



2010-07-09

# Evaluating the Usefulness of an Aural Gapped Listening Summary as a Measure of Academic Listening Proficiency

Sarah Elizabeth Mottaghinejad  
*Brigham Young University - Provo*

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>

 Part of the [Linguistics Commons](#)

---

## BYU ScholarsArchive Citation

Mottaghinejad, Sarah Elizabeth, "Evaluating the Usefulness of an Aural Gapped Listening Summary as a Measure of Academic Listening Proficiency" (2010). *All Theses and Dissertations*. 2210.  
<https://scholarsarchive.byu.edu/etd/2210>

This Selected Project is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact [scholarsarchive@byu.edu](mailto:scholarsarchive@byu.edu).

Evaluating the Usefulness of an Aural Gapped Listening Summary  
as a Measure of Academic Listening Proficiency

Sarah Elizabeth Appling Mottaghinejad

A selected project submitted to the faculty of  
Brigham Young University  
in partial fulfillment of the requirements for the degree of  
Master of Arts

Dan P. Dewey, Chair  
Neil J. Anderson  
Mark W. Tanner

Department of Linguistics and English Language  
Brigham Young University

August 2010

Copyright © 2010 Sarah Mottaghinejad

All Rights Reserved

## ABSTRACT

### Evaluating the Usefulness of an Aural Gapped Listening Summary as a Measure of Academic Listening Proficiency

Sarah Elizabeth Appling Mottaghinejad

Department of Linguistics and English Language

Master of Arts

For this project I sought to find a more effective means of evaluating academic listening comprehension. This involved doing an in-depth investigation of academic listening, the constructs involved in listening comprehension, and of methods of assessing listening comprehension. It also included a study of the concept of test usefulness (Bachman and Palmer, 1996), which consists of reliability, construct validity, authenticity, interactiveness, impact, and practicality, and is used to help select the most effective methods of assessing language abilities. Based on my review of listening comprehension testing methods, I created a method of assessing academic listening comprehension, Aural Gapped Listening Summaries (AGLS), produced a short version of the AGLS for piloting through BYU's English Language Center and credit exam for matriculated students, and then analyzed the results of this piloting to determine whether future investigation was merited. This project write-up includes a description of the development of the AGLS, the methods of administration, and students' cursory perceptions of the AGLS, as well as the results of the pilot test.

The AGLS involved students listening to an excerpt of a lecture followed by an aural summary of that lecture with every 8<sup>th</sup> word replaced by low-volume static. Then they were asked to type a word or phrase in a box on their computer screens that would best fill in the gap where the static was. Ranks on the AGLS were correlated with a standard listening test, which is administered every semester at Brigham Young University, and with students' individual perceptions of their listening abilities. Results showed that AGLS correlates moderately well with traditional measures of academic listening ( $r=0.7731$ ) while giving testers interesting information about student interlanguage in very little time. Results further showed that AGLS has a much higher reliability coefficient ( $r=0.9223$ ) in comparison to the other listening test. Therefore, although traditionally testers have had to write lengthy tests in order to get an adequate representation of students' listening abilities, it may be possible to obtain the necessary information about students' abilities with this more time-efficient measurement tool.

Keywords: academic listening, summary cloze, noise test, collocational competence, working memory, usefulness

## **ACKNOWLEDGEMENTS**

There are many people to whom I owe a great deal of gratitude; I hope I can do them justice. First, my thanks go to my grandfather for expecting nothing less than excellence, my mother for filling our house with books, and my father for not demanding anything. I should probably thank my husband, too, for his statistical advice and for taking care of our daughter for the hundreds of hours I dedicated to my research.

Thanks also go to Dr. Diane Strong-Krause who helped me come up with the idea in the first place and offered continued counsel even though she was not my thesis chair. Likewise, my thanks to my chair, Dr. Dan Dewey, and the rest of my committee, Drs. Neil Anderson and Mark Tanner, for their patience as I learned what it means to do academic research. Finally, thanks to Russell Hansen who did all the computer programming, and to all the ELLs who endured my test.

## Table of Contents

ABSTRACT .....	ii
ACKNOWLEDGEMENTS .....	iii
Table of Contents .....	iv
List of Tables .....	vi
List of Figures .....	vii
Chapter 1 Introduction .....	1
Research Problem .....	2
Studies Addressing this Problem .....	4
Purpose of this Project .....	5
Chapter 2 Literature Review .....	7
Usefulness .....	7
Reliability .....	7
Validity .....	8
The academic listening construct .....	13
Discrimination .....	14
Authenticity .....	14
Interactiveness .....	15
Impact .....	16
Practicality .....	17
Working Memory .....	18
Phonological Sequencing Capacity .....	20
Collocational Competence .....	21
Listening Assessment Techniques .....	22
Short Answer Method .....	22
True-False Method .....	22
Multiple-Choice Method .....	23
Summary Cloze Method .....	23
Aural Gapped Listening Summaries .....	25
Summary Cloze Studies .....	25
An Experimental Application of “Cloze” Procedure . . . to Listening Comprehension .....	26
An Experimental Application of “Cloze” . . . as a . . . Test of Listening Comprehension .....	27
A New Technique for Measuring Listening Comprehension .....	28
Testing Listening Comprehension in Japanese University Entrance Examinations .....	28
Chapter 3 Developing the Aural Gapped Listening Summary .....	32
Chapter 4 Pilot Testing .....	35
Test Subjects .....	35
Administration .....	36
Survey Questions .....	38
Chapter 5 Quantitative Evidence for Usefulness .....	40

Traditional Test.....	40
Measures of Central Tendency and Dispersion .....	41
Reliability.....	41
Comparison of Matriculated and EAP Students .....	43
Correlation of the AGLS to the Traditional Test .....	45
Variability Across Gender, Native Language, and Age .....	45
Discrimination.....	47
Survey Data.....	47
Chapter 6 Conclusion.....	49
Discussion of the AGLS's Usefulness.....	49
Reliability.....	49
Validity .....	50
Authenticity.....	51
Interactiveness.....	52
Impact .....	52
Practicality .....	53
Further Discussion .....	53
Recommendations.....	55
Limitations .....	56
Suggestions for Future Research .....	57
Conclusion .....	59
References.....	60
Appendix A Transcript of Lecture.....	70
Appendix B Transcript of Summary with Gaps .....	72
Appendix C Table of Acceptable Answers.....	73

