

Iterative Evaluation of a Large-Scale, Intelligent Game for Language Learning

W. Lewis Johnson, Carole Beal

Center for Advanced Research in Technology for Education

USC / Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292

Abstract. Educational content developers, including AIED developers, traditionally make a distinction between formative evaluation and summative evaluation of learning materials. Although the distinction is valid, it is inadequate for many AIED systems because they require multiple types of evaluation and multiple stages of evaluation. Developers of interactive intelligent learning environments must evaluate the effectiveness of the component technologies, the quality of the user interaction, and the potential of the program to achieve learning outcomes, in order to uncover problems prior to summative evaluation. Often these intermediate evaluations go unreported, so other developers cannot benefit from the lessons learned. This paper documents the iterative evaluation of the Tactical Language Training System, an interactive game for learning foreign language and culture. This project employs a highly iterative development and evaluation cycle. The courseware and software have already undergone six discrete stages of formative evaluation, and further formative evaluations are planned. The paper documents the evaluations that were taken at each stage, as well as the information obtained, and draws lessons that may be applicable to other AIED systems.

Introduction

Educational content developers conventionally draw a distinction between formative and summative evaluation of educational materials. Formative evaluation takes place during development; it seeks to understand strengths and amplify them, and understand weaknesses and mend them, before the educational materials are deployed. Summative evaluation is retrospective, to document concrete achievement [5]. Many view formative evaluation as something that should be kept internal to a project, and not published. This is due in part to the belief that formative evaluations need not involve learners. For example, Scriven [6] is frequently quoted as having said: “When the cook tastes the soup, that’s formative; when the guests taste the soup, that’s summative.”

Although the formative vs. summative distinction is useful, it does not provide much guidance to AIED developers. AIED systems frequently incorporate novel computational methods, realized in systems that must be usable by the target learners, and which are designed to achieve learning outcomes. These issues all warrant evaluation, and the “cooks” cannot answer the evaluation questions simply by “tasting the soup.” Yet one cannot use summative evaluation methods for this purpose either. Multiple evaluation questions need to be answered, which can involve multiple experiments, large numbers of subjects and large amounts of data. Meanwhile the system continues to be developed, so by the time the evaluation studies are complete they are no longer relevant to the system in its current form.

This paper documents the formative evaluation process to date for the Tactical Language Training System (TLTS). This project aims to create computer-based games, incorporating artificial intelligence technology, and each supporting approximately 80 hours of learning.

Given the effort required to create this much content, evaluation with learners could not wait until the summative stage. Instead, a highly iterative formative evaluation process was adopted, involving six discrete evaluation stages so far. Representative users were involved in nearly all stages. Each individual evaluation was small scale, but together they provide an accumulating body of evidence from which to predict that the completed system will meet its design objectives. The evaluation process has enabled the TLTS to evolve from an exploratory prototype to a practical training tool that is about to be deployed on a wide scale. These evaluation techniques may be relevant to other AIED projects that wish to make a smooth transition from the research laboratory to broad-based educational use.

1. Overview of the Tactical Language Training System

The Tactical Language Training System is designed to help people rapidly acquire basic spoken conversation skills, particularly in languages that few foreigners learn because they are considered to be very difficult. Each language training package is designed to give people enough knowledge of language and culture to carry out specific tasks in a foreign country, such as introducing yourself, obtaining directions, and arranging meetings with local officials. The curriculum and software design are focused on the necessary skills for the target tasks, i.e., it has a strong task-based focus [3]. The current curricula focus on the needs of military personnel engaged in civil affairs missions; however the same method could be applied to any language course that focuses on communication skills for specific situations. Two training courses are being developed so far: Tactical Levantine Arabic, for the Arabic dialect spoken in Lebanon and surrounding countries, and Tactical Iraqi, for the Iraqi Arabic dialect.

