sentiments and their relative prevalence among the different levels of expertise

Discourse Analysis

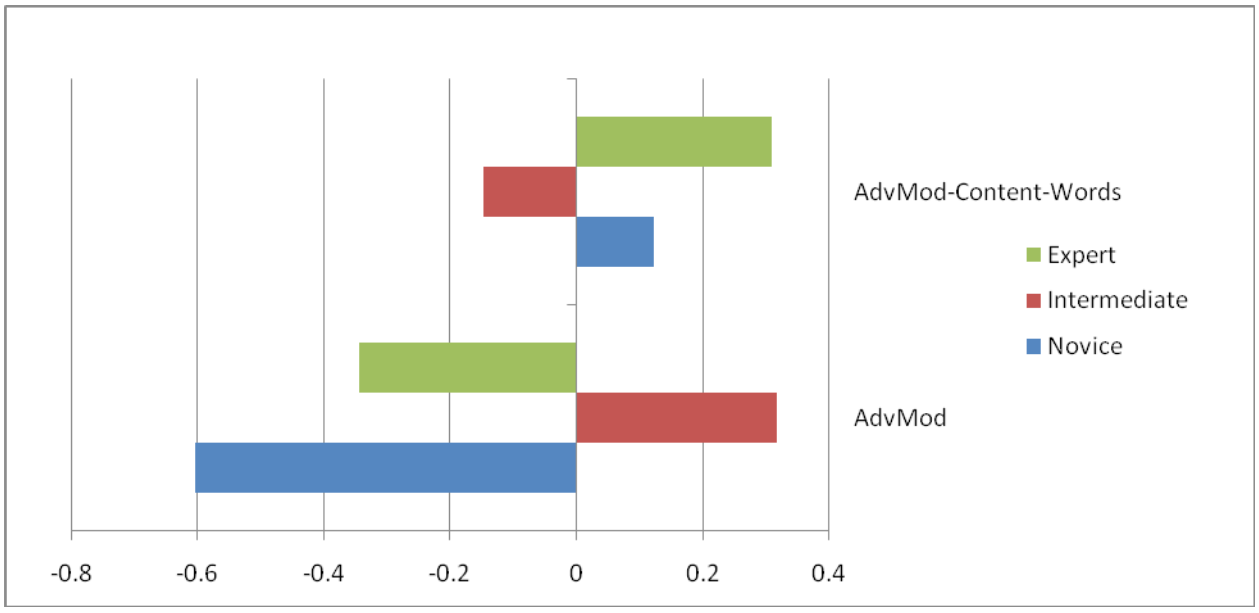

Figure 5- A figure comparing the frequency of adverb modifiers used in general speech versus content specific speech. AdvMod-Content-Words is a subset of AdvMod.

Combined Speech and Sentiment Analysis Expectation-Maximization Results

Table 3 - The Confusion Matrix attained when doing expectation maximization of the combined set of speech and sentiment features

Classified as → True class ↓	Novice	Intermediate	Expert
Novice	3	0	0
Intermediate	0	8	1
Expert	0	1	2
Total	3	9	3

Cluster Centroids for Combined Speech and Sentiment Analysis

Table 4 – EM cluster centroid values for the different features included in the analysis

	Novice	Intermediate	Expert
Duration	0.97	-0.30	-1.23
Filled Pauses	-0.1267	-0.5108	1.78
LIWC Neg	0.32	-0.52	1.3
SureLw	-0.68	0.26	0.65
Quality	-0.55	0.62	-1.1

Summary of Results

Table 5 – A Comparison of the results obtained when running EM on different sets of features. In all cases EM was run with 3 clusters

Type of Analysis	**Accuracy**	**Features of Interest**
Chance	0.33	N/A
Humans	0.47	Fluency, Number of Misconceptions, Confidence
Content Words	0.47	Mathcontent words, Chemistry Content Words
Dependencies	0.47	Adverb Modifiers
Dependencies of Content Words	0.47	Content Based Adverb Modifiers
Speech Analysis	0.73	Pitch, Filled Pauses, Duration, Disfluencies

