The Evolution of Assessment: Learning about Culture from a Serious Game

Matthew J. Hays[1], Amy Ogan[2], H. Chad Lane[1]

[1] Institute for Creative Technologies, University of Southern California, 13274 Fiji Way, Marina del Rey CA 90292, USA
[2] Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh PA 15213, USA
hays@ict.usc.edu, aeo@andrew.cmu.edu, lane@ict.usc.edu

Abstract. In ill-defined domains, properly assessing learning is, itself, an ill-defined problem. Over the last several years, the domain of interest to us has been teaching Americans about Iraqi business culture via a serious-game-based practice environment. We describe this system and the various measures we used in a series of studies to assess its ability to teach. As subsequent studies identified the limits of each measure, we selected additional measures that would let us better understand what and how people were learning, using Bloom's revised taxonomy as a guide. We relate these and other lessons we learned in the process of refining our solution to this ill-defined problem.

Keywords: learning, technology, assessment, measurement, ill-defined domain, culture, serious game

1 Introduction

As societies and their economic and humanitarian transactions have become more globalized, cross-cultural negotiation has emerged as an important ill-defined domain. Culture often dramatically affects people's expectations when they interact with others. These effects can be exacerbated because they are often implicit. That is, the role culture plays becomes salient only when expectations are violated—and one may not be able to identify cultural differences as the cause of interpersonal difficulty [2].

With several collaborators, we have developed a cultural training system called BiLAT. BiLAT is a serious-game-based learning environment that is designed to teach the preparation for and execution of meetings in a cross-cultural context [12, 13]. The immersive approach [14] and focus on practice [10] were motivated by cognitive psychology and the instructional design literature. Elsewhere, we detail the development and implementation of BiLAT [12, 13]. The present paper provides an overview of BiLAT and an accompanying intelligent tutoring system (ITS), but focuses on our assessments of learning from BiLAT, their evolution, and the lessons we learned along the way.

2 How can BiLAT improve intercultural competence?

Rulebooks, demonstration videos, and lectures are sufficient instructional tools in many learning contexts. When well designed, these mostly passive approaches are effective at conveying facts and examples. However, competence in ill-defined domains is dependent on *contextualized* understanding; learners must determine the circumstances under which particular solutions to problems are appropriate [7, 18]. Without direct experience or live role-play, this task is very nearly impossible [19]. Unfortunately, on-the-job training is rarely an option and live role-play is costly and difficult to scale up. Moreover, even if these were viable alternatives in terms of resources, it would be difficult to ensure consistent pedagogical content and provide appropriate learning scaffolds.

Fig. 1. Meetings with a police officer (left) and a businessman (right) in BiLAT

BiLAT simulates a business meeting in which cultural awareness, adherence to expectations, and relationship building are important. It can be used as a consistent, scalable, lower-cost alternative to role-playing. Figure 1 shows the BiLAT interface, in which learners research and engage in turn-based dialogue with virtual characters by selecting actions from menus.

Success in BiLAT depends on building trust with the virtual characters before discussing potential agreements, a basic principle of Arab business culture [20]. BiLAT therefore emphasizes the timing of actions and their context of use by modeling *meeting phases*, which determine when the actions a user can choose are appropriate. For example, one generally advisable social action is discussion of children; both cultures take pride in the achievements of their young. Talking about your—or your meeting partner's—children is a good idea near the beginning of a meeting, but not while negotiating the terms of an agreement.

Learners with little experience and no external guidance might become confused about when or whether an action is generally advisable. We therefore developed an ITS to help clarify these situations and more broadly support learners through their interactions with the virtual characters. The ITS takes the form of a virtual coach that assists the learner during the meeting. After each turn, the coach decides whether [12] and how [11] to provide feedback about past actions or hints about future actions.

Designing and developing BiLAT and the ITS were extensive, complicated processes. Determining whether the two systems function together as an effective training tool has been an equally intricate process. The next section of this paper details the ways in which we measured how BiLAT and the ITS improved learners' comprehension of and competence with Iraqi business-meeting culture.