## List of Tables

TABLE 2.1	<i>Summary of Historical Types of Validity</i> .....	10
TABLE 2.2	<i>Academic Listening Construct Matrix</i> .....	13
TABLE 2.3	<i>Advantages and Disadvantages of Listening Assessment Techniques</i> .....	26
TABLE 4.1	<i>Examinees' Native Languages</i> .....	36

## List of Figures

FIGURE 2.1 <i>Model of Phonological Short-Term Memory</i> .....	20
FIGURE 4.3 <i>AGLS Practice with Answers and Explanation</i> .....	37
FIGURE 4.4 <i>AGLS Instructions and Blanks</i> .....	38
FIGURE 5.1 <i>Nonnative English Speakers' Aural Gapped Listening Summary Totals</i> .....	42
FIGURE 5.2 <i>Nonnative English Speakers' Traditional Test Totals</i> .....	42
FIGURE 5.3 <i>Tally of Students who Passed Both, One, or Neither of the Tests</i> .....	43
FIGURE 5.4 <i>Differences between Groups</i> .....	44
FIGURE 5.5 <i>Correlation Between AGLS and a Traditional Test</i> .....	46
FIGURE 5.6 <i>Student Opinions of AGLS</i> .....	48



## **Chapter 1**

### **Introduction**

The college years represent significant challenges and opportunities for all people who decide to pursue higher education, especially for those studying in a second language. The language barrier for foreign students in the U.S. was widely recognized in the 1960s when the number of students was estimated at 90,000 annually (Gregory-Panopoulos, 1966). Now, approximately 40 years later, the Institute of International Education reports that there are more than half a million second language English speakers in the United States pursuing bachelor's, master's, and doctorate degrees in a variety of areas (Gardner and Witherell, 2006). The sheer numbers of non-native English speakers inundating American universities has forced administrations to acknowledge the linguistic needs of this population.

Many colleges and universities have programs in place to offer second language students additional support because the challenges for these students are compounded by the complexities of an academic register in a foreign tongue (Gardner and Witherell, 2006). These programs, which may be available pre- or post-entry, range from optional tutoring and writing centers with no proficiency prerequisites to mandatory pre-sessional tutoring and concurrent academic English instruction in addition to a minimum TOEFL score.

At Brigham Young University, in order to be admitted, students are required to have a TOEFL score of at least 580 for the paper-based test or 85 on the internet based test, and they are also offered elective academic ESL classes. Like many foreign language classes, students may “test out” of these classes and still receive academic credit. Approximately 33 students take this credit exam each year. It covers reading, writing, grammar, and listening; and can generally be completed within two to three hours. The goal of this exam is to discriminate between English

Language Learners (ELLs) whose English proficiency would not hinder them in academic life and those learners who would still benefit from further English for Academic Purposes (EAP) instruction.

Listening was the specific language skill chosen for this project because it is possibly the most important skill for learners of Academic English to master to enable higher academic success in a listening-intensive college environment.

### **Research Problem**

Though there are many tests to assess the English language skills of reading and writing, we are still trying to achieve consensus as to what listening is, what affects listening performance, or how to measure it (Bodie, Worthington, Imhof, & Cooper, 2008; Dunkel, Henning, & Chaudron, 1993). Indeed, there was once doubt that listening could be considered a separate linguistic skill at all (Buck, 1992; Vandergrift, 1999). Listening skills are so under-researched in comparison to the other language skills because they are difficult to quantify (Flowerdew & Miller, 2010). Vandergrift (2010) says that of the four skills, “listening is the least understood and most difficult to investigate” (p. 160). How does one study an aspect of linguistic competence that is supposed to have no performance characteristics? Moreover, according to Bodie, et al. (2008), the emphasis on listening skills and behaviors has preempted the need to discover the underlying competencies involved in listening. Bodie and her associates attempt to consolidate the literature on listening from the various fields of cognitive psychology, anthropology, communication, management, and psycholinguistics to produce a cohesive, comprehensive view that encompasses all of the latest research available. The majority of the research they review comes from the perspective that listening is a form of information processing, but it is also perceived as a behavior with individual personality traits and cognitive characteristics influencing the

selection, organization, and integration of information that constitutes listening (Imhof, 2004). It is important to acknowledge the real depth and complexity of listening because relegating it to comprehension alone introduces several problems in assessment. For instance, level of education, critical thinking skills, working memory capacity in the L1, and subjectivity of interpretation can all confound measures of comprehension (Bodie et al., 2008).

Lund (1991) describes a paradox of language practices—listening in the classroom is receiving more focus than ever, and still reading receives more attention in research (see also Buck, 1997; Imhof, 2010; Vandergrift, 2007). He relates that it has been presumed that comprehension is general and principles learned in reading automatically transfer to listening and vice versa. The salient principle from Lund’s research is that “reading and listening are indeed distinct modalities that develop on different schedules and require differentiated instructional techniques” (p. 201). The mechanisms for comprehension may or may not be the same, but learners approach reading and listening and apply strategies to them in vastly different ways, such as more attention to detail in reading and greater attention to main ideas in listening (Lund, 1991).

Most listening assessment practices are informed by reading assessment—almost everything done in listening was first done in reading (Buck, 1997; Lund, 1991; Vandergrift, 2006), yet the evidence indicates it should be the other way around. One major component of reading, after all, is the phonological representation of words. Reading is heavily dependent on the accurate perception and recognition of sounds (Birch, 2002). No matter what language is being read, whether English, which has a somewhat phonological alphabet, or Chinese which is morphemic, it is the phonological part of the brain that is being activated (Chan, 2001; Koda, 2005). In 1987, Wagner and Torgesen found that phonological competence plays a causal role in reading skills

acquisition, and now it is a well-established point of view (Koda, 2005; Richardson, Thomson, Scott, & Goswami, 2004). People learning to read in their first language likely have greater facility than in their second because they bring such a large auditory vocabulary to the task (Birch, 2002). Thus, learners' auditory vocabularies need to be addressed first, which can then be transferred into reading vocabulary. The strategies for decoding auditory information and tying it to meaning by ear may be similar to those needed by the eye, which further indicates the importance of starting reading instruction with materials that are similar to spoken language (Birch, 2002).

