

Culture in the Classroom: Challenges for Assessment in Ill-Defined Domains

Amy Ogan
Human Computer
Interaction Institute,
Carnegie Mellon
University
aao@andrew.cmu.edu

Vincent Alevan
Human Computer
Interaction Institute,
Carnegie Mellon
University
alevan@cs.cmu.edu

Christopher Jones
Modern Languages,
Carnegie Mellon
University
cjones@andrew.cmu.edu

Abstract. While cognitive tutoring has been successful in well-defined domains, much less work has been completed in cognitive tutoring in ill-defined domains. Skill assessment is one of the difficult challenges that must be solved before cognitive tutoring can become a reality in these domains. Assessment formats commonly used by instructors in ill-defined domains do not always have a correct answer or follow a finite set of solution paths. This necessitates clever ways of evaluating student responses to open-ended questions in which variability in both approach and final solution is a given. Additionally, there may be difficulty in covering the full problem space with a single assessment. In this paper we describe complementary assessments developed to evaluate a tutoring system in a classroom study. Critical analysis questions which are easily machine-interpretable and writing samples that address more nuanced understanding assess different skills that will be incorporated into a set of intelligent tutors that cover the cultural learning domain.

Keywords: ill-defined domains, assessment, intercultural competence

ASSESSMENT IN ILL-DEFINED DOMAINS

Cognitive tutors, a type of computer-based instruction that compares student actions to a model of correct and incorrect behavior and provides context-sensitive feedback and problem selection, have been effective at increasing student learning in real-world settings in domains with well-structured problems such as math and science domains. For example, use of the Algebra-1 Cognitive Tutor has been shown to increase student learning by at least one standard deviation over traditional classroom instruction [Koedinger, 1997]. The methodology which makes cognitive tutoring so successful in these domains, however, has made cognitive tutoring less practical in ill-defined domains. An ill-defined knowledge domain is one in which the following two properties hold: 1) each case of knowledge application involves the concurrent interaction of multiple conceptual structures such as schemas or organizational principles, each of which is individually complex, and 2) the interaction of these structures varies substantially across cases nominally of the same type, producing across-case irregularity [Spiro, 1995]. These properties push the limits of traditional cognitive tutoring methodologies, such as representing knowledge in the form of production rule systems. Although they present unique problems, ill-defined domains stand to benefit from intelligent tutoring as much as well-defined domains, because in these areas students profit greatly from tutoring strengths such as working at their own pace, receiving immediate feedback, and having privacy to practice skills before performing in front of the class.

In particular, skill assessment is one of the difficult challenges that that must be solved before cognitive tutoring can become a reality in ill-defined domains. Assessment is hard in these domains for several reasons. There may be difficulty in covering the full problem space. Ill-defined domains cannot be described in a finite set of production rules [Scandura, 2003]. This set of production rules enables assessment in well-defined domains to cover the complete set of skills specified as learning objectives. Ill-defined problems may have so many sub-problems that different people could focus on wholly different issues and still come up with viable solutions. A variety of problem types must be used to ensure that all skills are covered. Therefore we must choose the appropriate response types to elicit

useful student answers for each skill. Traditional tutors use standard response mechanisms such as multiple choice, or require numeric responses in quantitative domains. Much mathematical knowledge, for example, is well developed and linked to performance, and it is relatively simple to identify the correct answer for a range of questions requiring the understanding of basic concepts [Legree, 2005]. It is possible to leverage these techniques in ill-defined domains if we separate the skills of the domain which can be described as discrete or rule-based knowledge components from those which require an open, unconstrained environment to demonstrate in an ecologically valid manner.