3 How can we assess intercultural competence?

The same things that make intercultural interaction difficult to train are those that make its improvement difficult to measure [21]. Is a business meeting successful as long as a mutually beneficial outcome is reached? What if the negotiations came at the cost of the business relationship, making it the last agreement those two parties will ever reach? Perhaps one partner takes from the meeting a negative opinion about all members of the other's culture; is that still a successful meeting? Without hard and fast rules, determining the complete extent of a trainee's ability cannot be accomplished solely by checking multiple-choice responses against a key. Instead, multiple measures are needed to get a complete understanding of trainees' comprehension and competence. Dozens of quantitative studies investigating the effectiveness of non-technological cross-cultural training programs, many including several measures, have been undertaken with exactly this goal [4, 19]. Selecting a subset of these measures appropriate to evaluating learning from BiLAT required several iterations of empirical research. We also used Bloom's revised taxonomy of educational objectives as a framework for our decisions [1, 5]. This taxonomy is a widely accepted hierarchical classification that defines levels of learning, activities that promote learning at each level, and assessments of learning at each level. The rest of this section describes the measures we selected and how we used them to gauge BiLAT's effectiveness as an educational tool.

3.1 Measuring remembering and understanding: a situational judgment test

In Bloom's revised taxonomy, the two most basic levels of learning are remembering and understanding. *Remembering* is the ability to recall or recognize information in the format in which it was learned (i.e., without requiring transfer or application). Students can demonstrate remembering by providing definitions for key terms or labeling components of a system. *Understanding* can be thought of as remembering that has been freed from its original format. Students can demonstrate understanding by summarizing or generating additional examples of a category.

Situational judgment tests (SJTs) are appropriate for measuring remembering and understanding in ill-defined domains [17]. In a common SJT format, learners read several scenarios that describe various problems related to the training domain. Each scenario is accompanied by potential solutions to which learners provide Likert-scale ratings of advisability (i.e., 1 = "very unadvisable"; 10 = "very advisable") [6]. Responses are generated by several subject-matter experts (SMEs), who have substantial familiarity with the training domain. The consensus of the SMEs' answers

is the standard against which trainees' scores are compared [3]. The greater the correlation between a trainee and the SMEs, the greater the trainee's understanding.

Assessments of remembering and understanding must be tailored specifically to the content of instruction. Otherwise, the assessments begin to measure the ability o apply or transfer knowledge, which are at higher levels of Bloom's taxonomy. In the case of assessing learning from BiLAT, we needed measures that explicitly and exclusively addressed Iraqi business culture. A literature search revealed many measures of intercultural competence [23], but none specific to the topics we believed BiLAT taught. We therefore needed to develop one, and worked with several SMEs to create an SJT appropriate to measure learning from BiLAT and the ITS [9].

We used this SJT in several experiments. The participants' first task in each of these experiments was to complete the SJT. We then oriented the participants to the content of BiLAT by showing them a high-production-value video that depicted a live-action American-Iraqi meeting in which the American fails to adhere to the cultural norms of his host [8]. After the video, participants used BiLAT for several hours and then took the SJT again. Thus, we used the SJT in a pretest-posttest design; we defined learning as an increase in the correlation between participants' and SMEs' ratings from pretest to posttest. We found that BiLAT produced substantial overall gains in remembering and understanding [8, 9, 11, 13, 15].

According to Bloom's revised taxonomy, measures of remembering and understanding do not require the interactivity provided by BiLAT or the assistance provided by the ITS. Instead, passive approaches such as watching videos and listening to stories can affect measures of remembering and understanding. In a concurrent experiment, in which the video was shown *prior* to taking the SJT pretest, participants' scores on the pretest were as high as their scores on the posttest in our other experiments [22]. This result suggested that the video was affecting SJT scores. We tested this hypothesis in a subsequent experiment in which we administered the SJT, showed the video, and then again administered the SJT. Even without any use of the BiLAT system, there was a reliable improvement from pretest (M = .474, SE = .032) to posttest (M = .715, SE = .021): $F(1, 17) = 51.225$, $p < .001$, $\eta^2 = .751$. This result—and its magnitude—suggested that the SJT was highly influenced by the video. This result also meant that we could not determine the degree to which gameplay affected SJT scores in our prior studies. However, gameplay caused learning gains on other measures that should not be affected by the video [1]; these measures are discussed below.