Sentiment	0.87	LIWC Neg, Land, Quality, Try, SureLw
Sentiment+Speech	0.87	Filled Pauses, LIWC Neg, SureLw, Quality

Validating Features With the Control Group

The first question (about temperature control) was used as a control question to ensure that the features that we selected were not speaker specific. This control question was one that we believed everyone would feel comfortable answering, and would produce little variation across the spectrum from novices to expert. We ran the same set of analysis listed above, on the control data and found no correlations with expertise, which suggests that the features that we utilized were not limited to being speaker specific effects. Additionally we controlled for gender and level of education discovering the features that best predicted these. Once again, we found that none of the features that we used for predicting expertise were correlated with the features that accurately predicted male/female or undergraduate/graduate.

Discussion

Human Ratings

The fact that humans expressed great difficulty with this task is both unsurprising and unsettling. It is unsurprising in the context that the initial labels of expertise were themselves, informed by human judgment. A large part of the consideration behind the ratings had to do with a student's formal instruction in STEM discipline, but the reality is that those classifications alone do not always constitute a good basis acknowledging expertise. Instead, observing an individual in practice, as was the case for the expertise labels that we assigned, can help ensure more holistic labeling. The variability of human ratings is disconcerting because this is often the way students in project based learning environments are assessed, if they are assessed at all. Without the presence of consistent and on-going assessment of student work it can be difficult for students to learning effectively.

The Nature of Informal Discourse and the Relative Unimportance of Content Words

When considering the informality of the interviews, the irrelevance of content words is not unexpected, and may be very indicative of the target implementation environment. As people engage in casual conversation, there may be a ceiling to the amount of domain-specific nomenclature they use. This limit may be out of a desire to simplify things for the listener and make sure that they understand. It may also be that college students are all at a relatively equal level of scientific terminology when they engage in casual conversation.

Furthermore, this result is in line with learning theory about informal discourse. The language of science and mathematics are decidedly different from the language of everyday conversation (Brown and Spang, 2008). Because of this, it is difficult to assume that students will employ noticeably different levels of science and mathematics terminology when not in a formal setting. In addition, the types of questions that we asked the research subjects did not neatly fall into any particular field of study. What's more, the nature of the open-ended design space, is that people are expected to bring previous knowledge from a variety of backgrounds and use that to solve a problem. As such, it could be perfectly conceivable for a computer scientist, chemical engineer and mechanical engineer to all come

up with expert solutions to a problem using completely different nomenclature. In fact, this lack of reliance on domain-specific information may prove to be quite useful for doing this type of analysis in multi-disciplinary or inter-disciplinary spaces. One no longer has to be concerned with having content lexicons for all of the content areas represented by the students. Nor must they create content lexicons for every possible area that the students might explore with their projects.

Inconclusiveness of Discourse Analysis
While human raters were keen to identify misconceptions and reward well constructed approaches to solving this problem, the technique that we utilized for perceiving discourse was largely inconclusive. Novices, intermediate and expert users all used a varying number of word dependencies, suggesting that the measure that we used was more closely tied to speaker behavior than to expertise. What is interesting however, is the difference in the ways that experts tend to use adverbial modifiers. Whereas expert users were the least likely to use adverbial modifiers in a general context, they are the most likely to use adverbial modifiers in the context of content words. This suggests that the experts do have a better handle on the material in that they are able to be identify more causal relationships between entities of conceptual importance to the exercise. Intermediate users on the other hand tend to shy away from using adverbial modifiers when discussing content matters. This idea is further complicated by the relative prevalence of adverbial modifiers in novice speech. It is our supposition that this is closely linked to student self-perceived expertise, which we will discuss in more detail in the following section.