The most appropriate way to identify the extent of student knowledge for some skills in an ill-defined domain, however, may be through open-ended responses. In similar situations medical professionals do not collect exactly the same data and do not follow the same paths of thought [Charlin, 2004]. The multiple choice test format requires a unique right solution to problems, a correct conclusive answer to give to data specified in the problem. Assessment formats commonly used by instructors in ill-defined domains do not require a single correct answer nor follow a solution path with a limited number of alternatives. For example, a student asked to synthesize study of immigration issues in France could draw on a wide variety of historical, journalistic or even experiential data in formulating a valid response. This issue of format is a problem because machine interpretation of correctness, which is a strength of cognitive tutoring, is difficult to execute on open-ended or verbal responses. It is simple to programmatically determine whether student responses to closed question types such as multiple choice formats are correct. Even brief short answers with a limited range of possible correct answers may be reviewed with simple natural language processing. Item construction in formal, well-defined knowledge domains can easily incorporate general knowledge and expertise, and item revision based on the use of item statistics can maximize characteristics such as reliability and validity [Legree, 2005]. Limitations in these techniques, however, necessitate clever ways of evaluating student responses to open-ended questions in which variability in both approach and final solution is a given. Even when this is successfully accomplished, the grading of open-ended responses may still be construed as subjective and also sets limits on standardization across students.

In this paper we describe an approach to the difficulties in assessment associated with ill-defined domains. We discuss two distinct types of assessment, critical analysis questions that are easily machine-interpretable and writing samples from a discussion board that demonstrate deeper knowledge, that are utilized to evaluate a tutoring system while conducting a study in a real classroom setting. We then describe how these assessment formats will be incorporated into a system of complementary intelligent tutors.

INTERCULTURAL COMPETENCE

The domain in which we focus our work, intercultural competence, is particularly ill-defined and difficult to assess in an objective manner [Kramsch, 1993]. Although there is no current consensus on the exact definition of this relatively new term, intercultural competence in general refers to the abilities to “reflect and gain insight on native perspectives, opinions, and values; reflect critically and engage with *otherness*” [Scarino, 2000]. In an attempt to include these higher-order skills in every language classroom, the American Council on the Teaching of Foreign Languages (ACTFL) has set forth a number of content standards regarding what students should know and be able to do in the document *Standards for Foreign Language Learning in the 21st Century* [ACTFL, 1996]. A significant number of these standards focus on cultural understanding, e.g. Standard 3.2: “Students acquire information and recognize the distinctive viewpoints that are only available through the foreign language and its cultures”. While less familiar than the four skills of reading, writing, listening and speaking traditionally associated with language acquisition, cultural learning is a critical part of the second language classroom and is the foundation upon which all other language skills are based [Kramsch, 1993].

SKILLS

Student assessment refers to the documentation of whether the learning objectives of a domain were achieved. To define the learning objectives in a discipline, first we must specify the skills that are involved in domain mastery and then we can determine how best to assess them. While experts are not in complete agreement, the literature on intercultural competence suggests a culture-general skill set (e.g. Kramsch, 1999) which involves several distinct proficiencies that students are expected to master:

- 1) Critical analysis of culture: This refers to the ability to relate products, practices, and behaviors to cultural attitudes and values. Also to the ability to notice elements of a foreign culture

- 2) Cultural perspective-taking skills: The ability to look at and express cultural elements from outside one's own cultural space
- 3) Taking an ethnorelative stance toward culture: Demonstrating tolerance for cultural differences

For our tutoring system and assessment, we will focus on the first two skills, i.e. a critical analysis of culture and cultural perspective-taking, which rely less on stable personality traits such as openness than does the third, long-term skill of developing tolerance for cultural difference. These skill definitions are abstract and it is questionable whether they can be described with formalizable rules that always produce correct answers. Demonstration of these skills is open-ended and the knowledge may even be co-constructed, with consensus not always being reached. These skills are distinct from one another, and they necessitate different methodologies to assess whether students have acquired them. In particular, the second skill is difficult to assess with current cognitive tutor methodology which describes domains as a complete set of production rules covering the problem space that students do or do not know at any stage in their learning trajectory.

