Assessing Aptitude for Learning with a Serious Game for Foreign Language and Culture

W. Lewis Johnson and Shumin Wu

Alelo Inc., 11965 Venice Blvd., Suite 402, Los Angeles, CA 90066 USA
ljohnson@alelo.com, shuminwu@usc.edu

Abstract. The Tactical Language and Culture Training System is interactive environment for learning foreign language and culture, designed to help people quickly acquire spoken communication skills. It is a serious game, combining interactive game experiences as well as interactive lessons. As part of our research, we wish to understand what individual learner characteristics predict successful learning with this approach, and investigate whether the approach can be improved so that a wider range of learners can learn effectively with it. This paper reports on an experiment, to assess which learners learn most effectively with TLCTS, and attempt to identify the individual factors that predict successful training with TLCTS. A group of US Marines participated in a session of focused training with Tactical Iraqi™, an Iraqi Arabic course designed for military use. Performance scores and interaction logs were analyzed to determine which learners were most successful, and why.

Keywords: Student Assessment, Serious Game, Language Learning.

1 Introduction

The Tactical Language and Culture Training System is a serious game platform that helps learners quickly acquire knowledge of foreign language and culture through a combination of interactive lessons that focus on particular skills, and interactive games to practice and apply these skills. The system makes extensive use of intelligent tutoring and other artificial intelligence technologies, including automated speech recognition, spoken dialog and animated agents, natural language process and learner modeling. TLCTS is very widely used. At least twenty thousand copies of TLCTS courses have been distributed, and tens of thousands of learners have used them to date. The most widely used course is the Tactical Iraqi™, which teaches colloquial Iraqi Arabic.

A study by the Marine Corps Center for Lessons Learned (MCCLL) currently is documenting strong evidence of Tactical Iraqi™'s effectiveness. It examines the experience of the 2nd Battalion and 3rd Battalion, 7th US Marine Regiment (2/7 and 3/7 Marines), who trained with Tactical Iraqi™ prior to their most recent tour of duty in Iraq. The 3/7 attracted the attention of MCCLL because it did not suffer a single combat casualty during its most recent tour of duty. In the opinion of the 3/7 officers, the training greatly increased the battalion's operational capability as it enabled it to

B. Woolf et al. (Eds.): ITS 2008, LNCS 5091, pp. 520–529, 2008.

operate more efficiently, with an increased understanding of the situation and better relationships with the local people. They felt that the Marines who trained with Tactical Iraqi™ achieved a substantial level of language proficiency, so much so that they deserved to receive college credit for the language proficiency they gained. These results, while preliminary, suggest that Tactical Iraqi™ training led to improved on-the-job performance (a Kirkpatrick level 3 result) [12] and this in turn contributed to improved organizational outcomes (a Kirkpatrick level 4 result). These results follow earlier experimental studies that provide scientific evidence that Tactical Iraqi™ produces learning gains [14].

However, there remain a number of questions that previous research does not answer. Can one predict what types of learners will benefit the most from training with TLCTS? What patterns of learner behavior and performance are predictive of success with TLCTS?

This paper presents preliminary results from a field study attempting to address these questions. A group of Marines took part in a focused training session with Tactical Iraqi™, attempting to identify which individuals show the most promise of learning effectively with the software. This work is an example of what Chan has referred to as adoption-based research [2]: research that contributes to, and is predicated upon, the successful adoption of effective learning systems. Studies such as these, conducted with learners in authentic educational settings, are necessary to understand how educational software performs in practice, and is a necessary step toward transition of learning technology into regular field use.

2 System Overview

The following is brief overview of some of the main capabilities that TLCTS training systems provide. More detail may be found elsewhere [9, 10, 11].

Figure 1 shows images of TLCTS trainers in use. Current systems run on Windows PCs equipped with a headset microphone. Each course includes a Skill Builder, consisting of a set of interactive lessons, each of which focuses on communicative tasks. The top left of Figure 1 shows a Tactical Language learning lab installed for the U.S. Army 3rd Infantry Division at Ft. Stewart, GA. The top right of Figure 1 shows a typical Skill Builder lesson page. The learner can hear recordings of example phrases, and practice saying those phrases. The integrated speech recognizer, trained on language learner speech, gives the learner feedback as to whether or not their speech was intelligible and matched the target phrase. Learners practice in a series of exercises that progressively prepare learners for employing their language and cultural knowledge in conversational settings.