Expertise as a Measure of Certainty (clearer description of what is in the different categories)
The findings from the speech and sentiment data seem to point to a complementary result around user certainty. On the sentiment side, more expert individuals tended to use more SureLw[1] words as to suggest less uncertainty. This makes sense from a cognitive perspective since the expert is likely to have had more exposure to the types of problems being posed and can therefore be more assured in the validity of her solution. The novice, who may have the proper inclination in designing their system, may struggle because they have yet to see a similar system implemented, and therefore express uncertainly about the feasibility of their approach. This is precisely what was observed with the two research subjects that occupied extremes of the expertise spectrum. One student was an undergraduate student majoring in the humanities, who produced the same solution as an engineering PhD. What differs in their presentation of the solution, however, is how certain they are of its feasibility. The uncertain student goes so far as her berate her own idea and even calls it crazy. The engineering student produces no such expressions of infeasibility, but instead offers the solution in a very controlled systematic fashion. These two students also demonstrated the dichotomy in speaking efficiency and mean pitch that also appeared as an indicative feature in the results. The engineering student maintained a steady, relaxed pace, exhibiting very little excitement or displeasure. He chartered a very direct path through how he would design the system. The humanities student, on the other hand expressed a full range of emotions; everything from excitement and laughter to apparent confusion. The student was also frequently distracted by side thoughts about how she used to wonder how they performed the separation of glass, plastic, paper and metal, when she was younger.

Interpreting the results for Try, Quality and even normalized disfluencies is a bit more challenging, as these did not follow a linear path from novices to experts. Instead, the peak value for these was seen among the intermediate level research subjects. Novice and expert values were, surprisingly, quite similar. In the case of normalized disfluencies, the experts and intermediate individuals actually had more than novices. One possible reason for this is that more experienced users have a self-initiated mandate to keep talking as a way of affirming their level of understanding. They have self-image that

[1] SureLw words are described in detail at http://www.wjh.harvard.edu/~inquirer/SureLw.html and includes words like absolute, actual, certain, essential, even, indeed, inevitable, etc. There are 175 of these words in total.

they feel compelled to maintain, but when they are faced with complex problems that require them to think on the spot, they have to find ways to fill the time. In contrast, a novice may be more comfortable sitting quietly, because they have accepted their lack of knowledge in this area. In many respects they can hide behind their professed ignorance in the subject matter.

In the case of the Quality sentiment—which is loosely a proxy for the number of descriptors and modifiers a student uses—we observed that individuals of intermediate knowledge level were the most inclined to use these qualifiers, whereas both experts and novices were far less likely to use them. This we directly attribute to the student knowledge levels. The novice student does not have a strong enough grasp of the material to incorporate a bounty of qualifying words into their utterances. The intermediate student, in contrast, who is just beginning to develop mastery in a subject area, may just be learning new terms and concepts. As a practical application of that learning, and as a desire to sound informed, the student may find herself using a great many descriptors, even when unnecessary or unwarranted. Finally, the expert, having already developed mastery of the subject matter can sufficiently and succinctly describe things without requiring the use of unnecessary modifiers. We should note that we deem many of these modifiers as unnecessary to solving the design challenge because of what we observed in the discourse analysis. The reader will recall that experts used the most adverbial modifiers in the context of content words, but used the fewest when looking at the total percentage of adverbial modifiers used. The intermediate user displayed the opposite behavior, using the most general adverbial modifiers, and the least content-related adverbial modifiers. In some respects this may even suggest that the intermediate user feels uncomfortable talking about the actual content in much detail, but will try to use a great deal of detail about superficial matters. It may also be the case the intermediate user refers to things in less exact language by using pronouns. (We did not take these into consideration when looking at dependency relationships.) But the fact that novices tended to use a relatively larger number of adverbial modifiers provides some evidence that intermediate users may either be avoiding content specific details, or else are inclined to ignore those details.

