Research and Development of Online Adaptive Placement Test of Listening Comprehension: A Preliminary Report

Wen-Shuenn, Wu
Instructor, Department of Foreign Languages and Literature, Chung Hua University

Abstract

A prototype of an online adaptive placement test of listening comprehension has been established. This test system, which is the second stage of a web-based adaptive English placement test platform, is now being developed to help universities group freshman students into classes which match their English proficiency. In a computer-adaptive language test (CALT), each examinee takes a unique test that is tailored to his ability level. It is generally believed that CALT is faster, more reliable and more cost-effective. While most computer-based test models use computer servers to randomly choose next suitable test items, more carefully designed regulations are applied to the item selection from the test bank so that a listening comprehension test will be more efficient and reliable. Different types of listening test will also be introduced. Finally, Some practical issues such as the classification of the word list, definitions of item codes, and design of test makers’ interface are also discussed.

Key Words: listening comprehension, computer-adaptive language test, computer adaptive test, placement test.

Received: April, 2005 ; In revised form: Jun 15, 2005
Introduction

It has been a tendency that more and more universities in Taiwan would like to give freshmen an English placement test so that they can be grouped into appropriate levels of English competence. In all likelihood, a web-based English placement test is the only solution to the laborious and time-consuming task of classifying freshmen in the first week of the fall semester. It has been more than three years since Chung Hua University adopted web-based English placement tests made by a computer information company to group freshmen into three levels. Students of level A have better English proficiency and only have to take one-year general English courses and one-year more advanced English for specific purposes (ESP). Students of level B are assumed to have intermediate English proficiency and are required to take two-year general English courses and one-semester ESP. Students of level C, who usually have lower English competence, are asked to take three-year general English courses. English placement tests adopted by Chung Hua University have some distinctive features. First, it is a web-based English placement test with listening comprehension, grammar, and reading comprehension. The English placement test with test items randomly chosen by computers is administered at the computer center of Chung Hua University usually in the first week of a fall semester. Second, in order to encourage students to study hard and take fewer English courses, students can be upgraded to a higher level if they receive a score higher than cutting points to certain levels of a placement test given in the end of freshman year or sophomore year. Third, to make sure that all freshmen, sophomores, or juniors come to take this placement test, the test result take 30% of their final average of English courses.

While the web-based English placement test can efficiently help us group students into proper level of English proficiency, it has a couple of serious drawbacks. First of all, it is quite expensive. The university has to pay more than NT$800,000 each year to give two placement tests. Second, the Department of Foreign Languages and Literature can not control the quality of test items and is not given a complete, statistically valid analysis report after the test. Finally, the computer information company which offers us test items has not fully taken the advantage of computer. The English placement tests that we have administered are randomly linear ones rather than systematically adaptive ones. In addition, the computer information company has not applied multimedia to current English placement tests except audio files in the listening comprehension.

Literature Review

It is no doubt that listening comprehension is an essential skill that EFL learners should acquire as early as possible. As Dunkel (1986) pointed out that most researchers of listening comprehension agreed that “listening comprehension should be the focal methodology in foreign/second language instruction, particularly at the initial stages of language study” (p.99).
Many researches on factors affecting listening comprehension have been conducted (Boyle, 1984; Samuels, 1984; Flowerdew & Miller, 1992). Some of these factors include unfamiliar lexis (Kelly, 1991; Tsai, 2004), speech rate (Griffiths, 1991; Zhao, 1997), accents (Major, Fitzmaurice, Bunta, & Balasubramanian, 2002), and previous background knowledge (Chiang & Dunkel, 1992).

Since listening comprehension plays an important role in facilitating language learning, what are some effective methods to assess whether EFL learners achieve this skill? In chapter 3 of Assessing Listening, Gary Buck (2001) states that assessing listening can be historically divided into three approaches:

1. discrete-point approaches such as phonemic discrimination tasks, paraphrases recognition, and response evaluation
2. integrative approaches such as reduced redundancy, listening cloze, gap-filling tests, dictation, sentence-repetition tasks, and translation
3. communicative approaches with authentic texts and authentic tasks

However, some approaches or activities that emphasize listening for words rather than listening for meaning are discouraged by scholars such as Jack C. Richards. A lot of researchers of listening comprehension also suggest that top-down process instead of bottom-up process should be applied to listening comprehension strategies. According to Long (as cited in Chiang & Dunkel, 1992), while little has been done about listening from schema-theoretic perspective, interaction of listener’s life experience and prior knowledge about the world with spoken discourse greatly enhances the process and construction of meaning.