ASSESSMENT APPROACHES

In a typical classroom, teachers may assess these skills through various methods such as observation of behavior, writing samples, developing portfolios of written and oral work, or simply through classroom discussion, all of which result in "an analyzed profile" of the student [Tyler, 1949]. It is suggested that it may be helpful for instructors to employ triangulation, the use of a set of these assessment techniques to "cross check" competence to provide evidence of the greater validity of the various measures used [Deardorff, 2004]. These techniques are difficult to reproduce with traditional cognitive tutoring methods. Prior assessment in computer-based instruction in the area of cultural competence has developed along two distinct trajectories, neither of which covers the full scope of skills involved in cultural learning. Several quantitative studies have focused solely on knowledge that can be objectively assessed, such as facts learned through instructional video. In a controlled study presented by Herron [2000], students watched French video as an 'advance organizational tool' to gather information that was then used in answering cultural assessment questions. This type of evaluation measures whether the student has learned discrete knowledge components about the culture. While these questions target practices and products, such as the question "What do French people typically eat for a midday snack?", they are too narrow to address more controversial ideas such as the core values and attitudes of the target culture, and ignore the skills involved in more interactive tasks such as cultural perspective taking. It is tempting to reduce the complexity of a domain by narrowing the instructional scope to objectively assessable facts. However, if real complexities exist and their mastery is important, this reduction is an inappropriate oversimplification and can lead to conceptual misunderstanding [Spiro, 1995].

On the other hand, those who take an extreme constructivist approach believe that learning is a personal interpretation of the world and therefore objective measurement is nearly impossible [Merril, 1991]. Some work on computer-based pedagogical approaches has been done following these theories, such as large-scale, resource intensive projects like "Cultura" or "A la rencontre de Philippe" [Furstenberg, 2001]. In a different take on cultural instruction, these projects invite students to construct their own knowledge of core values and attitudes in a different culture. This is accomplished by students, largely without instructor intervention, answering questionnaires about their own culture and communicating with a French classroom to evaluate the authentic cultural descriptions provided by the questionnaires from the other class. This is an extremely motivating and deeply informative method of cross-cultural learning, but presents difficulties in scaling up in implementation due to the great deal of overhead involved with linking classes across continents. To avoid this overhead, a comparable type of constructive learning is done using similar existing material in a UC Berkeley study, *Teaching Text and Context Through Multimedia* [Kramsch, 1999]. More importantly, both of these curriculum approaches lack assessment of student learning beyond surface-level self-report by students of perceived quality of learning and motivation, e.g. a scale-based response to "Do you feel like you learned a lot?" Few empirical or descriptive classroom studies have been completed on work that subscribes to these theories.

SYSTEM DESIGN

We incorporated these skills into a tutor we developed to teach this domain. The Cognitive Tutor Authoring Tools (CTAT) are a set of tools built by researchers at Carnegie Mellon University and Worcester Polytechnic Institute that facilitate rapid development of intelligent tutors (for a full description of the CTAT system, see [Aleven, in press]). Using these tools, developers can build tutors

that use model tracing to compare student action to a model of correct and incorrect steps and provide individualized hints and feedback. One advantage of CTAT is the ability to build "example-tracing tutors", which allow authors to visualize and employ the processes found in full intelligent tutors without complex AI programming. Building an example-tracing tutor is accomplished first by designing a Flash interface with specialized interface widgets that communicate with the example-tracing system. Next, the author uses the interface to demonstrate the correct and incorrect steps students may take when completing a problem, and the provided Behavior Recorder tool automatically creates a graph of the demonstrated behaviors. After the author generalizes the problem-solving steps recorded in the graph and attaches hint and feedback messages, the graph is ready to be used for tutoring. These tutors can then be easily deployed to the Internet. One of the new Flash widgets developed especially for the tutor described here is a video component that logs all actions performed on the video (e.g., play, pause, "rewind").

To find the cultural content we present in the tutor, we requested suggestions of feature films demonstrating cultural attitudes or behaviors from French instructors who utilize such authentic material in their classroom. Instructors were asked to provide brief descriptions of appropriate scenes that fall under several chosen themes of the French culture. The films were documented to find one- to two-minute video clips that present cultural information and afford a natural moment to pause and ask students to make a prediction about the events of the second half of the clip. Clips with these "teachable moments" were chosen such that the prediction is dependent on cultural knowledge and not the narrative content of the film.

