

## **ELLS EVALUATION PLAN**

(Draft by Yong Zhao, Michigan State University, Alan Ginsburg, US DE)

### Overview

The evaluation for the whole ELLS project is divided into two broad categories: (a) developmental evaluation and (b) implementation evaluation, corresponding to the main stages of the project. The development evaluation will assess the effectiveness of project components while the implementation evaluation will assess the effectiveness of the whole project. The purpose of the assessment of one or more of the ELLS components is to provide timely, systematic, and useful information to the design team and technical work group so as to ensure the quality of the development. The aim of the systems evaluation is to assess the overall effectiveness of the program in supporting second language learning so as to provide strategic program information to help make judgments about how well the program works and areas to focus to improve its effectiveness.

### Evaluation Plan

#### ELLS Stages and Objectives

The overall goal of the evaluation is to evaluate how well the ELLS project is meeting its objectives at different stages of implementation (see Table 1). The ELLS project evaluation will be conducted at critical moments of the development, implementation, and deployment stages. Depending on the stage, we will use a combination of focus groups, online surveys, observations, interviews, and tests to collect gather data and provide continuous feedback to the design and development team so that they can make more informed design and implementation decisions, which should result in a better product.

**Table 1: E-Language Learning System (ELLS) Performance Objectives by Key Stage of System**

<b>ELLS Stage and Component</b>	<b>Performance Objectives</b>
<b>1. Developmental Stage</b>	
Content	<ul style="list-style-type: none"> <li>• Administrators/Teachers perceive content aligns with local language instructional standards/assessments for age group</li> <li>• Administrators/Teachers believe content will yield one year of growth</li> </ul>
<b>Information Technology</b>	
<ul style="list-style-type: none"> <li>• Speech</li> </ul>	<ul style="list-style-type: none"> <li>• Text to speech translation is sufficiently clear and natural to be useable for language learning.</li> <li>• Speech recognition over limited speech domain is sufficiently accurate for students of different abilities to be useable.</li> </ul>
<ul style="list-style-type: none"> <li>• Communication</li> </ul>	<ul style="list-style-type: none"> <li>• E-mail/chat functions provide appropriate safeguards for students.</li> <li>•</li> </ul>
<ul style="list-style-type: none"> <li>• Platform</li> </ul>	<ul style="list-style-type: none"> <li>• Learning management system provides required student assessments.</li> <li>• Student record system provides required</li> <li>• Platform effectively integrates different system components without excessive loss of system responsiveness.</li> </ul>
Instructional Design	<ul style="list-style-type: none"> <li>• Users are engaged in story and characters</li> <li>• Gaming format is effectively integrated into instruction</li> </ul>
Completed modules	<ul style="list-style-type: none"> <li>• Individual modules produce appropriate learning outcomes</li> <li>• Individual modules perform on a variety of machines as intended.</li> </ul>
<b>2. Implementation Stage</b>	
Fully Operational System: Pilot test	<ul style="list-style-type: none"> <li>• Fully operation system in pilot sites meets functionality requirements.</li> <li>• Fully operation system in pilot sites produces designed learning outcomes.</li> </ul>
Fully Operational System: Public release	<ul style="list-style-type: none"> <li>• International and national delivery and distribution system meets functionality requirements.</li> <li>• Fully operational system produces desired learning outcomes</li> </ul>

Strategies and Settings for the Evaluation

The evaluation will use four different strategies to accomplish its goals: (1) applied lab test, (2) field test, (3) pilot implementation evaluation, and (4) full implementation evaluation.

### Applied Lab Tests

The E-Language system will utilize a number of new technologies, such as speech recognition and speech synthesis, to facilitate language learning. However the effectiveness and technological feasibility have not been well examined in language applications. Thus it is necessary to conduct developmental evaluations about their effectiveness and functionality before they should be used. For this purpose, the formative evaluation team will conduct a number of experiments in a controlled environment to assess the effectiveness and technical capacities of these new technologies.

New technologies that will need development assessments include:

- Speech technologies
- Machine translation technologies
- Automatic glossing technologies
- Computer-mediated communication technologies

Lab tests will also be carried out assess user reactions to prototypes of the content. There will be points when empirical evidence is needed to make decisions about certain design issues. For example, while in general people may believe full animation is more appealing and engaging, but it is more costly than slide-style animation. Experiments can be conducted to compare the effects of full animation and slide-style animation. Design issues that may need empirical evidence include:

- Effectiveness of different qualities of animation
- Effectiveness of different qualities of sound/audio
- Effectiveness of different qualities of imaging
- Effectiveness of different design features

Lab tests differ from field tests in a number of ways. First, they are more controlled. They are conducted in a lab setting instead of a natural learning situation. Second, they use relative smaller numbers of participants. We could use a small sample of potential learners. Third, they are used to gather data about a specific issue rather than the whole system or product.