The TLTS includes the following main components [8]. The Mission Game (Figure 1, left side) is an interactive story-based 3D game where learners practice carrying out the mission. Here the player's character, at middle left, is introducing himself to a Lebanese man in a café. The player is accompanied by an aide character (far left), who can offer suggestions if the player gets stuck. The Skill Builder (Figure 1, right) is a set of interactive exercises focused on the target skills, in which learners practice saying words and phrases, listening to and responding to sample utterances. A virtual tutor evaluates the learner's speech and gives feedback that provides encouragement and attempts to overcome learner negative affectivity [7]. A speech-enabled Arcade Game gives learners further practice opportunities. Finally, there is a hypertext glossary can show the vocabulary in each lesson, the grammatical structure of the phrases being learned, and explains the rules of grammar that apply to each utterance.



Figure 1. Views of the Tactical Language Training System

2. Evaluation Issues for the TLTS

The stated goal of the TLTS project is to enable learners with a wide range of aptitudes to acquire basic conversational proficiency in the target tasks, in a difficult language such as Arabic, in as little as eighty hours of time on the computer. We believe that achieving this goal requires a combination of curriculum innovations and new and previously untested technologies. This raises a host of evaluation issues and difficulties. It is hard to find existing courses into which TLTS can be inserted for evaluation purposes, because the TLTS curriculum and target population differ greatly from that of a typical Arabic language course. Most Arabic courses place heavy emphasis on reading and writing Modern Standard Arabic, and are designed for high-aptitude learners. The TLTS Arabic courseware focuses on spoken Arabic dialects, and is designed to cater to a wide range of learners with limited aptitude or motivational difficulties. The TLTS employs an innovative combination of gaming and intelligent tutoring technologies; this method needed to be evaluated for effectiveness. It incorporates novel speech recognition [11], pedagogical agent [7] and autonomous agent technologies [14], whose performance must be tested. Because of the large content development commitment, content must be evaluated as it is developed in order to correct design problems as early as possible. It is not even obvious how much content is needed for 80 hours of interaction.

Then once the content is developed, additional evaluation questions come up. Standard language proficiency assessments are not well suited for evaluating TLTS learning outcomes. The most relevant assessment is the Oral Proficiency Interview (OPI), in which a trained interviewer engages the learner in progressively more complex dialog in the foreign language. Since TLTS learners apply language to specific tasks, their score on an OPI may depend on the topic that is the focus of the conversation. So-called task-based approaches to assessment [3] may be relevant, but as Bachman [1] notes, it is difficult to draw reliable conclusions about learner proficiency solely on the basis of task-based assessments. Thus TLTS faces a similar problem to other intelligent tutoring systems such as the PUMP Algebra Tutor [9]: new assessment instruments must be developed in order to evaluate skills that the learning environment focuses on. Finally, we need to know what components of the TLTS contribute to learning effectiveness; there are multiple components which may have synergistic effects.

3. Evaluating the Initial Concept

The project began in April of 2003, and focused initially on Levantine Arabic, mainly because Lebanese speakers and data sets are readily available in the United States. Very early on, an interactive PowerPoint mockup of the intended user interaction was developed and presented to prospective stakeholders. This was followed by simple prototypes of the Mission Game and Skill Builder. The Mission Game prototype was created as a “mod” of the Unreal Tournament 2003 game, using the GameBots extension for artificially intelligent characters (<http://www.planetunreal.com/gamebots/>). It allowed a learner to enter the virtual café shown in Figure 1, engage in a conversation with a character to get directions to the local leader’s house, and then follow those directions toward that house. The Skill Builder prototype was implemented in ToolBook, with enough lessons to cover the vocabulary needed for the first scene of the Mission Game, although not all lessons were integrated with the speech recognizer.

This prototype then was delivered to the Department of Foreign Languages at the US Military Academy (USMA) for formative evaluation. The USMA was a good choice for assisting the evaluation because they are interested in new technologies for language learning, and they have an extensive Arabic language program that provides strong training in spoken Arabic. They assigned an experienced Arabic student (Cadet Ian Strand) to go through the lesson materials, try to carry out the mission in the MPE, and report on the potential value of the software for learning. CDT Strand was not a truly representative user, since he already

knew Arabic and had a high language aptitude. However he proved to be an ideal evaluator at this stage—he was able to complete the lessons and mission even though the lessons were incomplete, and was able to evaluate the courseware from a hypothetical novice’s perspective.