In classroom foreign language instruction, three stages of listening strategy or technique are usually used as a good format of listening comprehension class (Dunkel, 1986; Field, 1998):

1. **pre-listening**: predicting the content of the spoken message
2. **listening**: identifying key words and using background knowledge to process selective listening
3. **post-listening**: checking accuracy of comprehension and examining functional language

Nevertheless, in reality this good format of listening comprehension class can not be ideally reflected through listening test. Most conventional paper-and-pencil listening tests, which usually do not have pre-listening warm-ups and often do not have post-listening activities, evaluate whether listeners can decode auditory input, discriminate similar phonemes, interpret stress and intonation, recognize paraphrases, and respond appropriately. Most traditional listening tests are recorded beforehand and play once only during test. Test takers usually do not have sufficient time to anticipate what they are going to hear before the listening; hence, they hardly activate their prior knowledge to correlate the listening or make prediction.
about the content of what they will hear. Therefore, theoretically and ideally, listening comprehension is viewed as an active process in which learners concentrate on selective aural input and interact what they hear with their prior knowledge and experience to generate meaning (O’Malley, Chamot, & Kupper, 1989). It can be true in a well-planned listening comprehension class with pre-listening, listening, and post-listening stages and an experienced teacher as a facilitator and instructor. However, in traditional, group-administered, fixed-length paper-and-pencil listening tests, most language learners struggle for success in listening comprehension measured by correct responses to questions. Take the most popular multiple-choice questions in a listening comprehension test for example. Test takers have to decode spoken discourses within limited and pressed period and read the three or four possible answers to find the most appropriate one. In other words, in an objective listening comprehension test (e.g. multiple-choice questions, true/false items, gap-filling formats, or dictation), listening is more like a receptive skill rather than an active process. Researchers can not fully understand how those answers have been arrived at – by identifying all the specific and implied meaning of utterances or making a lucky guess.

**Design of Online Adaptive Placement Test of Listening Comprehension**

The principle of computer-adaptive test (CAT) is that each examinee is usually given a test item of medium-difficulty level. If the examinee responds correctly, the next item received is more difficult. If the examinee misses the item, an easier question is offered in turn. In other words, specific test items are presented by the computer according to the estimated ability of the examinee and his/her response to previous items (Dunkel, 1999a, p.92, 1999b, p.79).

Ideally, online adaptive placement test of listening comprehension (OAP TLC) should use Item Response Theory (IRT) as the underlying psychometric model for test development. However, IRT-based CALT is an interdisciplinary field of study, involving at least three different areas of knowledge – English, computer science, and mathematics. It really needs a team work and can hardly finish within one year. Alternatively, a compromise was reached; a corpus-based word list was used for test makers to prepare more statistically objective test items.

This OAP TLC is divided into three interfaces – student interface for examinees to take an English placement test, teacher interface for teachers or test makers to prepare, edit and upload their test items, and administrator interface for system administrators to manage all the test items. The following CALT aspects were considered when OAP TLC was first developed. Since OAP TLC shares some similar features with the online adaptive vocabulary test system (OAVTS) that I previously developed, both tests also share a similar interface design with slight modification for OAP TLC.

**1. Difficulty Level of English Words**
It is generally believed that the size of English words that an EFL learner has will definitely affect his/her listening comprehension competence. Most EFL learners should have the experience of concentrating too much on the words that we don’t know so that we miss parts of listening that follow. To most EFL learners, unfamiliar vocabulary words and speech rate are probably two major factors that will affect the competence of listening comprehension (Tsai, 2004).

The following dictionaries were referred to determine the difficulty level of English words shown in the script of listening test items. (Wu, 2005)

A. Collins COBUILD English Dictionary for Advanced Learners (2001) (abbrev. as Collins)

The lexical database of OAPTLC used the same one as OAVTS and was classified into six difficulty levels.