These same ideas may be applied to assessment. Conventional assessment practices may be justified in testing reading, but research should be done in assessing listening first, and then that research may be used to inform reading assessment practices. The major criticism of conventional assessment practices is that, "a narrow focus on the right answer to comprehension questions . . . does little to help students understand and control the processes leading to comprehension" (Vandergrift, 2007, p. 191). It is inadequate to find ways that only partially assess listening. There must be a way to assess listening that will help students control and augment their individual listening competencies.

### **Studies Addressing this Problem**

In Flowerdew's (1994) review of academic listening, he states that what we think of as "bottom-up" and "top-down" processing can be very misleading. No one can really say what the difference is between "higher" and "lower" levels of comprehension. He also expresses regret regarding how little work has been done to analyze the discourse structure of academic lectures. Alderson and Bachman (2001) say the assessment of listening is critical to teaching as well as to

testing language proficiency, and yet little has been written about specific constructs, underlying abilities, or designing and validating listening assessment instruments.

There have been many studies conducted that address this problem of finding adequate ways to assess listening skills (Brindley, 1998; Buck, 1988, 1992, 1997, 2001). Ellis (1996) attempts to measure the underlying competence in listening proficiency and traced everything from vocabulary acquisition to learning grammar forms back to phonological short-term memory. Yi'an (1998) identifies the listening processes leading to comprehension, but admitted that the multiple-choice method of testing she implemented in her study "posed threats to the construct validity of the test" (p. 40). Questionnaires, stimulated recall, interviews, and reflective journals have all been used to gain insights into the listening experience (Brindley, 1998; Brown, 1995; Brown, 2002; Buck, 1988; 1997; 2001; Ellis, 1996; Lewkowicz, 1991; Mackey & Gass, 2000; Vandergrift, 2006; 2007; 2010; Yi'an, 1998), but these are often difficult to quantify. Other studies have examined the interaction of task-type, topic, and listening processes in responding to assessment items (Buck, 1988; 1992; Vandergrift 2007; Yi'an, 1998). Of the few studies that have been conducted, however, none have produced entirely satisfactory results (Tafaghodtari & Vandergrift, 2008). Tests still ask narrowly-focused questions, assess skills on the periphery of listening competence, and fail to adequately quantify the listening process (Buck, 2001).

### **Purpose of this Project**

The goal of this project was to develop a useful academic listening assessment technique and find evidence (to be described in the Review of Literature) that it has the potential to circumvent these problems inherent in other assessment techniques. The challenge was to produce a discriminating assessment instrument that maintains reliability and validity while being both time- and cost-efficient. The Aural Gapped Listening Summaries (AGLS) was created in an

effort to meet this challenge because it seemed to be a way to provide a more direct measurement than other techniques. The AGLS showed quite a bit of promise based on past studies of summary clozes, and it is hoped that it can perform the same function as BYU's credit exam for listening proficiency in much less time and with equal or greater validity and reliability, or be used as a tool to screen students for academic work. In this assessment technique, examinees listen to a short, lecture-style listening passage followed by a summary of that passage. In the summary, the first sentence is left intact, but thereafter every  $n^{\text{th}}$  word is removed and replaced by low-volume static. In the case of this study, every 8<sup>th</sup> word is gapped. Examinees listen to the summary, phrase by phrase, and must produce the appropriate word to fill in the blanks. The AGLS was developed as a reaction to the need for a highly reliable, valid, and yet efficient listening measurement tool.

This project explored the questions, *What is the academic listening construct? To what extent does Aural Gapped Listening Summary exhibit the qualities of usefulness (reliability, construct validity, authenticity, interactiveness, impact, and practicality) at BYU? To what extent does an Aural Gapped Listening Summary correlate with more traditional measures of academic listening? To what extent does an Aural Gapped Listening Summary discriminate between people whose English proficiency would not hinder them at an English-speaking university and those who would benefit from further EAP instruction?* The expectation was that the AGLS would be a useful measure of academic listening proficiency and that it would discriminate well between groups.

## Chapter 2

### Literature Review

The purpose of this project was to develop and determine the usefulness of an Aural Gapped Listening Summary (AGLS) in discriminating between those university students who need further listening instruction in order to be successful in English-based higher education and those who do not.

#### Usefulness

According to Bachman and Palmer (1996), the intended use of a test is the most important consideration when designing it, and usefulness is comprised of a balance of “reliability, construct validity, authenticity, interactiveness, impact, and practicality” (p. 17). The discussion of each of these elements of usefulness will be followed by an overview of working memory, phonological sequencing capacity, and collocational competence, which are all good indicators of second language proficiency. Then, there will be a review of several listening assessment techniques which will transition into a discussion about cloze testing and AGLS. Finally, with the intention of justifying the use of AGLS, there will be a detailed summary of four studies that are somewhat similar to the current, exploratory research.

**Reliability.** Reliability is the most “measurable” requirement for a useful test, in that it is relatively easy to estimate the degree of reliability in a test mathematically, and it can be described as consistency in measurement. It is also essential to the concept of validity since no test can have any degree of validity without also displaying some extent of reliability. According to Mackey and Gass (2000), reliability is defined in the following ways:

- + Rater Reliability: assesses the degree to which different raters (or the same rater twice) give consistent estimates of the same language ability.
- + Test-Retest Reliability: assesses the consistency of scores between repeated trials.
- + Parallel-Forms Reliability: assesses the consistency of the results of two tests which are supposed to measure the same linguistic construct.
- + Internal Consistency: assesses the consistency of results across items within a test.

Some factors that might make scores less reliable would be if the test were too difficult and if students were only guessing the answers (this is obviously a higher risk for multiple choice than for other types of items). Other factors that would affect the reliability would be ambiguous questions, having more than one correct answer, and poor rater training. Factors specifically related to listening comprehension would be rate of speech, volume, and sound quality. There are three important things to remember when discussing reliability: (1) reliability is not dichotomous—there are degrees of reliability, (2) test *scores* rather than tests may be considered reliable, and (3) without some degree of reliability, test scores cannot be considered valid (Mackey & Gass, 2000).