Two kinds of interactive games are included in TLCTS training systems. The bottom right of Figure 1 shows the Arcade Game in Tactical Pashto™, in which learners navigate their characters through a town by giving spoken commands in Pashto. The bottom left shows the Mission Game in which learners communicate with non-player characters using speech and gesture in order to carry out a mission. In this scenario, from Tactical Iraqi™, the player is instructing Iraqi non-player characters in the proper procedure for manning a security checkpoint.

Fig. 1. Images from the Tactical Language and Culture Training System (TLCTS)

TLCTS users receive tutorial feedback in the Skill Builder on their use of language. Depending on the type of exercise, the system can give feedback on pronunciation, morphological and grammatical forms, word choice, or cultural pragmatics, as in this example. In the games, on the other hand, the use of tutorial feedback is limited, as it was found to interfere- with game play. Instead, in the Mission Game feedback is integrated into the responses of the non-player characters in the game. The game display signals whenever the character's attitude changes, and the changes in attitude can influence the way the character responds to the learner.

3 Research Opportunity and Research Questions

The US Marine Corps Training and Education Command (TECOM) is currently conducting a multi-year study, called SEPTR (Simulation Enhanced Pre-deployment Training and Rehearsal) to evaluate the use of simulation-based training in preparing units for deployment overseas. The goals of the study are to test and validate existing training simulations, identify opportunities for improvement of those systems, and identify requirements for future training systems. The study is also expected to develop model methods for integrating simulation-based training effectively into training programs. These methods may then be disseminated across the Marine Corps, leading to the adoption of computer simulations as a standard method for training.

Because of TLCTS early success, TECOM selected it for inclusion in the SEPTR evaluations. The 2/7 Marines agreed to participate in SEPTR, and agreed to include

TLCTS into its current predeployment training program. However like any unit preparing for deployment overseas, the 2/7 has many skills to train, and very little time to complete the training. The battalion therefore decided to organize a "masters program", in which two Marines per squad would receive intensive training in language and culture. The challenge then was to quickly identify the Marines that were likely to benefit the most from TLCTS training, and enroll them in an intensive program of 40 or more hours of training with Tactical Iraqi™.

The standard method for assessing language aptitude is to employ a language aptitude test, e.g., the Defense Language Aptitude Battery (DLAB) used to determine who may pursue training as a military linguist. Such language aptitude tests are only moderate predictors of learning outcomes, typically yielding correlations of between 0.4 and 0.6 with learning outcomes as determined by a variety of outcome measures [4]. This is partly because other factors such as motivation influence language learning outcomes [5], but it also may be because current aptitude batteries may not test the full range of abilities relevant to language learning. In fact, the DLAB does not engage subjects in speaking the language, or using it for face-to-face communication in culturally appropriate ways. In contrast, TLCTS places particular emphasis on face-to-face communication in simulated social encounters. It is therefore not clear how strong a predictor the DLAB would be for the skills that TLCTS trains.

Therefore, instead of DLAB we decided to use a sample of material from Tactical Iraqi™ itself to assess likelihood for success in the masters training program. All candidate Marines would complete several hours of beginner-level training with TLCTS. The curriculum selected for this assessment would introduce the candidates to aspects of the phonology, morphology, syntax, and pragmatics of Iraqi Arabic, as well as non-verbal gestures and other cultural information relevant to face-to-face encounters with people in Iraq. We would then collect and analyze the data from the training sessions, including quiz scores, estimates of learner skill mastery, interaction logs, and speech recordings. We would also collect background information on each candidate, as well as self-assessments of their interest and motivation to learn Arabic. These data would allow us to answer the following questions:

1. Which candidates were most successful in their training?
2. Which characteristics of individual learners were conducive to success?
3. What patterns of training behavior led to success?