3.2 Measuring the ability to apply knowledge: an in-game transfer task

The third level of Bloom's revised taxonomy is applying. *Applying* is the ability to solve a problem similar to those solved during training and modify what has been learned in order to transfer it to another situation. It can be thought of as extending understanding to novel applications.

We were able to use BiLAT itself to measure learners' ability to apply their knowledge. After participants used BiLAT for up to 100 minutes to solve a problem in an Iraqi marketplace, we disabled the ITS and asked participants to solve a new

problem with a different character. Our measure of learning in this transfer task was the probability that participants would select an inappropriate action during their meetings; lower probabilities indicated greater mastery. We chose this measure rather than a pretest-posttest design because interacting with the BiLAT system involves becoming familiar with the interface and how the system models various concepts like trust-building. To the extent that this familiarity affects the likelihood of making errors, pretest scores would have been artificially deflated and would have created the illusion of greater learning gains than were actually generated.

We had three goals in disabling the ITS in this transfer task. First, feedback from the coach could have decreased error probabilities over the course of the task. This effect would have inflated our estimates of learning and would have added noise to the data. Second, the ITS is *designed* to fade over time; it is intended to support practice in a way that helps the learner no longer need support, like training wheels on a child's bicycle. Third, there is no coach in the real world. Assessing learning under similar conditions thus added external validity to our measurements.

In one study, we used this transfer task to compare the pedagogical value of two different coaches. One coach provided action-level, easy-to-follow feedback (e.g., "don't give gifts that contain alcohol"). The other coach provided conceptual feedback, which required learners to more deeply contemplate their potential actions (e.g., "make sure that your gifts are culturally appropriate"). Otherwise, the coaches behaved identically. We found that, while the coaches were active, they were equally helpful; participants made as many errors in the market scenario with the conceptual coach as with the specific coach. However, in the transfer task (without the ITS), a different pattern emerged. Participants who had been assisted by the conceptual coach were reliably less likely to make errors. The deeper thought that the conceptual coach encouraged led the participants to be better able to transfer their understanding to a new character—but did not differentially affect their SJT scores [11]. In other words, both groups of participants had the same *amount* of remembered knowledge, but the conceptual coach enabled better *application* of that knowledge to new situations.

Unfortunately, there are significant drawbacks to using an in-game measure. Primarily, it invites the criticism that we are "testing to the teach." From that perspective, the new scenario cannot be considered a true transfer task. Indeed, without other measures, one could make the argument that people who use BiLAT may not be learning anything more than how to use BiLAT. To that end, the next section describes yet another measure we used to evaluate the efficacy of BiLAT and the ITS as an instructional system.

3.3 Measuring the ability to analyze: a cultural assimilator

The fourth of the six levels in Bloom's revised taxonomy is analyzing. *Analyzing* is the ability to deconstruct and examine instructional materials. It results in the student understanding why some solution can be applied to a particular set of problems. This understanding allows the student to infer the *causes* of problems and what makes particular solutions appropriate.

By definition, analyzing extends beyond the learner's experiences and reaches throughout the training domain. Thus, measuring learners' analytical skills does not require assessments to be tailored precisely to the content of the training system (vs. remembering or understanding). As a result, we were able to use an existing measure rather than creating our own. The measure we chose was the cultural assimilator (CA) created by Cushner and Brislin [7]. The implementations of this and many other CAs appear similar to the SJT, in that each item consists of a scenario that occurs in a target culture. However, whereas the SJT asks learners to rate various solutions to the problem described in the scenario, the CA asks learners to select the best *explanation* of the problem. On each item, selecting a culturally sophisticated explanation yields a score of two points; explanations reflecting some insight are worth one point; and inappropriate selections are worth zero points. Selecting a two-point explanation requires deep understanding for two reasons. First, the situations in the CA are not those encountered in gameplay, and so the gameplay experience must be analyzed in order to extract the needed information. Second, some one-point explanations are "attractive lures," meaning that people with a less sophisticated understanding of cross-cultural interaction will be likely to select them instead.