Using the CTAT tools, we created a tutor to display the cultural film clips and prompt students with questions to assist them in noticing cultural features of the film. First, students can review details about the film they are about to see, including film credits, a brief plot summary of the movie, and a paragraph of context for the clip they will be viewing. As an illustration, one video used was *Monsieur Ibrahim*, a film from 2003 that deals with issues of cultures clashing among immigrants in Paris. The next screen presents the video where students view the first half of the clip. In the scene from *Monsieur Ibrahim*, a boy Moses walks into a neighborhood convenience store and continues a conversation with the elderly proprietor about the etymology of their names. The proprietor gently explains to Moses that he is Muslim, and not Arab. The boy asks, "Then why does my father say 'Go to the Arab's?'". At this point the video pauses and students respond to a set of questions (see Fig.1) in which they: 1) predict the next event that will occur in the clip from a drop-down menu, 2) provide a more extensive natural language explanation of their choice, and 3) state what they believe an appropriate response to the situation might be in their own culture. The first question is presented in a menu format to suggest several appropriate responses and lightly constrain students to actual cultural possibilities. For this video clip, two of the possible responses include 'The neighbors don't take the time to get to know me' and 'Anyone in this profession is labeled an Arab'. Unlike a traditional cognitive tutor, any reply is accepted in the drop-down menu. In the second question, students are given space to explore their hypothesis and provide evidence from the clip or their knowledge of the French culture to support their reasoning. They may rewind and review the video up to this point as often as they like. The tutor does not provide feedback.

When students have provided answers to all three questions, the succeeding portion of the video clip plays and the ensuing cultural event is revealed. At the end of the video clip, a second set of questions appears that asks students to make an assessment of whether they were correct in their prediction or not. If they were not correct, they are asked to revise their prediction. Finally, students are given a set of characterization questions about the clip that may be answered with 'true', 'false', or 'maybe'. These questions are tutored with hints and error messages, as well as success messages that provide a summarization of the evidence for a correct answer. For example, one question states, "Monsieur Ibrahim lives in an isolated immigrant community", which would be correctly answered with 'false'. At this point, students may rewind and review the full clip as often as they like.

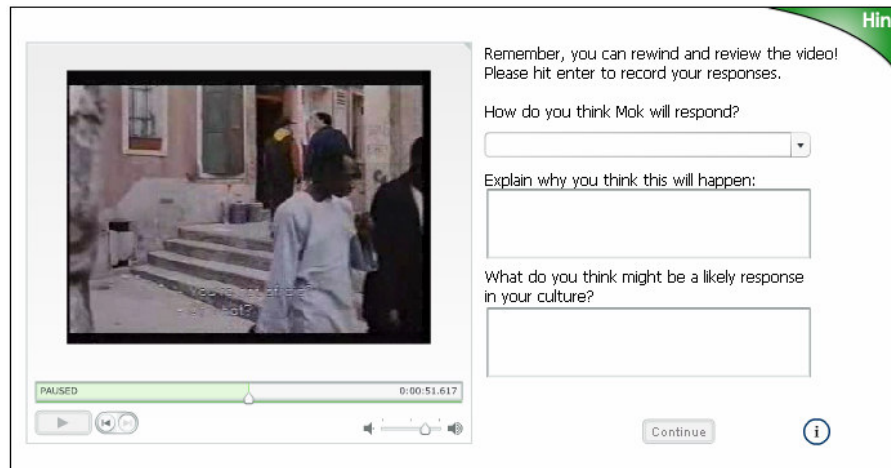


Figure 1: Screenshot of the initial set of tutor questions

Following the video, students participate in an asynchronous online class discussion where they use what they have seen to reflect on cultural differences, similarities or assumptions the class has about the French culture, and ask questions about the meaning of behaviors they have seen in the clips. The prompt that students see for *Monsieur Ibrahim* is, “Post at least one original post and one reply with questions, analysis, or other thoughts about the immigration issues in France you've seen! Think about what are the racial and ethnic stereotypes in France that you have seen depicted in this film to get started.” The responses students provide regarding their personal experiences spark interesting comparisons with valuable information from multiple cultures. The writing samples posted to the discussion board form one part of the assessment for evaluating student improvement using the tutor, along with a post test that measures the cultural knowledge that students have acquired from the components in the film clip.