The choice of a version of the training system itself an assessment tool is unusual, but affords a number of advantages. It tests a wide range of cognitive abilities relevant learning language, wider than what is typical of language aptitude tests. It gives us an opportunity to determine whether trainees are able to assess their own language performance, plan their learning activities to achieve mastery, and recognize when they have successfully mastered the target language skills. Meanwhile, by taking part in the assessment the candidates are learning language and cultural skills that are potentially valuable to the trainees. This enhances motivation, both at the individual level (individual candidates are more likely to have intrinsic motivation to learn the language skills and do well) and at the organizational level (the officers in the battalion are more willing to set aside training time for the candidates to participate, and are more likely to have interest in successful training outcomes).

The data collected from this assessment, as well as from the training sessions of the candidates who are ultimately selected to complete the masters training program, gave us the opportunity to further investigate some research questions that are of concern to our research. One is the following:

4. Are game-based learning techniques useful in promoting learning?

Although games have attracted significant interest in educational circles, evidence of their effectiveness is mixed. This has led some educational researchers to question their value. Distractive elements [3] and learning-unrelated reward systems [13] are blamed for lowering productivity of learning activities. Undesired behaviors were reported where learners tried to use "shortcuts" to succeed in games by exploring the system properties instead of the learning materials [1]. Other researchers are optimistic about learning by playing games, but suggest games should be paired with traditional classroom curriculums and practices [6]. Previous studies and experience with TLCTS courses has also produced mixed results. Reports from TLCTS users indicate that they consider the game and simulation elements of TLCTS courses to be important and without them TLCTS would not have been chosen to be part of the SEPTR study. However an evaluation of an earlier version of Tactical Iraqi™ indicated that trainees actually rated the Skill Builder more highly than the game components [14]. We hypothesized that the subjects in the earlier study did not receive a proper orientation briefing regarding proper use of the learning software, and that the content focus of the game experiences needed to be adapted to make it more relevant to their jobs and missions. We wished to see whether better initial orientation, and recent improvements to the Mission Game, would result in improved attitudes toward the game experiences. We also wished to collect data on learner interaction with the games, to see whether there are opportunities to further improve the structure of the game experiences, and/or incorporate automated guidance and feedback, to help learners make most productive use of the games.

4 Study Procedure

The 2/7 officers selected 49 Marines to take part in the initial assessment, and organized them into two groups of approximately 25. The session proctor gave an initial twenty-minute orientation, demonstrated the software, and explained how to use it for training. The proctor told the candidates to strive to master the material, reviewing and repeating the learning materials, exercises, and quizzes as necessary.

Candidates then spent ten minutes completing a short questionnaire. The questionnaire asked whether the candidates had been deployed to Iraq before, if so how many times, and how motivated they were to learn Arabic. These questions were asked because motivation has previously been found to affect language learning outcomes [2], and in the case of Tactical Iraqi™ previous evaluations showed that trainees who had previously been deployed to Iraq had higher motivation to learn Arabic [8]. Candidates were asked to report their background information that reveal their maturity, experience, and/or job responsibilities, and training experience, which we hypothesized might influence how the candidates learn. The candidates then trained for approximately 45 minutes in the Tactical Iraqi™ Skill Builder. They were directed to

focus on four lessons: *Getting Started* (a tutorial), *Meeting Strangers* (vocabulary, phrases and etiquette relating to meeting strangers, possessive morphological endings), *Introducing Your Team* (Arabic terms for military ranks, phrases and etiquette relating to making introductions, definite articles, demonstratives, grammatical gender and agreement), and *Pronunciation Lesson 1* (easy Arabic consonants, long vs. short vowels, single vs. double consonants). The proctor provided the candidates with occasional technical assistance, but otherwise left them to train on their own. After a 10 minute break, the candidates were then directed to resume training in the Skill Builder for another 45 minutes. They were then directed to spend twenty minutes in the Mission Game, and then take another ten-minute break. Finally, the candidates completed another 30 minutes of Skill Builder training.