- **Level 0**: includes basic and the most frequently used words, such as the, a, of, you, this, is, are, etc.
- **Level 1**: includes words in the frequency band 5 and 4 of Collins.
- **Level 2**: Longman labels the most frequently used spoken words (S1~S3) and written words (W1~W3). If they have not been categorized into Level 1, then they belong to Level 2.
- **Level 3**: contains words of the frequency band 3 of Collins if they do not belong to Longman’s S1~S3 and W1~W3, plus Macmillan’s two-star entries and three-star entries if they are not in Level 1 and Level 2.
- **Level 4**: includes Macmillan’s one-star entries plus Collins frequency band 2.
- **Level 5**: contains words labeled as frequency band 1 of Collins.

2. **Definition of Item Codes**

The representation of each code is defined as follows:

- 1st code: type (vocabulary, grammar, listening, reading, etc.)
- 2nd ~ 4th codes: difficulty parameter (these three codes are reserved for the item difficulty of IRT logistic model)\(^3\)
- 5th code: approaches (phonemic discrimination, paraphrase recognition, response evaluation, listening cloze, etc.) (Buck, 2001)
- 6th ~ 7th codes: category (general, business, science, medicine, law, environment, etc.)
- 8th code: functions (request for help, request for permission, request for information, understand excuses, understand comments, understand suggestions, understand apologies, understand opinion, etc.) (Hughes, 2003, p.161)
- 9th code: invariable N means “number”.

151
10th ~ 15th codes: serial numbers

All these item codes are given by test makers except the 2nd to 4th codes which are reserved for the item difficulty calculated by an IRT-based logistic model. These item codes will be recorded by database software (e.g. Windows SQL 2000 in Windows-based server or MySQL in Linux-based server) and can be easily retrieved in the future for data analysis.

3. Design of Interface for Teachers or Test Makers

A network-based, user-friendly interface for teachers was designed so that teachers can concentrate on preparing test items. Teachers do not have to memorize the details of the item codes; they only have to choose a proper one from pull-down menus. A foolproof device was also put in OAPTLC; test developers have to assign a difficulty level to a test item before they can submit the question to the test bank. In the future, when a certain IRT model is decided, test makers won’t have to assign a difficulty level to a test item.

![Figure 1. A screen shot of test maker’s interface](image)

**Question Types**

On the one hand OAPTLC should take full advantage of computers and Internet to show striking discrepancies between paper-and-pencil listening tests and computer-adaptive listening tests, but on the other hand computers are not so smart as to deal with open-ended
responses, short-answer questions, or gap-filling on summaries, etc. There are still some technical limitations on question types because of scoring. Here are some question types that have been developed in the prototype of OAP TLC.

1. One Photograph with Four Spoken Sentences

This type of questions is similar to the first section of listening part in the Test of English for International Communication (TOEIC). Examinees will hear the recording only once. In the photograph section, recording will be played 5 seconds after the photograph is shown; there is no play button on the screen. The purpose of this listening section is to offer examinees a context with a photo and hope that they can recognize those frequently used words in certain situations. Because the four statements are not shown on examinee’s computer screen, they have to concentrate on listening and use their short-term memory to find out one statement that most closely matches the photograph. Here is an example:

After you see a photo like the right one, you will hear:
(A) A farmer is fertilizing a tree.
(B) A lumberjack is cutting a tree.
(C) A firefighter is putting down a forest fire.
(D) A logger is transporting a tree.

![Figure 2. A photo shown on examinees’ screen](image)

2. Spoken Utterances and Question Response

In a response evaluation, test takers hear a question followed by three possible responses and choose the most appropriate response. Both the question and the three possible responses given orally. Most questions in this section are wh- questions and are usually related to examinee’s daily life experience. Here is an example:

Test-takers hear:

*Why is there no electricity today?*

They hear:

(A) I prefer to cook with gas.
(B) The power company shut it off for two hours.
(C) The elections are Friday.
Some researchers may have reservation about this type of listening tasks because test takers may understand the utterance perfectly but do not know how to respond it properly (Kitao & Kitao, 1996). In addition, it seems that this kind of test items “simply requires understanding the literal meaning of a number of short, decontextualised utterances” (Buck, 2001, p.65). However, it is a more communicative type of task than many listening tasks such as phonemic discrimination tasks. It is also an essential listening ability to understand wh-questions and quickly respond them properly.

3. Short Conversations and Question Response

A short conversation is presented aurally and then a question is also given aurally. Here is an example:

Test-takers hear:
Man: You must change planes in Los Angeles.
Female: Is there a direct flight?
Man: No, but you could change planes in Seattle instead.
Narrator: What would the woman prefer?
They read:
(A) To take a direct flight.
(B) To change planes in Seattle.
(C) To exchange her money.
(D) To go by train.