An alternative approach at this stage could have been to test the system in a Wizard-of-Oz experiment. Although Wizard-of-Oz experiments can be valuable for early evaluation [13], they have one clear disadvantage—they keep the prototype in the laboratory, under the control of an experimenter. By instead creating a self-contained prototype with limited functionality, we obtained early external validation of our approach.

4. Adding Functionality, Testing Usability

Several months of further development and internal testing followed. The decentralized architecture of the initial prototypes was replaced with an integrated multi-process architecture [8]. Further improvements were made to the speech recognizer, and the lesson and game content were progressively extended. Then in April 2004 we conducted the next formative evaluation with non-project members.

Seven learners participated in this study. Most were people in our laboratory who had some awareness of the project; however none of them had been involved in the development of the TLTS. Although all had some foreign language training, none of them knew any Arabic. All were experienced computer game players. They were thus examples of people who ultimately should benefit from TLTS, although not truly representative of the diversity of learners that TLTS was designed to support.

The purpose of this test was to evaluate the usability and functionality of the system from a user’s perspective. Each subject was introduced to the system by an experimenter, and was videotaped as they spent a one-hour session with the software, using a simplified thinking aloud protocol [13]. Afterwards the experimenter carried out a semi-structured interview, asking the subject about their impressions of different parts of the system.

No major usability problems were reported, and none appeared on the videotape. The subjects asserted that they felt the approach was *much* better than classroom instruction. Subjects who had failed to learn very much in their previous foreign language classes were convinced that they would be able to learn successfully using this approach. The subjects also felt that the game and lesson components supported each other, that if they had spent more time in the lessons it would help their performance in the game.

At the same time, a number of problems emerged, both in the instructional design and in the use of the underlying technology. The pronunciation evaluation in the Skill Builder was too stringent for beginners; this created the impression that the primary learning objective was pronunciation instead of communication. The feedback of the pedagogical agent was repetitive and sometimes incorrect. Because we had designed the pedagogical agent to act human-like, instances of repetitive, non-human-like behaviour were especially glaring. Some subjects were unsure of where to go in the game and what to do. There was also a general reluctance to play the game, for fear that it would be too difficult. Once they got to the game, they had difficulty applying the knowledge that they had acquired in the Skill Builder.

These evaluations led to system improvements. The library of tactics employed by the pedagogical agent was greatly extended, pronunciation accuracy threshold was lowered, and speech recognition performance was improved. More simulated conversation exercises were added to the Skill Builder, to facilitate transfer of knowledge to the Mission Game. An introductory tutorial was created for the Mission Game, in order to help learners get started.

5. A Comparative Test with Representative Users

A more extensive test was then conducted in July of 2004 with representative users. It was structured to provide preliminary evidence as to whether the software design promotes learning, and identify what parts of the software are most important in promoting learning.

The following is a brief overview of this study, which is described in more detail in [2]. Twenty-one soldiers at Ft. Bragg, North Carolina, were recruited for the study. The subjects were divided in four groups, in a 2x2 design. Two groups got both the Skill Builder and Mission Game, two got just the Skill Builder. Two groups got a version of the Skill Builder with pronunciation feedback, two groups got no pronunciation feedback. This enabled us to start to assess the role that tutorial feedback and gameplay might have on learning outcomes. Due to the limited availability of test computers each group only had six hours to work with the software over the course of a week, so learning gains were expected to be limited.

The group that worked with the complete system rated it as most helpful, considered it to be superior to classroom instruction, and in fact considered it to be comparable to one-on-one tutoring. On the other hand, the group that got tutorial feedback without the Mission Game scored highest on the post-test. It appeared that combination of performance feedback and motivational feedback provided by the virtual tutor helped to keep the learners engaged and focused on learning. Some reported that they found the human-like responses to be enjoyable and “cool”. Apparently the shortcomings that the earlier study had identified in the tutorial feedback had been corrected.