5 Study Results

Of the 49 participating Marines, one was excluded from the analysis presented here because he did not complete the survey questionnaire. Each subject was assigned a score between 1 (low) and 5 (high) for his performance in each of the three learning environments: Skill Builder, Arcade Game, and Mission Game. The Skill Builder scores were assigned according to the number lessons attempted, the number of lessons completed with a high quiz score (80% or better), and the number of individual language and cultural skills that the learner model indicated were fully mastered. The Arcade Game scores were assigned according to the number of levels played, completed, and the number of hints requested by the learner to complete the level. Similarly, the Mission Game scores were assigned according to the number of scenes played, the number of scenes completed, and the number of hints the learner used to complete the scene. Overall performance scores were computed based on the environment performance scores and time spent within each learning environment, using the following formula:

$$OverallPerformanceScore = \sum_{env} (T_{env} \times Score_{env}) / \sum_{env} T_{env} \ , \qquad (1)$$

where *env* represents the three learning environments, T_{env} is the time spent in a particular environment, and $Score_{env}$ is the assigned score for this environment. Note that the overall performance scores are continuous values computed out of ordinal environment performance scores.

5.1 General Results: Which Candidates Were Most Successful?

Certain observations were recorded during the proctoring sessions. First, although the candidates were instructed to focus on the Skill Builder lessons, some trainees still remained in the two game environments that interested them until the proctor specifically directed them back to the lessons. Secondly, some trainees left early for various reasons. Thirdly, some trainees who had used TLCTS before tended to skip Skill Builder lessons and devoted more of their training time to the game environments.

Therefore, actual training time (as determined from the log data) had a relatively high variance (time in Skill Builder: M = 1.08 hrs, SD = 0.72 hrs; time in Mission Game: M = 0.92 hrs, SD = 0.55 hrs; time in Arcade Game: M = 0.36 hrs, SD = 0.36 hrs). And this in turn resulted in high variance in performance scores in each environment (Skill Builder score: M = 2.92, SD = 1.46; Mission Game score: M = 2.92, SD = 1.49; Arcade Game score: M = 2.48, SD = 1.38).

Thus we needed to compute a summary score that can fairly reflect the trainees' overall performance. We hypothesized that the result the trainee could achieve in a particular learning environment was proportional to the time he has spent in this environment. Therefore, the aforementioned method (1) was introduced to compute the overall performance score and is expected to counteract the noise that perturbs the accuracy of otherwise simply computed average score that would be used as the overall performance score. We argue this method is valid because language and culture skills taught/practiced in these environments are closely related. For example, we regard those who invested most of their training time in one environment and accomplished great results as good learners even though they might have scored low in other environments due to time constraints. On the other hand, if a trainee evenly distributed his time but only does averagely in each environment, we view this trainee as a mediocre performer.

As a result, the average overall performance score for this population (N=48) is close to the medium category (M=2.91, SD=1.13, %95CI = [2.585, 3.241]). We found 10 most successful candidates who achieved high performance scores (>4.0). 1 out of 10 scored 5 in all the three environments; 3 out of 10 scored 5 in two environments, and the rest 6 scored 5 in one environment. The best candidates spent on average 2.5 hours pure training time with the system (SD = 0.43 hrs).

5.2 Which Individual Characteristics Were Conducive to Success?

The 11 characteristics we examined are categorized into 4 groups. The personal trait category includes *age*, *education*, self-reported *motivation to learn Arabic language and culture,* and *experience of training with TLCTS before*; the military experience category includes *rank*, *time in service*, *experience of deployment to Iraq*; the linguistic category includes *language spoken other than English* and *language formally studied*; the music ability category includes self-rated *musical talent*, *ability to sing or play instrument*, and *experience of formal music training*.