Reliable test scores are generally obtained from longer exams because these tests contain enough data points that it would be unlikely for students to exhibit skills that they do not actually have. It is important to note that length can promote reliability but does not guarantee it. Further, a test can't be too long because eventually a student's test fatigue will start decreasing reliability. That's an important concept to this study because one of the goals is to be time-efficient without sacrificing reliability.

**Validity.** Construct validity is the second requirement for a useful test. While there are many types of validity historically discussed in the literature, Bachman and Palmer (1996) limit their



discussion to construct validity because it can be argued that most categories of validity are subsumed by construct validity and those that are not fall under the heading of Bachman and Palmer's other five criteria for usefulness. However, this review of validity research will include those traditional types of validity to give historical perspective to the discussion.

Validity is the extent to which adequately reliable scores are interpreted and used appropriately, whereas construct validity refers to whether a scale measures or correlates with the theorized linguistic competence. The primary use of language tests is to make inferences about language ability and thence decisions about individuals based on those inferences (Bachman & Palmer, 1996). Tests arbitrate educational policies, teaching pedagogy, institutional decisions, and principles of language theory and research (Cumming & Berwick, 1996). For this reason, test validation is one of the most important aspect of language assessment.

With all the research on the concept and process, validity has been fairly well-defined but is difficult to pin down because it is a unitary construct with many overlapping facets. Historically speaking, the types of validity have been divided into anywhere from five to eighteen subcategories, not all of which are mutually exclusive, and any combination of which could be used to validate a test. The canonical types that Angoff (1988) proposed are summarized in Table 2.1, arranged in alphabetical order.

Although consequential validation is a major concern today, consequences of a test do not seem to be included in Angoff's historical discussion of validity, nor are cognitive complexity (what Bachman and Palmer [1996] might call interactiveness) or fairness. Linn, Baker, and Dunbar's (1991) categories of transfer and generalizability (generalizing the results of a test to other populations, tasks, or time periods) could fall under concurrent, construct, content, convergent, criterion, divergent, population, task, and temporal validity. Their content quality

Table 2.1  
*Summary of Historical Types of Validity*

<b>Type</b>	<b>Definition</b>
concurrent validity	correlation of the test against another; simultaneous measure of the same construct that has been previously validated
construct validity	correlation of the measuring instrument against the theoretical language ability it is trying to measure—includes both convergent and divergent validity
content validity	verification that the test contains a satisfactory sampling of the target subject matter
convergent validity	correlation among different methods of testing the same construct
criterion-related validity	correlation of test performance and real-life performance—includes both concurrent and predictive validity
divergent (discriminant) validity	little to no correlation between different constructs measured by the same methods
ecological validity	the extent to which the testing situation, methods, and materials, as well as the examinees' exhibited behaviors, approximate real-life; authenticity
face validity	the appearance of validity to the test participants and users
factorial validity	strong correlation of each measurement item with the one construct it is related to, and a weak (or insignificant) correlation with all other constructs; established by factor analysis (Grefen, 2005)
operational validity	matching what the test measures to the test's function
population validity	variation in validity coefficients across populations; i.e., being able to apply test results to other people
predictive validity	correlation between observed scores on the test with future performance
task validity	the amount of variation in validity coefficients across tasks
temporal validity	the amount of variation in validity coefficients across time

*Note.* Adapted from Angoff, 1988 p. 21-28.

and coverage are included in content and ecological validity, while meaningfulness is really included under all of Angoff's categories.

Messick (1989, 1994, 1996) changed the way language assessment professionals talk about validity by proposing that validity is a unitary construct with many facets, and by suggesting that we cannot tell what our tests really measure; we can only ask what evidence there is for the interpretation and use of test scores (Alderson & Banerjee, 2002).

All of these various components of validity still leave unanswered questions: *Which aspects are more important? If an exam fulfills some, but not all, validity requirements, should it still be*

*considered valid? How do you determine if it fulfills any of the requirements? And can a test really be valid at all?* Because of issues like these, the definition of validity has seen many revisions since the inception of validity research in psychometrics approximately 80 years ago—and in the last 20 years in particular. Most researchers have come to believe that *construct validity* is *the* central principle in validity and all others are appendages to it (Moss, 1992). The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurements Used in Education (NCMUE) state that construct validity is “the most important consideration in test evaluation” (AERA, 1999, p. 17) and validity is considered “a unitary concept requiring multiple types of evidence to support specific inferences made from test scores” (Moss, 1992, p. 234).

Because of the centrality of the construct in validation, particularly in Bachman and Palmer’s (1996) model of test usefulness, it helps to know what a construct is, as well as what the specific construct to be assessed is before gathering validity evidence for a rating scale or testing technique. A construct is a theoretical linguistic competence such as nursing English or academic English listening proficiency. According to Alistair and Ahmed (2009), this is probably the most important component of validation because it is more proactive than other validation measures—it focuses and directs the test writer in the construction period. The construct influences the choice of what to include in the test and is the light with which to judge how important different elements are. However, there is no clear agreement on how to define the competencies involved in listening. Because of Nichols’ (1948) definition of listening as the retention of orally presented information, the emphasis in listening research has been on measurement, leapfrogging the development of a sound theoretical basis for the listening construct (Bodie et al., 2008), hence the numerous taxonomies of listening situations with no real

hierarchy and no way of knowing if they sufficiently encompass this nebulous idea called listening. The problem with defining a listening construct is that it must be based on an observable product or output (Bachman & Palmer, 1996), while listening by nature is unobservable because of its nonproduction. As soon as second language production is involved, the listening task also becomes a speaking or writing task—at the very least, spelling or pronunciation could be significant hindrances in the assessment task. To avoid testing irrelevant constructs, specific definitions are needed to “provide a basis for using test scores for their intended purpose(s), to guide test development” (Bachman and Palmer, 1996, p. 98), and to enable construct validation. There are different ways to define a construct, including the competence-based, task-based, and task-competence-based approaches (see Bachman & Palmer, 1990; 1996; Buck, 2001).