Sometimes the listener must go beyond understanding literal words and make a pragmatic inference in order to choose the correct answer. Buck (2001) points out that inference is at the core of language processing and many scholars also advocate that listening comprehension tests should go beyond literal meanings. However, he further mentions that it is not easy to make inference question items because the answers are not explicitly stated in the text (p.147).

4. Mini-Lectures

Understanding academic lectures, technical reports, or business presentations is one of the ultimate goals for English learners. Ideally, advanced ESL learners should acquire both listening comprehension and note-taking skills. In reality, note-taking ability is not easy to evaluate. Some researchers propose gap-filling on summaries for short speech. Instead of being given a complete written text, test makers are given a summary of the passage they are going to hear. in which some of the important content words have been replaced by blanks (Buck, 2001, p.71). Although this technique is highly recommended, Lewkowicz (as cited in Buck, 2001) also found two major problems – (1) it is not easy to make this kind of listening tests, (2) it does require pre-listening activities, which is impossible for a listening comprehension test.
Furthermore, it is not easy to objectively score the words offered by examinees. For instance, whether we should ignore spelling mistakes is still controversial. An alternative option is, again, multiple-choice questions so that computers can quickly and easily respond with feedback in the end of a test. Here is an example we used from a piece of special English news from Voice of America (VOA) website (http://www.voanews.com/specialenglish/2005-04-06-voa2.cfm). Test-takers hear:

I’m Gwen Outen with the VOA Special English Education Report. We continue our reports for students around the world who want to attend a college or university in the United States. This week, we answer questions from two listeners. Richard Lin from China wants to know how to get a scholarship to an American college. Amarkhuu Ayulguisaikhan of Mongolia wants to know the differences among different kinds of financial aid. They are assistantships, grants, scholarships and fellowships.

An assistantship is a job a student does. In exchange, the student receives money or attends classes for free. Graduate students usually get assistantships. The student works about twenty hours a week helping a professor. The student may teach classes, help grade papers and tests, or do research in a laboratory.

A grant is a gift of money to pay for some or all of the costs of college. Unlike loans, grants do not have to be re-paid. Private groups or organizations generally give grants to students who need the money.

Scholarships and fellowships also do not have to be re-paid. A scholarship is financial aid to undergraduate students; a fellowship is the same kind of aid for graduate students. Generally, scholarships and fellowships go to students with special abilities or athletic skills. Some scholarships are based on financial need. Others go to students who live in a certain area.

For example, the University of Missouri in Columbia has two financial aid programs for international students only. The Global Tiger Scholarship is supported by the group representing former university students. In return for scholarship money, the international student agrees to help the group during the school year. The other international scholarship or fellowship at the University of Missouri is called the grant-in-aid program. It provides money to students who need the help, get good grades and take part in university activities.

To get these scholarships, students must complete forms found on the university’s Web site. The address is missouri.edu. Information about scholarships from other colleges and universities is listed on their own Web sites.

For general information about how to get financial aid, go to a Web site called FinAid. The address is finaid.org.

This VOA Special English Education Report was written by Nancy Steinbach. This is Gwen Outen.
They read:

*If an undergraduate who has special athletic skills, what is the financial aid that he/she can most likely receive?*

(A) An assistantship.

(B) A grant.

(C) A scholarship.

(D) A fellowship.

It seems that there is a general consensus among listening researchers that we should use authentic texts for communicative test items. However, it is not easy to find copyright-free authentic texts with realistic spoken language from the target-language use situations. Therefore, many test-developers often look for broadcast materials such as radio programs with scripted texts (Buck, 2001, p.87). Luckily, materials from VOA website are copyright-free. The terms of use in VOA state that “You are welcome to use any material that is published by voanews.com, or you may link to any of the web pages that Voice of America has published on the internet. There is no need to request further permission” (from http://www.voanews.com/english/disclaim.cfm). There is only one condition. You are not allowed to abridge or edit any VOA material which you may use.