T-tests show that 32 trainees who identified their motivation greater or equal to 4 outperformed the 14 trainees having motivation below 4 ($t(44) = 2.012$, $p = 0.50$). Older trainees (>=20 year old) scored lower than younger ones (<20), but the difference is not statistically significant ($t(46) = -1.491$, $p = 0.14$). No significant difference was found for education, either. The 21 trainees who received some college education had performance close to the 27 trainees who only received high school degrees ($t(45.75) = -0.383$, $p = 0.715$). Interestingly, former TLCTS trainers did not have superior performance than fresher users do. Rather, they scored a little lower than those who have never trained with TLCTS before ($t(46) = -0.123$, $p = 0.902$) as they would be expected to. The proctor observed that some former trainees devoted little effort to the Skill Builder lessons and played a lot in the game environment, but they were not able to complete the entire game, probably because their language skills had

decayed. Additionally, it also could be that some of the former trainees did not learn much in the previous experience, or only spent a little time on the system. Finally, among the former trainees there was a cluster of trainees who had both very low motivation and performance.

In the military experience category, rank did not effect the training results, as the average scores for three groups of different ranks are approximately the same (Rank > E-3 Score: M = 2.88, SD=1.46; Rank = E-3 Score: M = 2.91, SD = 1.17; Rank = E-2 Score: M = 2.95, SD = 1.04). However, the group with less than one year of *time in service* and the group with more then one year had statistically different performance ($t(45) = 1.961$, $p = 0.056$). As for experience of deployment to Iraq, there is no significant finding between the group with the experience and the group without ($t(44) = -.822$, $p = 0.416$).

Those who had studied another foreign language performed at a level that was close to those who did not ($t(46) = 0.115$, $p = 0.909$).In the language experience category, only 4 trainees speak a language other than English, so it is impossible to draw conclusions about the role of foreign language fluency.

In the music ability category, no significant effect is found. Trainees who rated their music talents higher seemed to score slightly lower than those who identified themselves as "I have no talent in music" ($t(46) = -0.551$, $p = 0.584$). Similarly, trainees who reported practicing singing or playing instrument were outperformed by their non-practicing counterparts ($t(45) = -1.091$, $p = 0.281$). However, those having taken formal music training scored a little higher ($t(45) = 0.430$, $p = 0.669$). But those results are not statistically significant to verify hypotheses.

In summary, characteristics such as *motivation* and *time in service* seem promising to be conductive to success. We do not find significant effect with other characteristics. The findings are reinforced when we take a look at the group of those successful candidates. We found out among the 10 best trainees, 90% reported high motivation, and 70% served in military more than 1 year. T-tests on the best candidate group and the other trainee group also show that motivation has significant effects on the overall performance ($t(44) = 2.381$, $p = 0.021$), while the effect of time in service seems not statistically significant ($t(9.07) = 1.036$, $p = 0.372$).

5.3 What Patterns of Training Behavior Led to Success?

We examined the activity patterns of the successful candidates against the rest of participants. It was found that successful learners did particularly well in Skill Builder lessons, compared with the rest of the trainees (quizzes completed: $t(9.65) = 2.654$, $p = 0.025$; skill mastered: $t(46) = 2.691$, $p = 0.100$). We believe that this provided them with good foundations to be able to apply the language and culture skills they learned from the lessons to the other game environments. In the Arcade Game, 60% of them never requested a single hint to complete a level, and therefore were never penalized by minus points because of hint requests.

Log files show that they also performed in-game learning. For instance, 60% of them used this strategy: when playing the mission scenes, they first heavily used the hint facility to go through them, and then replayed the scenes and finally completed them The best performer group requested 59.10 mission game hints on average, compared with the other performer group which used only 20.97 hints on average

($t(9.87) = 2.382$, $p = 0.039$). As we can see the successful learners used different strategies in the Mission Game and Arcade Game. The difference between these two games explains the distinction of their behaviors. In the Mission Game even though the aide agent can offer hints on the expected speech in English and Arabic, the learner would not be able to memorize it if he/she did not build up enough skill level from the Skill Builder lessons due to the complexity of the speech. Therefore, they need to request hints often. In the Arcade Game, especially the beginner levels, expected utterances are relatively short and simple, and therefore medium-leveled skills can be directly applied.