STUDY DESIGN

A full study was recently completed that used the assessments we developed to test the tutor. Data collection was done automatically through the Pittsburgh Science of Learning Center (PSLC) and Open Learning Initiative infrastructures. The PSLC (<http://www/learnlab.org>) is a “national resource for learning research” that includes authoring tools for online courses and support for running “in-vivo” learning experiments, that is, learning experiments in real educational settings, as opposed to the lab. Elementary French I Online is a PSLC-targeted course that is being developed as part of the Open Learning Initiative at Carnegie Mellon University, which is a program that provides a collection of “openly available and free online courses and course materials.” This combination of web-based distribution and experimental rigor allows a number of innovative instructional components to be designed, evaluated, and deployed as part of the French Online Course.

The study was conducted in the Carnegie Mellon and University of Pittsburgh Elementary I French Online classes to test the hypothesis that attention-focusing techniques increase cultural learning in an online environment. Thirty-two students were randomly assigned within each class to either an experimental group that used the system described above or a control group that viewed the same video clips without intervention. The control group watched the clips as often as they liked, but without the attention-focusing caused by pausing the video at a critical moment and responding to the three sets of questions about the content. In each of three assignments spaced throughout the semester, groups were assigned two film clips to watch and discuss in place of a typical cultural reading and writing assignment for the class. Film clips corresponded to cultural themes that were explored in the classroom, such as immigration, employment, and education in the French culture. The materials for each assignment were linked from a single course webpage that provided students with background information about the theme. According to the format of the rest of the online course material, students worked through each assignment at their own pace. This information was written in English and was adapted from the cultural course materials that the tutor replaced. Students in the control condition performed exactly the same sequence of instruction, except that the “tutors” linked off their course materials simply presented the video clips without the attention-focusing techniques.

ASSESSMENT

Prior to the first assignment, demographic data was collected for each student. In each assignment, students first took a pre-test that explored their knowledge of basic information about the theme. For example, one question related to the theme of education asked ‘The baccalauréat is...’ (the baccalauréat is a French examination similar in some respects to the American SAT). The response was open-ended and allowed students to demonstrate all of their knowledge about each topic, while avoiding the testing threat of influencing what they notice in the film clip. Students completed tutor sessions with video clips from two different movies relating to the lesson theme. All actions that students performed in the tutor were recorded. After each tutor, students accessed the discussion board for that video clip and were required to post at least twice and encouraged to return to the discussion board to read other replies and to post again. All discussion posts were stored in a database, along with a link to the original post if it was a reply. Finally, students completed a post-test for the assignment that included the analytical questions described in the next section along with self-report on scales that measure items related to intercultural competence such as world-mindedness and attributional confidence [Sampson, 1957]. Our work on assessment is currently focused on the analytical post-test questions and the discussion board writing samples. This mixed assessment methodology addresses a set of concerns about ill-defined domains.

The analytical post-test questions are designed to be able to leverage traditional cognitive tutor techniques of assessment. The questions were developed with the help of a French citizen and language instructor through a component analysis of the cultural elements in the film clips. Cultural elements in the film were decomposed into a hierarchy that covered the main concepts in the theme. Each element was then formed into a question requiring cultural analysis that was situated within the context of the film. Each post-test covered three questions in true/false format relating to each film in the theme as well as two questions in multiple choice format that asked students to compare and contrast across films in the theme. In addition, several of the questions were followed by a short answer component which asked students to explain why they made the choice that they did. See Figure 2 for examples of the two types of questions. This type of assessment allows us to cover more completely the intercultural competence problem space by ensuring that students are assessed on their knowledge of the key cultural elements relating to each theme, specifically relating to values and attitudes that are deeper than the behavioral elements assessed in such work as the Herron study.