Buck (2001) proposed an all-purpose listening construct as one that assesses

the ability to 1) process extended samples of realistic spoken language, automatically and in real time; 2) understand the linguistic information that is unequivocally included in the text; and, 3) make whatever inferences are unambiguously implicated by the content of the passage. (p. 114)

This is a vague definition and adaptable to any listening assessment situation, but it *does* need to be adapted to specific scenarios if it is to be of any use. For example, Buck’s construct as applied to AGLS would be described as assessing the ability to 1) process a two-and-a-half minute speech sample of realistic, spoken, and academic language at a normal speech rate, automatically and in real time; 2) understand the linguistic information that is unequivocally included in the text; and, 3) make linguistic inferences about missing words based on the content of the passage and examinees’ own collocational competence.

***The academic listening construct.*** Buck (2001) identified five types of knowledge involved in listening comprehension that much of the literature has in common: pragmatic, semantic, syntactic, lexical, and phonological. Psycholinguistic theory states that these interact to facilitate each other. Richards (1983) first made a distinction between conversational listening and academic listening. As such, academic listening has received even less attention than listening as a whole. Eleven years after Richards, Lynch (1994) proposed the four different—yet similar—categories of grammatical, discourse, pragmatic, and sociolinguistic knowledge in his Academic

Table 2.2

*Academic Listening Construct Matrix*

<b>Language Competency Required</b>	<b>Formal Lectures</b>	<b>Student Presentations</b>	<b>Instructions/ Assignments</b>
<b>Grammatical Knowledge:</b>			
phonological modification			
stress/intonation			
spoken vocabulary/slang			
oral syntax			
repetitions and false starts			
<b>Discourse Knowledge:</b>			
discourse markers			
rhetorical schemata			
story grammars			
asides/jokes			
separating main points/details			
<b>Pragmatic Knowledge:</b>			
basic functions/conveying ideas			
manipulating, learning, or creating			
indirect meaning/hints			
pragmatic implications			
text-based inferences			
<b>Sociolinguistic Knowledge:</b>			
appropriate linguistic forms			
informal			
idiomatic expressions			
local dialect			
cultural references			
figures of speech			

*Note.* Adapted from Lynch, 1998, p. 271.

Listening Construct Matrix (see Table 2.2). The Table is designed to help researchers and educators better characterize their target constructs.

There is much that listening comprehension shares with reading comprehension, such as decoding and the receptive nature of both skills. However, listening does have features unique to itself, such as phonological and lexico-grammatical features, as well as a distinct rhetorical structure and real-time processing. It is this real-time processing that is most interesting for this study. Listening texts occupy time rather than space. They must be perceived and understood as the words are uttered (Flowerdew, 1991). Much of what comprises the academic listening construct will be delineated in more detail as part of the authenticity discussion.

***Discrimination.*** Inherent in the construct validity of this study is discrimination because the goal is to separate students into two proficiency groups. Ferguson (1949) defined discrimination as “the number of relations of difference” (p. 61) test administrators can draw from an exam. In other words, discrimination is a test item’s ability to separate students into clearly defined groups, and a test that is designed to discriminate should be able to separate students into those who typify the proposed test interpretation and those who do not. For the AGLS, this means separating students into university-ready students and non-university-ready students. This is relevant to a discussion about validity because it is essentially an item’s ability to precisely identify those examinees with the intended proficiency (construct).

***Authenticity.*** Authenticity is considered by some as part of validity, while others think it subsumes both validity and reliability; it is Bachman and Palmer’s (1996) third requirement for a useful test. The question of authenticity naturally arose in the 1970s when communicative language teaching was becoming more popular (Lewkowitz, 2000), and has been debated ever since. Authenticity is generally defined in the field of linguistics as any text created for a native

speaker by a native speaker, but in testing it has more to do with how closely the testing method has approximated real life (Brown, 2003). Even though there is not strict agreement on the definition of authenticity, there is a certain amount of accord among assessment theorists as to its importance (Lewkowicz, 2000). There is a temptation to speak of authenticity as a dichotomous, absolute quality, but like reliability and validity, authenticity comes in degrees (Breen, 1985).

In order to achieve some extent of authenticity in listening tasks as contrasted with reading tasks, one must consider rate of speech, phonological modifications, and the discourse structure of spoken language. If listening is assessed by reading the written word aloud, that ignores the differences between written and spoken text (Buck, 1992). With regard to academic listening specifically, there is a certain amount of planning involved in lectures, which gives them some of the characteristics of written discourse, but as a spoken genre lectures also share much with interpersonal conversations. There are false starts, redundancies, repetitions, gap fillers like “um,” and “so,” and cues from body language (Flowerdew & Miller, 1997). Lectures are also often accompanied by visual aids. One particularly distinctive feature of academic listening is “micro-structuring.” Lectures are organized by intonation contours, usually comprised of incomplete clauses, and pauses or micro-level discourse markers like “and,” “so,” “but,” “now,” and “okay” are frequently used as organizational signposts (Flowerdew, 1994). Lecturers may or may not ask students questions and modify the lectures according to immediate student needs, but this is one side of authentic academic listening that is very difficult to capture in the testing environment.

**Interactiveness.** Interactiveness is the fourth consideration for a useful test. It refers to the extent that an examinee’s mental resources are engaged by a task. These mental resources include linguistic knowledge, metacognitive strategies, background knowledge, and affective

schemata. Real life linguistic tasks may vary as to what they require of their participants, which is why Bachman and Palmer separate it from authenticity. Interactiveness is the link between authenticity and construct validity. When attempting to measure authenticity and interactiveness, one must consider the characteristics of the examinees, the features of the real-world task that the testing situation is supposed to predict performance in, and the assessment task itself.

**Impact.** Impact, or washback, is yet another matter to consider when evaluating the usefulness of a test. It could also be considered part and parcel of validity, and it has to do with the consequences of test use and interpretation. Hughes (1989) defines washback as “the effect of testing on teaching and learning” (p. 1). When dealing with the impact of a test, one would analyze the values, ideologies, and broader framework guiding the construct, as well as the long-term effects that actually result from test implementation (Messick, 1989).