6 Study Changes and Future Work

After the assessment data described in this article were collected, the 2/7 Marines received word that they might have to deploy to Afghanistan instead of Iraq. The 2/7 therefore called a halt to the Iraqi assessment, and made plans to initiate it again with the Dari language spoken in Afghanistan. This is an example of the challenges inherent in conducting *in vivo* evaluations of learning software in the context of training practice. Such evaluations have greater external validity than studies in controlled laboratory settings, but they must adapt to the constraints of the organization's training activities.

Our future work includes the plan to collect more data from other Marines units to find out whether they were successful in their training. We also plan to observe their final live training exercise, in which they must interact with Iraqi role players. This will help to determine how effective their training really was.

7 Conclusions

A critical lesson we learnt from design of game-base training is how to design learning environments to optimize pacing. ITS research doesn't often consider the question how to keep learners engaged for extended periods. This of course is a key issue for computer games, which are typically designed specifically to promote extended play. The experience with Tactical Iraqi shows that this is a critical issue, and the game elements help to maintain a sustainable learning pace.

One of the attractions of game-based learning is that games promote motivation. Our results indicate that motivation is overall a key predictor of learning success. However the experience shows that games also motivate learners to engage in learning of their choice, rather than follow a designated program of instruction. We conclude from this that we need to provide learners with that freedom of choice, yet we should also provide learners advice of what to work on next, to make sure that they are being productive at all times. And that in turn requires instructional planning capability that adapts to the learner's choices, and a learner modeling capability that is works robustly regardless of the learner's choices.

Acknowledgments. This work was sponsored in part by the US Marine Corps, Program Manager for Training Systems (PM TRASYS).

References

1. Baker, R.S., Corbett, A.T., Koedinger, K.R.: Detecting Learner Misuse of Intelligent Tutoring Systems. In: Proceedings of the 7th International Conference on Intelligent Tutoring System, pp. 531–540 (2004)
2. Chan, T.W.: The four problems of technology-enhanced learning. In: Plenary address to AIED 2007 (2007)
3. Conati, C., Klawe, M.: Socially Intelligent Agents to Improve the Effectiveness of Educational Games. In: Proceedings of AAAI Fall Symposium on Socially Intelligent Agents - The human in the loop (2000)
4. Ellis, R.: The study of second language acquisition. Oxford University Press, Oxford (1994)
5. Gardner, R.: Social psychology and second language learning: The role of attitudes and motivation. Edward Arnold, London (1985)
6. Henderson, L., Klemes, J., Eshet, Y.: Just Playing a Game? Educational Simulation Software and Cognitive Outcomes. Journal of Educational Computing Research 22(1), 105–129 (2000)
7. Johnson, W.L.: Serious use of a serious game for language learning. In: Luckin, R., et al. (eds.) Artificial Intelligence in Education, pp. 67–74. IOS Press, Amsterdam (2007)
8. Johnson, W.L., Beal, C.: Iterative evaluation of a large-scale, intelligent game for language learning. In: Artificial Intelligence in Education, IOS Press, Amsterdam (2005)
9. Johnson, W.L., Beal, C., Fowles-Winkler, A., Lauper, U., Marsella, S., Narayanan, S., Papachristou, D., Vilhjálmsson, H.: Tactical Language Training System: An Interim Report. In: Lester, J.C., et al. (eds.) Intelligent Tutoring Systems: 7th International Conference, ITS 2004, pp. 336–345. Springer, Berlin (2004)
10. Johnson, W.L., Marsella, S., Vilhjálmsson, H.: The Tactical Language Training System. In: Proceedings of I/ITSEC 2004 (2004)
11. Johnson, W.L., Vilhjálmsson, H., Marsella, S.: Serious Games for Language Learning: How Much Game, How Much AI? In: Artificial Intelligence in Education. IOS Press, Amsterdam (2005)
12. Kirkpatrick, D.F.: Evaluating Training Programs: The Four Levels, Berrett-Koehler, San Francisco (1994)
13. Prensky, M.: Digital Game-Based Learning. McGraw-Hill (2001)
14. Surface, E.A., Dierdorff, E.C., Watson, A.: Special Operations Language Training Software Measurement of Effectiveness Study: Tactical Iraqi Study Final Report. Special Operations Forces Language Office, Tampa, FL, USA (2007)