**Limitation of the Study**

Dunkel (1999a) points out that listening comprehension proficiency CATs have not been popular because it takes a lot of time and effort to construct them. This is obvious the first limitation of this study. Especially, if a listening comprehension placement CAT based on IRT psychometric model was going to be developed, researchers and experts of English, computer science, and statistics or mathematics should collaborate to construct a testing platform. Second, because the project has been done on a tight budget, this prototype of OAPTLC can only use copyright-free listening materials (such as VOA news story) or royalty-free photos (such as clip art of Microsoft Office online resource) from Internet or use low-end recording software and entry-level hardware to make test items. Third, since the restriction of budget and manpower, no field-test of test items in paper-and-pencil format has been done for the purpose of calibrating the test items. Except the more statistically objective corpus-based word list, test makers also had to listen to recordings a couple of times to decide the difficulty level of test items. Some considerations to take into account are speech rate, accent, examinee’s prior knowledge about the listening topics, and phonological modification of English (such as assimilation, elision, or intrusion).

**Conclusion**

It is a trend that more and more listening tests will make use of the advantages of
computer in the future. Creating a computer-based listening test or an online adaptive listening test take much more money and time. It needs many significant advantages to make that effort worthwhile. The primary advantages of computer-based test delivery include shorter testing times, simultaneous score feedback, and repetitive availability (Buck, 2001, p.255). Since multimedia is one of the strengths of computers, the delivery of interactive multimedia to mass examinees over a network or over the Internet is possible in the near future if the network bandwidth is highly upgraded. The incorporation of video into foreign language instruction and its effect on an ESL/EFL context has been widely researched (MacWilliam, 1986; Garza, 1991; Secules, Herron, & Tomasello, 1992; Meskill, 1996), but the application of multimedia to listening comprehension CALTs seems to be still in its infancy. This is just a prototype of OAPTLC. In the next version of OAPTLC, IRT statistical parameters of the test items will be estimated after their trialing, test scripts will be professionally recorded, and multimedia will be properly integrated into a listening comprehension CAT.

Acknowledgment
The research in this article is funded by a faculty research grant (CHU-93-H-007, November 1, 2004 – October 31, 2005) from Chung Hua University, Hsinchu, Taiwan.

Notes
1. English courses offered in junior for students of level B and C are not credit-bearing ones so that these English courses won’t influence their major credits of graduation.
2. We usually use the weekend of orientation week to administer a computer-based placement test for freshmen. The test is concurrently given in five to six computer rooms, each of which contains about 70 students. This placement test lasts about 70 minutes.
3. The difficulty parameter, denoted by \( b \), of IRT logistic model is defined as the point on the ability scale at which the probability of correct response to the item is .5. The theoretical range of the values of this parameter is \(-\infty < b < +\infty\). However, typical values have the range \(-3 < b < +3\). We are going to use 0 to represent minus and 1 to represent plus. Therefore, -1.5 will be shown as 015.

References


Hypermedia & Telecommunications. Montreal, Canada.
線上適性聽力分級測驗的研究與發展: 期初報告

吳文舜
中華大學外國語文學系講師

摘要

本文的主要目的在探討並設計一套能符合臺灣地區大專院校的「線上適性聽力分級測驗」。此「線上適性聽力分級測驗」是「網路適性化英語分級測驗平台」的第二部份。目前正在研發的「網路適性化英語分級測驗平台」，初期是希望能幫助本校大一新生入學時，在第一週上課時就能依照學生的英語能力分配到適合他們程度的英語課程。在電腦適性化語言測驗 (computer-adaptive language test) 中，每一位考生所考的題目並不完全一樣，下一道題目的出現，是根據這一道題目答對與否而有所變化。如果這一題答對，下一道題目將會比較困難；相反的，如果這一題題目答錯，下一道題目將會比較簡單。電腦適性化語言測驗一般被認為比較快速可以測出學生能力、比較可靠、比較節省成本。大部份電腦化測驗 (computer-based test) 是由電腦伺服器用亂數挑選的方式從題庫中挑選題目給考生，考生考試的題目出現順序雖然不一樣，但是所有考生都考相同的題目。此次「線上適性聽力分級測驗」是希望設計出一套電腦在題庫中選題時能遵照較嚴謹的規範，用比較科學的方式，挑選出接近學生程度的題目。本文也會針對「線上適性聽力分級測驗」上的題型做一分析整理。聽力題庫腳本中的英語用字難易度、題目編碼所代表的意義，以及出題介面等相關問題也會有所介紹。

關鍵詞：電腦適性化語言測驗、電腦化測驗、聽力測驗、聽力分級測驗

收件日：2005 年 4 月 14 日；通過刊登日：2005 年 6 月 15 日