A multiple choice question on the theme of education:	Based on the clips you just saw, <i>Être et avoir</i> and <i>Le Péril jeune</i> are similar in EVERYTHING BUT: a) Students in both schools are expected to take responsibility for their learning b) Students in both schools are expected to obey authority c) Students in both schools are expected to have a positive attitude
A true/false question from the theme of employment:	From the factory scene in <i>Ressources humaines</i> , we know that social mobility is possible in French society.

Figure 2: Examples of post-test analytical questions

The writing samples from the discussion boards comprise the second main type of assessment. On each discussion board students were asked to post in general with cultural observations and were given one specific stimulus about the theme to help them begin, such as:

“Post at least one original post and one reply with questions, analysis, or other thoughts about the immigration issues in France you've seen! Think about what racial and ethnic stereotypes in France you have seen depicted in this film to get started.”

The writing on the discussion board addresses the limitations of closed format questions by allowing students to demonstrate intercultural competence skills in a less constrained, and therefore more ecologically valid, manner. To evaluate the quality of writing in the discussion posts, we hand code the responses into three major categories of good and bad cultural writing that were developed for a similar intercultural competence writing task [Steglitz, 1993]. In a general sense, a category 1 response gives no indication that the author recognizes the role of culture in describing behaviors or values of a culturally different other and may take the form of judgments or advice. Category 2 responses acknowledge that there are general cultural influences that may be the cause of what they notice, while category 3 responses relate behaviors and values to specific cultural phenomena, such as power

relationships or individualism. These student writing samples can be used to both measure learning outcomes directly and monitor behaviors that may lead to future learning. For example, students may demonstrate knowledge of a correct French perspective as well as demonstrate the ability to take a different perspective, even if the particular one that they take is not valid from the French point of view. This writing also serves as a record or portfolio of fine-grained learning that may assist in determining progress on intercultural competence skills over the course of the semester.

RESULTS AND DISCUSSION

ASSESSMENT VALUE

The analytical post-test questions have the ability to confirm or disprove an experimental hypothesis. In a preliminary analysis of the post-test question data from the French Online study, we found that 9 of the top 10 scoring students are from the experimental condition. While students in the control condition scored an average of only 64% on the post-tests, students in the experimental condition scored an average of 72%. These results partially confirm the hypothesis of the experiment, that attention-focusing techniques increase cultural learning in an online environment. To fully validate this result across cultural skills, we must also look at the other assessment type. In an average taken for each student of scores on Steglitz's scale for all discussion posts, the experimental condition once again outperforms the control condition with an average of 1.61 out of 3 to an average of 1.36 out of 3.

Although both assessment types show an advantage for the experimental condition, we argue that the analytical questions and the writing samples each add a unique value to the assessment of skills in this domain. The analytical post-test questions are designed to assess whether students have understood the knowledge components picked up as cultural elements in the video. They determine if the student can accurately characterize the French perspective. Because students may not express objective knowledge while writing in the discussion, the analytical questions confirm that students have learned these key cultural elements. For example, the following discussion post is written as an opinion without evidence, taken from a personal perspective. It does not demonstrate that the student has achieved deep cultural understanding or is able to approach the events in the film from a perspective that culture might have an influence on behavior:

“I agree with Eiben's point. Who are we to judge people? I believe that we should not judge people from what we hear or what they look like. Plus it is the parent's duty to raise the child to better understand the world around them. If the boy's father had not mentioned to him that the store clerk was Arabe he would not have mention it to the clerk.”

This same student, however, received an 87% average on the analytical questions. While he did not demonstrate perspective taking skills in his writing, he has shown that he was able to extract cultural information from the film clips.