### Field tests of ELLS components

Field tests are evaluations conducted in more natural instructional settings, such as a school or classroom or the learner's home instead of artificial research settings. Field tests will be used to evaluate the effectiveness of complete prototypes.

The major differences between field tests and implementation tests are along three dimensions. First, field tests are conducted at a smaller scale. Second, field tests are conducted on prototypes instead of real products. Third, field tests are shorter. While field tests will be used to assess the effectiveness of a series of prototypes, they are not necessarily continuous.

### Pilot implementation tests

Pilot tests evaluate implementation of the ELLS system in a limited number of sites under conditions with extensive monitoring and technical assistance. Pilot tests are real-world beta tests to assess overall system implementation, identify problems, and obtain initial estimates of learning outcomes in pilot sites. Pilot tests are similar to beta tests, a common practice in commercial product development.

### Full implementation tests

System tests are the final stage of evaluation. They are used to assess the effectiveness of the complete system, when it is used in real instructional settings as a significant part of the language instruction for an extended period of time.

The following sections describe organizations and activities for each stage.

### Applied Lab Tests: Design and Development

At the first stage, the design team will need information from the intended audience of the product to make decisions on issues related to the overall system, including the storyline that will drive the product, cast of characters that will be used throughout the game, and system level user interface. The primary focus of the evaluation at this stage is on collecting information about user interest and engagement in the story, their preferences for certain interface and set of characters. It is also necessary to assess student learning from prototypes at this stage. Experiments, and a focus group approach and surveys will be used to collect the information needed.

### Applied R& D Experiments

Objectives: Applied experiments will be used to assess the effectiveness of applying new technologies and design strategies to language learning.

Organization: Two applied experimental language labs should be established, one in China and the other in the US. These labs will conduct applied research on issues faced by the design teams or issues raised by the US Department of Education and Chinese Ministry of Education. The lab would support experiments to test the functionality of the new technologies such as the accuracy of speech software on different populations. The lab could also assess instructional design features early on for their engagement and learning content. Each lab should have a pool of participants (around 30) that can be assembled quickly to participate in these experiments. The participants should share common characteristics with the targeted audience, including age, language proficiency, and academic standing.

### Focus Groups

**Objectives:** Focus groups will be used to collect in-depth reactions and generate alternative designs and concepts to the prototypical design of the design team. They will be used:

- At the early stage of the project to determine the storyline, the overall interface, and designs of major characters;
- At the beginning of the development of each episode to determine the activities, settings, interface, and content of the episode;
- When prototypes are developed for each episode and decisions are needed to select the best design;
- And whenever there exists significant disagreement among the design team members and user input is needed.

**Organization:** Each focus group will consist of 10 to 20 members. There are three possible ways to organize the student and teacher focus groups, each has its advantages and disadvantages.

Option one:

- Four student focus groups in China, one in large modern cities (e.g., Beijing), one in medium sized but affluent cities (e.g., Shenzhen), one in rural areas (e.g., Yunnan Province);
- Four student focus groups in the US for the Chinese content, one in large urban inner city school without large Chinese population (e.g., Detroit), one in large urban inner city school with large Chinese population (e.g., New York), one in suburban school and one in rural school;
- Four student focus groups in the US for the ESL content, each representing a major geographical area;
- Members of the student focus groups will be selected from the same location and possible the same school;
- Each focus group will have about 20 members;
- Three expert teacher focus groups, one for English in China, one for Chinese in the US and one for ESL in the US;
- All focus group activities will be conducted onsite and face-to-face. A researcher will be present to conduct the sessions and collect the data.

Option two:

- One student focus group in China for the English content;
- Two student focus groups (one for Chinese and one for ESL) in the US;
- Members will be recruited from different locations and schools;
- Three expert teacher focus groups, one for English in China, one for Chinese in the US and one for ESL in the US;

- Focus group sessions will be mostly conducted online using video conferencing or other telecommunication technologies.

Option three:

- One student focus group in China for the English content;
- Two student focus groups (one for Chinese and one for ESL) in the US;
- Members will be recruited from schools in the same geographical area;
- Three expert teacher focus groups, one for English in China, one for Chinese in the US and one for ESL in the US;
- Focus group sessions will be mostly conducted in the same physical location and face-to-face.