Conclusions

This study has presented a set of domain-independent markers of expertise that can allow educators and researchers to recognize student learning through merely analyzing student speech. Using speech as a form of assessment certainly presents some challenges, but has the potential to introduce novel ways for understanding and predicting learning in open-ended learning environments. This ability to assess non-traditional learning should help open to door to more widespread adoption of experiential learning practices, and an associated increase in 21st century competencies.

Thus far our work has been exploratory in nature. We performed in-depth analysis on a small sample size in order to better inform the types of features that we need to be looking for in future work. Accordingly, one of the main areas for future research is to replicate this type of study on a much larger scale. Doing so will allow us to more rigorously validate the importance of features that were rendered statistically insignificant in this study.

In addition to scaling to a larger population, our current research efforts involve mining the features identified in this study in longitudinal research on a cohort of students. This study of student learning progressions will help us better identify learning pathways that students follow and the final trajectories that these student realize when follow those pathways. By better understanding and generalizing different student approaches to learning, education researchers will be better equipped to design transformational technology that empowers both teachers and students.

Finally, as noted in opening sections of this paper, this research falls into a larger body of work that is leveraging multi-modal learning data to assess student knowledge development in constructionist learning environments. At present we are analyzing speech, drawings, sentiment, collaboration, engagement and actions as a way to build a class of natural and holistic assessments.

References

Harvard Inquirer. http://www.wjh.harvard.edu/~inquirer/spreadsheet_guide.htm

Linguistic Inquiry and Word Count http://liwc.net/liwcdescription.php

Ameshi S. and Conati C.(2007). Unsupervised and Supervised Machine Learning in User Modeling for Intelligent Learning Environments. *Proceedings of the 2007 International Conference on Intelliget User Interfaces*, p.72-81.

Anaya, A. and Boticario, J. 2009. A Data Mining Approach to Reveal Representative Collaboration Indicators in Open Collaboration Frameworks. *In Proceedings of the 2nd International Conference on Educational Data Mining* (Jul. 1-3, 2009) Pages 210-219.

Beck, J. and Woolf, B. P. 2000. High-Level Student Modeling with Machine Learning. *In Proceedings of the 5th international Conference on intelligent Tutoring Systems* (June 19 - 23, 2000).

Beck, J. E. and Sison, J. 2006. Using Knowledge Tracing in a Noisy Environment to Measure Student Reading Proficiencies. *Int. J. Artif. Intell*. Ed. 16, 2 (Apr. 2006), 129-143.

Boersma, P., and Weenink, D. 2010. Praat: doing phonetics by computer [Computer program]. Version 5.2.03, http://www.praat.org/. Design Studies 19: 431-453

Brown, B. A., and Spang, E. 2008, Double talk: Synthesizing everyday and science language in the classroom. *Science Education,* 92: 708–732.

Brusilovsky, P. (1999). Adaptive and Intelligent Technologies for Web-based Education. Special Issue on Intelligent Systems and Teleteaching. Kunstliche Intelligenz.

Chen, Y., Liu, C., Lee, C., and Chang, T. 2010, "An Unsupervised Automated Essay Scoring System," *Intelligent Systems, IEEE* , vol.25, no.5, pp.61-67, Sept.-Oct. 2010

Chi, M., VanLehn, K., Litman, D., and Jordan, P. 2010. Inducing Effective Pedagogical Strategies Using Learning Context Features. *In: Proc. of the 18th Int. Conference on User Modeling*.

Clements, D. 1988 and Nastassi, B. (1988). Social and cognitive interactions in educational computer environments. *American Educational Research Journal*, 25(1), 87-106

Conati, C. and Maclaren, H. 2009. Empirically building and evaluating a probabilistic model of user affect. U*ser Modeling and User-Adapted Interaction* 19, 3 (Aug. 2009), 267-303. DOI= http://dx.doi.org/10.1007/s11257-009-9062-8

Craig, S. D., D'Mello,S., Witherspoon, A. and Graesser, A. 2008. 'Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive-affective states during learning', *Cognition & Emotion*, 22: 5, 777 — 788, First published on: 07 December 2007 (iFirst)

Dewey, J. (1902). *The school and society*. Chicago: The University of Chicago Press.