The discussion writing samples can give us insight into the other skills that make up intercultural competence. They allow a deeper investigation of whether students can actually do perspective-taking. With this method of assessment as well, we see differences in the skills that students demonstrate. One of the five lowest scoring students on the analytical questions, who received a 47% average, was able to articulate various French perspectives in immigrant communities and present evidence from the film in his writing:

“The two individuals walking through the neighborhood seemed to represent a dichotomy in French culture: the ignorant one who was afraid of the black people and the accepting one who lived amongst a scene of racially and ethnically diverse people. This shows that people in the French culture are not all on the same "wavelength." Hopefully, accepted diverse communities are the future of France and if they are, maybe the United States can take a hint from France in this particular sense.”

This student may not have noticed all of the specific cultural elements in the film clip to answer analytical questions, but was able to provide an overview of the main cultural point of the clip from outside of his own personal beliefs.

In a preliminary analysis of the data, it does appear that these assessment techniques measure different skills. In a first round of data coding, each post on the discussion boards was given a score according to the categories developed by Steglitz and then an average score was calculated for each student across all of their posts. Additionally, each student was given an average score across all of

their post test responses. No significant correlation was found between the two average scores (adjusted r squared = .066, $p > .15$). Separately, each type of assessment could provide insight into different student misconceptions. When analyzed together, the discussion board could be used as a process measure to determine how and when students are accurately able to answer the post-test questions.

APPLICATION TO TUTORING

Beyond being an effective measure of student learning, these assessment methodologies also have great potential for incorporation into cognitive tutors. The analytical questions are more than simply a post-test, but can be used within a tutor as a problem solving exercise following the presentation of cultural material. While evaluating writing samples is not common in existing tutoring methodology, we are currently training machine learning algorithms to identify critical features of the discussion writing that can predict whether it is a good example of cultural writing, using the categories developed by Steglitz as criteria. This scale has been validated using human raters, but no work has been done on machine interpretation of the writing samples. Our current work focuses on identifying and predicting intermediate features that are relatively easy to detect through machine learning that may predict student performance levels on the higher-level cultural proficiency categories. For example, one feature with significant predicting power seems to be the extent of argumentation. Students may write inferences, which include evidence from the film and a conclusion; opinions, which are conclusions with no evidence; or recall of information, which is evidence without conclusions about what they have seen. Using the kappa statistic, a measure from 0 to 1 that gauges reliability between raters (with 0 being no agreement and 1 complete agreement), we are currently able to achieve kappa values between .5 and .9 for predicting the intermediate features we have identified, using a set of machine learning algorithms such as Support Vector Machines (SVM) that are known to work well with text. Employing these intermediate features, we have improved from a low baseline kappa value of .01 using just the words in the post as features to a kappa value of .4 using the current framework with the intermediate features as predictors. Work continues in this area to improve prediction of both the intermediate features and to identify features that better predict the high-level categories. With this machine interpretation, we have the eventual goal of developing a tutor for the discussion board that will give on-line, non-deterministic feedback to students on how they might improve their cultural writing, based on the intermediate features that more conducive to guided feedback.

Cognitive tutors in this domain can benefit from using both types of assessment in conjunction with each other. It may be relatively easy to determine whether the discussion writing is of good general quality by using features that are less complicated to extract from writing, such as being on-topic or attempting to make inferences from evidence. It is a much more difficult natural language problem to know whether the inferences expressed are correct French perspectives, because there are so many ways of communication in free response. In this case, a correlation with student responses to the analytical questions may give a complete picture of the current state of student understanding. Together, these assessments speak to the initial concerns we describe for assessment in ill-defined domains. The writing samples allow students to explore the problem space with an open-ended, unconstrained format that allow for a variety of solution paths that address any number of sub-goals. Machine interpretation of correctness is easily accomplished with the analytical post-test questions, and we are working towards a procedure for evaluating the writing samples automatically in a way that is conducive to the cognitive tutoring goals of tracking student knowledge and providing context-sensitive feedback. We have made steps toward covering the full problem space of the domain by exploring multiple types of assessment that identify different types of student misconceptions and assess disparate skills which cannot be assessed in the same way. In addition, this assessment methodology has the ability to be applied to other ill-defined domains with similar properties. For example, in art or literary criticism, students also must have a strong foundation in the key concepts of the domain. Additionally, in such domains in which dialogue is a critical component, they must be able to express opinions and support them with evidence. We believe that this combined assessment methodology will be effective for intelligent tutoring in the domain of intercultural competence, and has the potential to extend to tutors in other ill-defined domains.