Tests are administered for a purpose, and thus imply goals and values associated with the results that impact people’s lives (Bachman, 1990). The consequences of a test can be positive or negative. Buck (1988), for example, raised the concern that “noise-tests,” where students listen to a passage with words replaced by a beep and supply the missing word, would cause students to focus on studying “mutilated passages,” which could not be good for their linguistic development and would be considered negative washback. Consider also the TOEFL; a nurse could be perfectly competent and able to function in an English-speaking hospital, but may not have the academic English ability to pass the TOEFL, and is, therefore, barred from a job. This negative impact, deciding that this nurse is not qualified to work in an English-speaking hospital based on a TOEFL score, is an invalid interpretation of the test results. Likewise, consider the Gaokao, which is a Chinese college entrance exam. It is considered the most pressured exam in



the world and students have been known to commit suicide because of it (Cumming and Berwick, 1996). In this instance, the use of test scores is consequentially invalid. Even tests, well-written, fairly administered, with scores interpreted and used appropriately, can have negative consequences.

On the other hand, tests may also have positive consequences, such as pinpointing areas of improvement for language learners, motivating them to progress even further, or winning them a higher salary. The three things to consider when assessing the impact of a test are the experience of preparing for and taking the test, the feedback examinees receive as a result of their performance, and the decisions made based on the test scores (Bachman & Palmer, 1996). It is a test-writer's responsibility to consider the impact, positive and negative, of a test and plan accordingly. A high-stakes test necessitates a lot of effort whereas a low-stakes test does not.

For the assessment of academic listening, positive effects would include students listening to authentic lectures and learning academic vocabulary to prepare for such an assessment. For this study specifically, a positive consequence would be identifying students who could benefit from concurrent ESL classes to help them cross the threshold into academic English. A negative impact that is a potential problem for any test is the stress an examinee may feel preparing for, taking, and receiving the results of the test. This is a very real problem for many students, and can hurt student performance and the reliability of test scores.

**Practicality.** After all the other criteria of a useful test have been considered, practicality is probably the biggest barrier to a well-written test becoming a useful test. It comes down to a deceptively simple equation of available resources divided by required resources. If the quotient comes out to be less than one, then the resources available are insufficient to appropriately

implement a test and the test is, therefore, impractical. How impractical the test is depends on how far from one the quotient is.

The “resources” that must be factored into the practicality equation include things like man-hours required to write the test and the technology and time needed to administer it. The resources a test consumes must not exceed the benefits of the test. One could argue that the assessment of academic listening cannot be authentic without the visual stimuli of a real lecture, but the resources may not be available to produce a test of such caliber, so the ideal may be sacrificed in the face of practicality. Practicality is different from the other criteria of usefulness in that practicality is not so much about the quality of the test as it is about whether the test can actually be created and implemented (Bachman and Palmer, 1996).

In sum, “usefulness” in testing is a balancing act where the test-writer evaluates the reliability, authenticity, interactiveness, and impact in light of the test’s construct and practicality.

### **Working Memory**

Now that it has been established what is required of a useful test, theoretical support for the construct validity of the study at hand will be provided.

According to Gass and Selinker (2008), “working memory refers to the structures and processes that humans use to store and manipulate information” (p. 250). According to Baddeley (2003), “The theoretical concept of working memory assumes that a limited capacity system, which temporarily maintains and stores information, supports human thought processes by providing an interface between perception, long-term memory and action” (p. 829). In 1974, Baddeley and Hitch described a central executive with two slave systems—the visuospatial sketchpad and the phonological loop. This view is not without opposition, but most theories

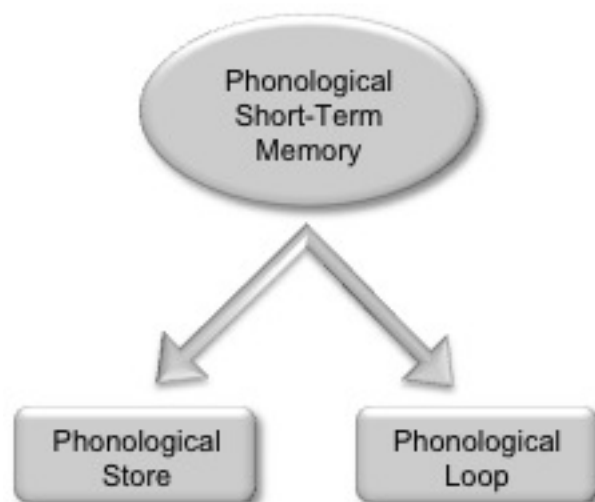
agree that working memory must be a system of limited capacity with peripheral storage systems (Baddeley, 2003).

The phonological loop and the central executive systems are indispensable in “monitoring comprehension and organizing input in a meaningful way” (Imhof, 2010, p. 106). The task for listeners is to store information as they create the mental representation of the text so it can be accessed later and corrected if necessary. Ohata (2006) noted “L2 short-term memory is often overloaded, causing words to be purged before they can be organized in L2 patterns and then interpreted” (p. 22). One early study, Glicksberg (1963), reported moderately high correlations between scores on a test of memory for linguistic input and listening comprehension. Loe (1964) investigated target language memory span for long sentences and grammatically complex sentences, and found that both native speakers and advanced students recalled sentences containing clauses better than sentences of the same length made up of a series of phrases. On the other hand, less proficient students found the sentences made up of phrases easier to remember. These results suggest that more proficient speakers have learned to make use of complex syntax to group linguistic data efficiently. Harris (1970) found relatively high correlations between scores of grammatical and content accuracy in short-term recall (now known as working memory) and tests of listening comprehension and usage. Rivers and Temperley (1978) pointed out that short-term memory is much more limited in a second language than in native language comprehension. Call (1985) reported that memory for sentences in isolation was an excellent predictor of listening success, and she concluded that “[working] memory for auditory input is an important component of listening comprehension” (p. 776). More recently, Gass, Roots, and Lee (2006) found that the more proficient the learner, the closer second language working memory matched first language working memory.