Another important lesson from this study was how to overcome learners’ reluctance to enter the Mission Game. We found that if the experimenter introduced them directly to the game and encouraged them to try saying hello to one of the characters there, they got engaged, and were more confident to try it. With the assistance of the virtual tutor, many were able to complete the initial scenario in the first session.

Improvement was found to be needed in the Mission Game and the post-test. The Mission Game was not able to recognize the full range of relevant utterances that subjects were learning in the Skill Builder. This and the fact that there are only a limited range of possible outcomes of the game when played in beginner mode gave learners the impression that they simply needed to memorize certain phrases to get through the game. After the first day the subjects showed up with printed cheat-sheets that they had created, so they could even avoid memorization. We concluded that the game would need to support more variability in order to be effective. On the evaluation side, we are concerned that the post-test that we used was based on the Skill Builder content, so that it did not really test the skills that learners should be acquiring in the game, namely to carry on a conversation.

We subsequently made improvements to the Mission Game language model and interaction so that there was more variability in game play. We also came up with a way to make the post-test focus more on conversational proficiency: to use the Mission Game as an assessment vehicle. If the virtual tutor in the game is replaced by another character who knows no Arabic, the learner is then forced to perform the task unassisted. If they can do this, it demonstrates that they have mastered the necessary skills, at least in that context. To make this approach viable, it would be necessary to log the user’s interaction with the software. Therefore logging capabilities were added to enable further analysis of learner performance.

6. A Longer-Term Test with Representative Users

Once these and other improvements were made to the system, and more content was added, another test was scheduled at Ft. Bragg, in October, 2004. This time the focus was on the following questions. (1) How quickly do learners go through the material? (2) How proficient are they when they complete the material? (3) How do the subjects’ attitudes and motivation affect performance, and vice versa? Question 1 was posed to extrapolate from the work completed so far and estimate how much additional content would be required to complete an

80-hour course. Question 2 was posed to assess progress toward achieving task-based conversational proficiency. In particular, we wanted to assess whether our proposed approach of using the Mission Game as an assessment tool was workable. Question 3 was of interest because we hypothesized that the benefits of TLTS result in part from improved learner motivation, both from the game play and from the tutorial feedback.

For this study, rather than break the subjects into groups, we assembled just one group of six subjects, and monitored them through three solid days of work with the program followed by a post-test. They were also soldiers, with a range of different aptitudes and proficiencies, although being members of the US Army Special Forces their intelligence was greater than that of the average soldier. Their ages ranged from 20 to 46 years, and all had some foreign language background; one even had some basic training in Modern Standard Arabic. Not surprisingly, all subjects in this study performed better than in the previous study, and performance was particularly good on vocabulary recognition and recall, understanding conversations, and simulated participation in conversations. They were also able to perform well in the Mission Game when employed as an assessment tool. They made better use of the Mission Game, and did not rely on cheat sheets this time. Overall, the utility of the Mission Game was much more apparent this time.

Although most of the subjects did well, two had particular difficulties. One was the oldest subject, who repeatedly indicated that he was preparing to retire from the military soon and had little interest in learning a difficult language that he would never use. The other subject expressed a high degree of anxiety about language learning, and that anxiety did not significantly abate over the course of the study.

Meanwhile, other problems surfaced. The new content that had been introduced in time for this evaluation still had some errors, and the underlying software had some bugs that impeded usability. The basic problem was that once the evaluation was scheduled, and subjects were accrued, it was impossible to postpone the test to perform further debugging. Given the choice of carrying out the test with a buggy version of the program and cancelling it altogether, the better choice was to go ahead with the evaluation and make the best of it.