D'Mello, S. K., Craig, S. D., Witherspoon, A., Mcdaniel, B., and Graesser, A. 2008. Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted*

Interaction 18, 1-2 (Feb. 2008), 45-80.

Dringus, L. P. and Ellis, T. 2005. Using data mining as a strategy for assessing asynchronous discussion forums. *Comput. Educ*. 45, 1 (Aug. 2005), 141-160.

Forbes-Riley, K., and Litman, D. 2010. Metacognition and Learning in Spoken Dialogue Computer Tutoring. *Proceedings 10th International Conference on Intelligent Tutoring Systems (ITS),* Pittsburgh, PA.

Forbes-Riley, K., Rotaru, M., and Litman, J. 2009. The Relative Impact of Student Affect on Performance Models in a Spoken Dialogue Tutoring System. *User Modeling and User-Adapted Interaction (Special Issue on Affective Modeling and Adaptation)*, 18(1-2), February, pages 11-43.

Freire, P. (1970). *Pedagogia do Oprimido* (17 ed.). Rio de Janeiro: Paz e Terra.

Freudenthal, H. (1973). Mathematics as an educational task. Dordrecht: Reidel.

Kafai, Y, Peppler, K., Chapman, R. 2009. The Computer Clubhouse: Constructionism and Creativity in Youth Communities. TEC series, Teachers College Press.

Klein, D and Manning, C. 2003. Accurate Unlexicalized Parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Liscombe, J., Hirschberg, J., and Venditti, J. 2005. Detecting Certainness in Spoken Tutorial Dialogues. In *Proceedings of Interspeech 2005—Eurospeech*, Lisbon, Portugal.

Litman, D., Moore, J., Dzikovska, M., and Farrow. E. 2009. Using Natural Language Processing to Analyze Tutorial Dialogue Corpora Across Domains and Modalities. *Proceedings 14th International Conference on Artificial Intelligence in Education (AIED)*, Brighton, UK, July.

Litman, D., and Forbes-Riley, K. 2009. Spoken Tutorial Dialogue and the Feeling of Another's Knowing. Proceedings 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), London, UK, September.

Manning, C. and Schütze, H. 1999. Foundations of Statistical Natural Language Processing. MIT Press.

Montessori, M. (1964). *The advanced Montessori method*. Cambridge, Mass.,: R. Bentley.

Montessori, M. (1965). *Spontaneous activity in education*. New York,: Schocken Books.

Papert, S. (1980). *Mindstorms : children, computers, and powerful ideas*. New York: Basic Books.

Pea, R.. (1987). Programming and problem-solving: Children's experiences with Logo. In T. O'Shea & E. Scanlon (Eds.), Educational computing (An Open University Reader). London: John Wiley & Sons.

Purandare, A., and Litman. D. 2008. Content-Learning Correlations in Spoken Tutoring Dialogs at Word, Turn and Discourse Levels. *Proceedings 21st International FLAIRS Conference*, Coconut Grove, Florida, May.

Rus, V., Lintean, M.,, and Azevedo, R.. 2009. Automatic Detection of Student Mental Models During Prior Knowledge Activation in MetaTutor. In *Proceedings of the 2nd International Conference on Educational Data Mining* (Jul. 1-3, 2009). Pages 161-170

U.S. Department of Education, Office of Educational Technology (2010). Transforming American Education: Learning Powered by Technology. *National Education Technology Plan 2010*. Washington, DC.

von Glasersfeld, E. (1984). An Introduction to Radical Constructivism. In *P. Watzlawick* (Ed.), The Invented Reality. New York: Norton.