These studies indicate three salient principles. First, working memory is limited for L2 input. Second, working memory capacity is closest to a second language learner's L1 capacity in the most proficient ELLs. Third, the complexity of sentences influences what can be retained in working memory. All of this is important because listening comprehension requires the ability to hold phonological information in working memory long enough to interpret the statement and manipulate the linguistic information contained therein. For reduced redundancy tests like AGLS, the redundancy can only be restored if the listener has the capacity to store the possible interpretations long enough to accept or reject the validity of each (Imhof, 2010).

### **Phonological Sequencing Capacity**

Phonological short-term memory is one of the slave systems of working memory that consists of the phonological store and the phonological loop (Baddeley & Hitch, 1974) (See Figure 2.1.). This is the “system that is responsible for the temporary storage and manipulation of speech” (Speciale, Ellis, & Bywater, 2004, p. 294). According to Ellis (1996), individual differences regarding phonological short-term memory (STM) determine how well learners acquire new vocabulary and grammar. Other sources, cited in Speciale, et al. (2004), maintain



*Figure 2.1 Model of Phonological Short-Term Memory*

that phonological sequence learning contributes to “the segmentation of speech into discrete units, identification of the lexical units of language, and the development of automaticity in their processing” (p. 294). The results of the Speciale et al. study verify the relationship between phonological STM and L2 vocabulary learning. The ability to hold aural input and its serial order long enough to allow sufficient processing to take place plays a critical role in listening comprehension, and O’Brien, Segalowitz, Freed, and Collentine (2007) found that phonological memory capacity actually predicts oral fluency gains in a second language. Since, among other things, the AGLS directly assesses a student’s ability to remember a phonological sequence (examinees have to hold each phrase in their minds as they figure out the correct answer), Ellis’s research, the Speciale et al. study, and the O’Brien et al. study indicate some degree of predictive validity for AGLS. Results from the AGLS could predict future success in university studies because of the vocabulary-intensive nature of most fields.

### **Collocational Competence**

According to Lewis (1997), fluency in a foreign language is a condition of the acquisition of a number of prefabricated phrases, and he regards this as the central feature of language learning. Others (Bahns & Eldaw, 1993; Fontenelle, 1994; Herbst, 1996; Lennon, 1996; Moon, 1992) distinguish fluency as a fundamental function of native speakers’ communicative competence. As speakers of a language, people in different communities favor certain expressions through and because of repeated use (Wray, 2000). Ellis (1996) said that, “Speaking natively is speaking idiomatically using frequent and familiar collocations, and learners thus have to acquire these familiar word sequences” (p. 97). Keshavarz and Salimi (2007) used a cloze test to examine collocational competence, and AGLS can also be construed to assess collocational competence to the extent that it is used as a strategy to complete this task when solely remembering breaks down.

## **Listening Assessment Techniques**

With the theoretical basis for the AGLS in mind, let's examine different listening assessment techniques. A test-taker's performance on a listening test is a function of the ability and assessment method (Yi'an, 1998). Any testing method should be studied and have its pros and cons weighed before writing a test. In the end, the appropriate assessment method is identified by the balance of Bachman and Palmer's (1996) six elements of a useful test. The AGLS was critically scrutinized and compared to other, more popular ways to assess listening, and what follows is an abridged description of those alternatives to the AGLS, along with definitions and criticisms.

**Short answer method.** Examinees listen to a passage and answer questions that require responses anywhere from a few words to paragraph length. If these are to be effective measures of listening, these answers must be very short so they do not rely too heavily on reading or writing ability. The rating of responses, however, can be work-intensive and subjective (an obstacle which is not as evident in the AGLS). Moreover, according to Brown and Hudson (1998), short answer tasks focus narrowly on assessing a few phrases or sentences and multiple answers are possible. The question must be written very carefully to constrain the number of appropriate responses so that examinees may produce similar correct answers.

**True-false method.** Examinees listen to a passage and decide if statements are true or false based on the information given in the passage. Evaluating whether something is true or false is a common purpose for listening, but the disadvantage is that test-takers have a 50/50 chance of getting the answer correct. This could be ameliorated by having a third "I don't know" or "Not enough information" option, but students invariably guess anyway and Barger and Doherty (1992) say that true-false items do not work well "because of the fleeting nature of the spoken

word and the natural and desired fact that listeners focus on what is said and not on what is not said” (p. 315). Another disadvantage of true-false items is that they emphasize trivial facts and details, and they tempt test writers to write “tricky” items (Brown & Hudson, 1998).

**Multiple-choice method.** Examinees listen to a passage and answer questions by choosing among a number of provided answers. Multiple-choice items are easy to score and are high in internal reliability but still open to possible guessing. It is difficult to develop well-formed questions, and the task does not resemble authentic language use. Because they are also pervasive, Brown and Hudson surmise that multiple-choice testing has limited the scope of skills teachers and language assessment professionals are testing.

The principal criticism of all three of the previous techniques, as stated by Vandergrift (2007), is that focusing on the minute answers to comprehension questions “does little to help students understand and control the processes leading to comprehension” (p. 191). It is hoped, though not conclusive, that the AGLS will help students do what Vandergrift suggests.

**Summary cloze method.** The final assessment alternative is the summary cloze. Cloze testing was the inspiration for the AGLS. There is considerable debate about what exactly cloze tests measure (Abraham & Chapelle, 1992; Brown, 2002), but they are generally used in testing reading comprehension. Every  $n^{\text{th}}$  word of the passage is removed and students have to fill in the blanks. They are different from normal fill-in-the-blank tests (also known as rational deletion clozes) because fill-in-the-blank can be isolated statements, and the words that are gapped are chosen by the test writer, not by counting out a regular number of words (Coombe, Folse, & Hubley, 2007). The theoretical basis for this technique is that there is a certain amount of redundancy built into texts to accommodate for mishearing, misapprehension, missed words, etc. and that a proficient learner should be able to restore redundancy to a text with blanked-out

information (Buck, 1988). What we hear is often a product of what we expect to hear, and the more proficient learners are, the clearer their expectations and the less those expectations are colored by their native languages. A summary cloze is just an adaptation of the cloze technique.