Another problem came up during analysis of the results: the log files that were collected proved to be very difficult to use. Questions that were easy to pose, e.g., “How long did each subject take on average per utterance in the final MPE test scene?” in fact proved to be difficult to answer. The log files that the TLTS generated had not been constructed in such a way as to facilitate the kinds of analyses that we subsequently wanted to perform. In a sense we relearned the lesson that other researchers have identified regarding interaction logs [10]: that log analysis is more than data collection, and attention must be paid both to the design of the logging facility and to the tools that operate on the resulting logs. Fortunately our iterative evaluation approach enabled us to learn this lesson quickly and correct the situation before subsequent evaluations.

7. Formative Evaluation of Tactical Iraqi

After having responded to the lessons learned from the previous test and corrected some of the errors in the Levantine Arabic content, we then temporarily put Levantine Arabic aside and focused on developing new content for Iraqi Arabic. There was a political reason for this (the desire to do something to improve the political situation in Iraq), a technical reason (to see if the TLTS was generalizable to new languages), and a pedagogical reason (to see if our understanding of how to develop content for the TLTS had progressed to the point where we could develop new courses quickly). Iraqi Arabic is substantially different from Levantine Arabic, and Iraqi cultural norms different from Lebanese cultural norms. Nevertheless our technical and pedagogical progress were such that by January 2005 we had a version of

Tactical Iraqi ready for external formative evaluation that was already better developed than any of the versions of Tactical Levantine Arabic that have been developed to date.

During January we sent out invitations to US military units to send personnel to our laboratory to attend a seminar on the installation and use of Tactical Iraqi, and to take the software back with them to let other members of their units use. Three units sent representatives. It was made clear to them that Tactical Iraqi was still undergoing formative evaluation, and that they had critical roles to play in support of the formative evaluation. During the seminar the participants spent substantial amounts of time using the software and gave us their feedback; meanwhile their interaction logs and speech recordings were collected and used to further train the speech recognizer and identify and correct program errors. All participants were enthusiastic about the program, and two of the three installed it at their home sites and solicited the assistance of other members of their unit in beta testing. Logs from these interactions were sent back to CARTE for further analysis.

Periodic testing continued through the spring of 2005, and two more training seminars were held. A US Air Force officer stationed in Los Angeles volunteered to pilot test the entire course developed to date in May. This will be followed in late May by a complete learning evaluation of the content developed to date, at Camp Pendleton, California. Fifty US Marines will complete the Tactical Iraqi course over a two week period, and then complete a post test. All interaction data will be logged and analyzed. Camp Pendleton staff will informally compare the learning gains from this course against learning from their existing classroom-based four-week Arabic course.

During this test we will employ new and improved assessment instruments. Participants will complete a pre-test, a periodic instrument to assess their attitudes toward learning, and a post-test questionnaire. The previous learning assessment post-test has been integrated into the TLTS, so that the same mechanism for collecting log files can also be used to collect post-test results. We have created a new test scene in the Mission Game in which the learner must perform a similar task, but in a slightly different context. This will help determine whether the skills learned in the game are transferable. We will also employ trained oral proficiency interviewers assess the learning gains, so that we can compare these results against the ones obtained within the program.

Although this upcoming evaluation is for higher stakes, it is still formative. The content for Tactical Iraqi is not yet complete. Nevertheless, it is expected that the Marines will make decisions about whether to incorporate Tactical Iraqi into their language program. Content development for the current version of Tactical Iraqi will end in June 2005, and summative evaluations at West Point and elsewhere are planned for the fall of 2005.

4. Summary

This article has described the formative evaluation process that was applied in the development of the Tactical Language Training System. The following is a summary of some of the key lessons learned that may apply to other AIED systems of similar scale and complexity. Interactive mock-ups and working prototypes should be developed as early as possible. Initial evaluations should if possible involve selected individuals who are not themselves target users but can offer a target user's perspective and are able to tolerate gaps in the prototype. Preliminary assessments of usability and user impressions should be conducted early if possible, and repeated if necessary, in order to identify problems before they have an impact on learning outcomes. In a complex learning environment with multiple components, multiple small-scale evaluations may be required until all components prove to be ready for use. Design requirements are likely to change based on lessons learned from earlier formative evaluations, which in turn call for further formative evaluation to validate them.