ACKNOWLEDGEMENTS

We would like to thank the Pittsburgh Science of Learning Center for funding this work, as well as the Cognitive Tutor Authoring Tools team at Carnegie Mellon University for providing incomparable assistance with developing specialized tools for the tutor. Michael West and Cary Campbell continue to accommodate us as instructors for the French Online courses.

REFERENCES

- ACTFL (American Council on the Teaching of Foreign Languages) (1996). Standards for Foreign Language Learning: Preparing for the 21st Century. New York: ACTFL.
- Aleven, V., Sewall, J., McLaren, B. M., & Koedinger, K. R. (in press). Rapid authoring of intelligent tutors for real-world and experimental use. Accepted for presentation at the 6th IEEE International Conference on Advanced Learning Technologies (ICALT 2006).
- Charlin, B. & van der Vleuten, C. (2004). *Standardized Assessment of Reasoning in Contexts of Uncertainty: The Script Concordance Approach*. Evaluation of the Health Professions. 28 (3), pp.304-319.
- Deardorff, D. (2004). The Identification and Assessment of Intercultural Competence as a Student Outcome of Internationalization at Institutions of Higher Education in the United States. Unpublished dissertation available at <http://www.lib.ncsu.edu/theses/available/etd-06162004-000223/unrestricted/etd.pdf>
- Furstenberg, G., et al. (2001). Giving a Virtual Voice to the Silent Language of Culture: The Cultura Project. *Language Learning & Technology*, 5(1), 55-102.
- Herron, C., & Dubreil S. (2000). Using Instructional Video to Teach Culture to Beginning Foreign Language Students. *CALICO*, 17(3), 395-429.
- Koedinger, K. R.; Anderson, J. R.; Hadley, W. H.; and Mark, M. A. (1997). Intelligent Tutoring Goes to School in the Big City. *Journal of Artificial Intelligence in Education* 8(1): 30-43.
- Kramsch, C., & Anderson, R. (1999). Teaching Text and Context through Multimedia. *Language Learning & Technology*, 2(2), 31-42.
- Kramsch, C. (1993) Context and Culture in Language Teaching. Hong Kong: Oxford University Press.
- Legree, P. J., Psotka, J. & Tremble, T. (2005). Applying Consensus Based Measurement to the Assessment of Emerging Domains (ARI Technical Report No. 1153). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.
- Merrill, M. D. (1991). Constructivism and instructional design. *Educational Technology*, May, 45-53.
- Sampson, D.L., & Smith, H.P. (1957). A scale to measure world-minded attitudes. *Journal of Social Psychology*, 45, 99-106.
- Scandura, J. (2003). Domain Specific Structural Analysis for Intelligent Tutoring Systems: Automatable Representation of Declarative, Procedural and Model-Based Knowledge with Relationships to Software Engineering. *Tech, Inst, Cognition and Learning*, Vol. 1, pp 7-57.
- Scarino, A. (2000). 'The Neglected Goals of Language Learning', *Babel*, 3 (34), Summer, pp 4-11.
- Spiro, R. J., Feltovich, P. J., Jacobson, M. J., & Coulson, R. L. (1995). Cognitive Flexibility, constructivism, and hypertext: Random access instruction for advance knowledge acquisition in ill-structured domains. In P. Stele, & J. Gale, *Constructivism in education* (pp. 85-108). Hillsdale, NJ: Erlbaum.
- Steglitz, I. (1993). The Identification and Assessment of Intercultural Competence as a Student Outcome of Internationalization at Institutions of Higher Education in the United States. Unpublished dissertation available at <http://www.lib.ncsu.edu/theses/available/etd-06162004-000223/unrestricted/etd.pdf>
- Tyler, R.W. (1949). Basic principles of curriculum and instruction. Chicago: University of Chicago Press.