Apparently, option one has the broadest representation and thus possibly the best external validity, however it is difficult to manage and quite costly. Option 2 preserves the advantage of broad representation of option 1 but is less costly and easier to manage than option 1. However, the online format may result in less comprehensive and in-depth responses. Additionally, it is also possible that the reduced membership also reduces the diversity of views. Option 3 is the most economical and easiest to manage of all, but it also has the least external validity due to its limitation of geographical and background representation.

### Surveys

Objectives: Surveys are a way to quickly collect information from a large group of respondents, although the information may not be as in-depth as that collected through focus groups. Survey data are easier to process as well. Surveys can be used after the design team has decided on a limited list of options for a particular issue, such as the storyline, interface, and characters but still needs responses from a broader audience.

Organization: Surveys can be delivered in one or both of the following ways:

- On-line. The surveys are published on the Web to collect data from either a pre-selected sample or the public, that is, anyone who is interested in providing feedback.
- On-site. The surveys are administered on-site to a pre-selected sample.

One advantage of on-line survey is that the data are already gathered in a database, so we do not need to enter the data later. Another perhaps more important advantage is that it can reach a broader audience. Participation is then not limited to certain geographical area. A disadvantage of online survey, especially when it is open to the public, is that the fact of needing to use a computer to provide the information has the potential of introducing sampling biases. For example, it is only possible for those who have a computer and are comfortable using computers to complete the surveys. Those who are comfortable completing surveys online may respond differently from those who are not able to complete the surveys online.

### Field Tests of System Components

As soon as a usable component prototypes are available, we will need to assess their effectiveness among targeted users. While we are still interested in user responses to the interface and storyline, we are more interested in finding out what students can learn and to what extent students are engaged in the learning process intended by the product. We will also examine which of the features and functions of the product work better than others. At this stage, we will continue to use the student groups previously established. However, these groups will now have an extended role. In addition to serving as focus groups to react to designs, these groups will serve as users of the prototype. Their role is two fold. One is to assess its functionality in practice. The second is to run mini-learning experiments to test out the learning value of different modules in practice.

### Pilot Implement Evaluation

Pilot test is like beta testing. The primary objective of evaluation at this stage are to:

- Evaluate the delivery system and
- assess the cognitive and affective outcomes of the product

Thus it can only begin when there is sufficient developed content and a production schedule to ensure an uninterrupted period when the learner can use the product as a main source of language learning. This does not mean we have to wait until the whole system is completed.

### Evaluation Design

A quasi-experimental design will be used. The experiment will take place in 12 schools: four in China for the English content, eight in the US (four for English and four for Chinese). Each school will select one class to participate in the experiment. Another class will be selected as the control.

### Outcome Measures

Outcome measures include:

- Language proficiency (all four skills, vocabulary, and grammar).  
Language proficiency assessment should more standard-based assessments that are commonly used.
- Interest in the target language;
- Engagement in the materials.

### Full Implementation Evaluation

At this stage, we will be interested in assessing the impact of the system and thus need to include more assessment that are commonly used, such as NAEP and other standard-based assessment tools.

When the pilot evaluation is completed and if the results are positive, we will expand the experiments to more schools and open the system to the public. At this stage, our main focus will be on assessing the overall impact of the project. While we will continue to collect data from new experimental sites, our scope of evaluation will be expanded to include other users as well.

At the summative stage, we will collect the following information to assess the overall impact of the project:

- Learning outcomes will be measured with standardized language proficiency tests;
- Learner engagement will be assessed with instruments used in stage 3 of the formative evaluation;
- Dissemination will be assessed by the number of users of the system.

### Summary of Evaluation Plan

The following table summarizes the formative evaluation plan:

Stage	Outcome/objective	Methods	Timeline
Lab	Effectiveness of new technologies and strategies; Interest in the story, interface, & characters; Alternative concepts and designs; Engagement in the design and concepts.	Experiments Focus Groups Surveys	March, 2002 for overall design & concept; Beginning of each new episode.
Field	System component operations Engagement in modules	Surveys; Observations; Interviews	Prototype stage for each new episode
Pilot	Learning outcomes; Affective outcomes. Pilot delivery and operations	Language proficiency assessment; Observation; Surveys; Focus groups	September, 2003 to July, 2004
Implementation	Learning outcomes;	Language	July 2004

	Affective outcomes. System delivery and operations	proficiency; Surveys; Interviews.	
--	--	---	--