According to Bloom, measures of analytical skill are affected by interactive learning tools but not passive instructional tools like stories and videos [5]. Thus, the CA should have been affected by gameplay but not by the orientation video. We have found evidence in recent studies that gameplay improves CA scores—and is especially helpful for learners who started out in the bottom half of scores on the pretest [16, 22]. On the other hand, we included the CA in the video-only study described above (immediately following the SJT, pre- and post-gameplay). Unlike with the SJT, we found that the video caused negligible change in CA scores from pretest (M = 9.375, SE = .460) to posttest (M = 9.333, SE = .462): $F < 1$ *ns*. In summary, BiLAT has been shown to improve scores on both the SJT and the CA, whereas the video only affects SJT scores. Together, these results suggest that BiLAT and the ITS combine to form an effective teaching system. These effects manifest at the level of analysis—and probably at higher levels in Bloom's revised taxonomy [5].

4 What did *we* learn through this process?

The development of BiLAT took years. It began with an intensive study of cross-cultural negotiation. This effort resulted in storyboards and board-game prototypes, which were developed and refined into a simulation prototype. This prototype was refined through systematic review by SMEs. We used a similar process to develop the training support provided by the ITS [12].

Likewise, the assessment of learning from BiLAT has undergone iterative development. Initially, we used only the SJT and found results consistent with our hypotheses; BiLAT appeared to be an effective pedagogical tool. As we conducted further experiments tied more strongly to learning theory, it became clear that the video could be affecting the SJT results. We directly tested this idea and found that at least some of the improvements in SJT scores in our earlier studies were likely driven

by the video. In subsequent studies, we introduced additional measures that evaluate deeper levels of learning from gameplay and the ITS. These measures highlighted the result that BiLAT does not simply provide training for *remembering* and *understanding*, but furthermore supports *applying* and *analyzing* in an intercultural domain. Above all, our experience emphasized the need to analyze learning in an ill-defined domain more completely and at a deeper level. As will often be the case in ill-defined domains, understanding student learning is almost certainly going to require employing multiple—and more refined—measures. In our experience, Bloom's revised taxonomy was an informative guide for this exploratory process.

As our work continues, we will further approach our problem from multiple perspectives. Mendenhall has created several dimensions of cultural assessment from a review of many different instruments [19]. Some of these dimensions include measures of learners' satisfaction with the instructional tool, which is absent from Bloom's hierarchy but is increasingly important as learners more frequently are required to manage their own learning. Even as we diversify and improve our assessments, we must also strive to be realistic about their limitations. Cross-cultural interaction will always be imperfectly measured by sets of questions or rubrics for behavioral change, regardless of their refinement. Researchers operating in such ill-defined domains must reconcile the need for better assessment with the reality of the difficulties inherent in such an endeavor, and continue to draw conclusions from their data with the appropriate amount of caution.

Acknowledgements

The project or effort described here has been sponsored by the U. S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

[1] Anderson, L.W., & Krathwohl, D.R. (eds.): A Taxonomy for Learning, Teaching and Assessing: A Revision of Bloom's Taxonomy of Educational Outcomes. New York: Longman (2001)

[2] Bennett, M.J.: Towards Ethnorelativism: A Developmental Model of Intercultural Sensitivity. In Paige, R.M. (ed.) *Education for the Intercultural Experience*, pp. 21–71, Yarmouth, ME: Intercultural Press (1993)

[3] Bergman, M.E., Drasgow, F., Donovan, M.A., Henning, J.B., Juraska, S.E.: Scoring Situational Judgment Tests: Once You Get the Data, Your Troubles Begin. *International Journal of Selection and Assessment, 14*, pp. 223-235 (2006)

[4] Black, J.S., Mendenhall, M.: Cross-Cultural Training Effectiveness: A Review and a Theoretical Framework for Future Research. *The Academy of Management Review, 15*, pp. 113-136 (1990)

[5] Bloom B.S., Krathwohl, D.R.: *Taxonomy of Educational Objectives, Handbook I: Cognitive Domain*. New York: Longman (1956)

[6] Chan, D., Schmitt, N.: Situational Judgment and Job Performance. *Human Performance, 15*, pp. 233-254 (2002)