Mostow [10] has observed that careful evaluation can be onerous, and for this reason researchers tend to avoid it or delay it until the end of a project. An iterative evaluation method is infeasible if it involves a series of onerous evaluation steps. Instead, this paper illustrates an approach where each evaluation is kept small, in terms of numbers of subjects, time on task, and/or depth of evaluation. The individual studies may yield less in the way of statistically significant results than large-scale evaluations do, but the benefit is that evaluation can be tightly coupled into the development process, yielding a system that is more likely to achieve the desired learning outcomes when it is complete. The experience gained in the formative pilot evaluations will moreover make it easier to measure those outcomes.

Acknowledgments

This project is part of the DARWARS Training Superiority Program of the Defense Advanced Research Projects Agency. The authors wish to acknowledge the contributions of the members of the Tactical Language Team. They also wish to thank the people at the US Military Academy, Special Operations Foreign Language Office, 4th Psychological Operations Group, Joint Readiness Training Center, 3rd Armored Cavalry Division, 7th Army Training Command, and Marine Corps Expeditionary Warfare School for their assistance in the evaluations described here.

References

- [1] Bachman, L.F. (2002). Some reflections on task-based language performance assessment. *Language Testing* 19(3), 461-484.
- [2] Beal, C., Johnson, W.L., Dabrowski, R., & Wu, S., (2005). Individualized feedback and simulation-based practice in the Tactical Language Training System: An experimental evaluation. AIED 2005. IOS Press.
- [3] Bygate, M., Skeehan, P., & Swain, M. (2001). *Researching pedagogic tasks: Second language learning, teaching, and testing*. Harlow, England: Longman.
- [4] Corbett, A.T., Koedinger, K.R. & Hadler, W.S. (2002). Cognitive Tutors: From research classroom to all classrooms. In P. Goodman (Ed.): *Technology enhanced learning: Opportunities for change*. Mahwah, NJ: Lawrence Erlbaum Associates.
- [5] The Center for Effective Teaching and Learning, University of Texas at El Paso. Formative and summative evaluation. <http://cetal.utep.edu/resources/portfolios/form-sum.htm>.
- [6] Scriven, 1991, cited in "Summative vs. Formative Evaluation", http://jan.ucc.nau.edu/edtech/etc667/proposal/evaluation/summative_vs._formative.htm
- [7] Johnson, W.L., Wu, S., & Nouhi, Y. (2004). Socially intelligent pronunciation feedback for second language learning. ITS '04 Workshop on Social and Emotional Intelligence in Learning Environments.
- [8] Johnson, W.L., Vilhjálmsón, H., & Marsella, S. (2004). The DARWARS Tactical Language Training System. Proceedings of IITSEC 2004.
- [9] Koedinger, K.R., Anderson, J.R., Hadley, W.M., & Mark, M.A. (1997). Intelligent tutoring goes to school in the big city. *IJAIED*, 8, 30-43.
- [10] Mostow, J. (2004). Evaluation purposes, excuses, and methods: Experience from a Reading Tutor that listens. *Interactive Literacy Education: Facilitating Literacy Environments Through Technology*, C. K. Kinzer, L. Verhoeven, ed., Erlbaum Publishers, Mahwah, NJ.
- [11] Mote, N., Johnson, W.L., Sethy, A., Silva, J., Narayanan, S. (2004). Tactical language detection and modeling of learning speech errors: The case of Arabic tactical language training for American English speakers. InSTIL/ICALL Symposium, Venice, Italy.
- [12] Nielsen, J. (1994). Guerrilla HCI: Using discount usability engineering to penetrate the intimidation barrier. http://www.useit.com/papers/guerrilla_hci.html
- [13] Rizzo, P., Lee, H., Shaw, E., Johnson, W.L., Wang, N., & Mayer, R. (2005). A semi-automated Wizard of Oz interface for modeling tutorial strategies. UM'05.
- [14] Si, M. & Marsella, S. (2005). Thespian: Using multiagent fitting to craft interactive drama. AAMAS 2005.