[7] Cushner, K., Brislin, R.W.: *Intercultural Interactions: A Practical Guide* (2nd ed.). Thousand Oaks, CA: Sage Publications (1996)

[8] Durlach, P.J.: Issues in Deployment of Serious Games. *Proceedings of the 31st Interservice/Industry Training, Simulation, and Education Conference* (2009)

[9] Durlach, P.J., Wansbury, T.G., Wilkinson, J.G.: Cultural Awareness and Negotiation Skills Training: Evaluation of a Prototype Semi-Immersive System. *26th Army Science Conference* (2008)

[10] Ericsson, K.A., Krampe, R.T., Tesch-Romer, C.: The Role of Deliberate Practice in the Acquisition of Expert Performance. *Psychological Review, 100*, pp. 363-406 (1993)

[11] Hays, M.J., Lane, H.C., Auerbach, D., Core, M.G., Gomboc, D., Rosenberg, M.: Feedback specificity and the learning of intercultural communication skills. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*, pp. 391-398 (2009)

[12] Hill, R.W., Belanich, J., Lane, H.C., Core, M., Dixon, M., Forbell, E., Kim, J., Hart, J.: Pedagogically Structured Game-Based Training: Development of the ELECT BiLAT Simulation. *Poster presented at the 25th Army Science Conference* (2006)

[13] Kim, J.M., Hill, R.W., Durlach, P.J., Lane, H.C., Forbell, E., Core, M., Marsella, S., Pynadath, D. V., Hart, J.: BiLAT: A Game-Based Environment for Practicing Negotiation in a Cultural Context. *International Journal of Artificial Intelligence in Education* (in press)

[14] Lane, H.C.: Metacognition and the Development of Intercultural Competence. *Proceedings of the Workshop on Metacognition and Self-Regulated Learning in Intelligent Tutoring Systems at the 13th International Conference on Artificial Intelligence in Education*, pp. 23-32 (2007)

[15] Lane, H.C., Hays, M.J., Auerbach, D., Core, M., Gomboc, D., Forbell, E., & Rosenberg, M.: Coaching Intercultural Communication in a Serious Game. *Proceedings of the 16th International Conference on Computers in Education*, pp. 35-42 (2008)

[16] Lane, H.C., Hays, M.J., Auerbach, D., Rosenberg, M.: Investigating the relationship between presence and learning in a serious game (under review at ITS2010)

[17] Legree, P.J., Psotka, J.: Refining Situational Judgment Test Methods. In *Proceedings of the 25th Army Science Conference* (2006)

[18] Lynch, C.F., Ashley, K., Aleven, V., Pinkwart, N.: Defining "Ill-Defined" Domains: A Literature Survey. In Aleven, V., Ashley, K., Lynch, C., Pinkwart, N. (eds.) *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems*, pp. 1-10 (2006)

[19] Mendenhall, M.E., Stahl, G.K., Ehnert, I., Oddou, G., Osland, J.S., Kuhlmann, T.M.: Evaluation Studies of Cross-Cultural Training Programs: A Review of the Literature from 1988-2000. In Landis, D., Bennett, J.M., Bennett, M.J. (eds.) *Handbook of Intercultural Training* (3rd ed.), pp. 129-144, Thousand Oaks, CA: Sage Publications (2004)

[20] Nydall, M.K.: *Understanding Arabs: A Guide for Modern Times* (4th ed.). Boston: Intercultural Press (2006)

[21] Ogan, A., Aleven, V., Jones, C.: Culture in the Classroom: Challenges for Assessment in Ill-Defined Domains. In Aleven, V., Ashley, K., Lynch, C., Pinkwart, N. (eds.) *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 8th International Conference on Intelligent Tutoring Systems*, pp. 92-100 (2006)

[22] Ogan, A., Kim, J., Aleven, V., Jones, C.: Explicit Social Goals and Learning in a Game for Cross-Cultural Negotiation. In *Proceedings of the Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education* (2009)

[23] Paige, R.M.: Instrumentation in Intercultural Training. In Landis, D., Bennett, J.M., Bennett, M.J. (eds.) *Handbook of Intercultural Training* (3rd ed.), pp. 85-128, Thousand Oaks, CA: Sage Publications (2004)