For summary cloze exercises, examinees listen to a passage, and then read a summary of the passage with every  $n^{\text{th}}$  word deleted. Filling in gaps in a summarized version allows for flexibility in text and topic and produces a large number of items for increased reliability. Test marking is objective but not restricted to exact words or phrases. A cloze may also have a number possible answers, which is one disadvantage of short-answer questions, but that could actually be a good thing in a cloze test because it can provide more information about students' interlanguage than a more traditional assessment method. For example, students may try to use a word that is a common collocation with the words before or after the gap, or they may use a synonym; thus showing the stages in their linguistic development. (See Yamauchi, 1990 for a discussion of the diagnostic power of cloze testing.)

Summary clozes have two very attractive advantages over other assessment formats in that they are comparatively easy to write and grade. All it takes is selecting an appropriate passage (which must be done regardless of the assessment technique), writing and recording the summary, and setting the gaps. Even though it is technically a constructed response test, there is a finite number of acceptable responses and it is generally quite easy for a native speaker of English to determine whether a response is syntactically and semantically appropriate (Buck, 1988). Furthermore, summary clozes have an excellent correlation with other whole-language assessment formats such as the TOEFL (in its older versions, at least) (Oller, 1983). Chapelle and Abraham also came to the same conclusion in a more recent study (1990). Another advantage of cloze testing is that, according to Templeton (1977), clozes have higher reliability



coefficients than other measures of proficiency and consistently discriminate well between groups.

Cloze gaps, however, are not easy to set and require careful pilot testing and moderation of the rating rubric. Lewkowicz (1991) correlated the results of his summary cloze with those of a more traditional listening exam and said that “though there is significant overlap between the two, the listening summary cloze tests skills other than listening” (p. 29). Some researchers conjecture that summary clozes may end up testing reading comprehension instead of listening comprehension, because, traditionally, students listen to a passage and read a summary with gaps. Others object that they are too cognitively demanding.

**Aural Gapped Listening Summaries.** It is the cognitive demands of the summary cloze that make it so attractive, and in response to other language testers’ scruples about using summary clozes due to the reading aspect, it is not difficult to overcome that specific drawback by making the summary cloze an *aural* summary, instead of a summary that is read. This technique, dubbed Aural Gapped Listening Summary (AGLS), uses a recorded summary of a lecture that ELLs have listened to, in which every eighth word has been replaced by low amplitude static to indicate the missing word. The cognitive demands of this task make it ideal for discriminating between two groups of ELLs. (See Table 2.3 for a comparison of the advantages and disadvantages of the five assessment techniques discussed.) Most importantly, the AGLS can be said to involve working memory, phonological sequencing capacity, and collocational competence.

### **Summary Cloze Studies**

Only four other studies like this one were found. They provide some significant insights into and evidence of the usefulness of the AGLS as a measure of second language listening.

Table 2.3

*Advantages and Disadvantages of Listening Assessment Techniques*

<b>Technique</b>	<b>Advantages</b>	<b>Disadvantages</b>
Short Answer	+ easy to produce, quick to administer	+ contains numerous possible answers + narrowly focused
True-False	+ provides simple indication of what has been understood	+ contains high guessing factor + contains “tricky” items & trivial details
Multiple-Choice	+ ensures less guessing + encompasses a wide range of principles	+ contains moderate guessing factor + inauthentic + overused and limiting
Summary Cloze	+ easy to construct, quick to administer + flexible + requires background knowledge in text	+ contains numerous possible answers + may test reading rather than listening
AGLS	+ easy to construct, quick to administer + flexible + requires background knowledge in text + contains numerous possible answers	+ may assess spelling or pronunciation on some items

**An experimental application of “cloze” procedure . . . to listening comprehension.** The first study was conducted by Dickens and Williams in 1964. The Cloze Procedure was a relatively new concept at the time—introduced by Taylor in 1953. Because Taylor had so much success with the technique in measuring readability and reading comprehension, and because most research concerning listening began with reading, the extension of the Cloze Procedure to listening comprehension was natural. Dickens and Williams took two professionally recorded “speeches”—one expository and one persuasive—from the Sequential Tests of Educational Progress (STEP) created by Educational Testing Service (ETS), and replaced words with random noise at five-second intervals and added five seconds of silence at the end of each sentence to give test-takers a moment to record their answers. Two groups of ESL students (n=126 and n=127) listened to the two passages. One group listened to the intact passages and then took a

traditional multiple-choice test, whereas the other group listened to the passages with the blanks and wrote down appropriate words to fill the gaps. The correlation between test subjects' scores on the different passages for the traditional test was  $r=0.37$ , while the correlation for the subjects taking the cloze was  $r=0.73$ . Though the split-half reliability estimates for the traditional test were quite low ( $r=0.45$ , persuasive;  $r=0.60$  expository), the numbers were much higher for the cloze ( $r=0.80$  persuasive,  $r=0.70$  expository). Dickens and Williams concluded that an "oral cloze procedure appears to have some advantages over certain multiple choice tests as a research technique for studying the comprehension of spoken messages" (p. 108).

**An experimental application of "cloze" . . . as a . . . test of listening comprehension.** The second study was conducted as part of a dissertation defended in 1966 by Gregory-Panopoulos. This exploratory research was a reaction to dissatisfaction with the then-popular Brown-Carsen Listening Comprehension Test. Gregory-Panopoulos was working with an analog tape of a 20-minute-long lecture, and physically cut out every fifth word and spliced in tone-tape of equal length to the discard using Scotch tape. He experimented as Dickens and Williams had done before him with different sounds to avoid jolting examinees, and he left the first, last, and 17 other paragraphs intact, for a total of nine "mutilated" (p. 57) paragraphs. Examinees were given a paper with numbered blanks, introduced to the technique, and asked to listen to the lecture, simultaneously filling in blanks. Gregory-Panopoulos correlated scores on the cloze with scores on the University of California English Placement Test, the Brown-Carsen Listening Comprehension Test, the California Reading Test, and STEP: listening portions. He also correlated his test with a reading test because he cited research claiming that reading ability is a good predictor of listening ability. He reported an internal reliability coefficient of  $r=0.926$  where  $p<.01$ , which was significantly higher than the coefficients for the tests that the cloze was

correlated with. Gregory-Panopoulos concluded that “the Cloze Procedure appeared to validate satisfactorily” with the Brown-Carsen listening test and the California reading test with correlation values as high as  $r=0.793$ .

**A new technique for measuring listening comprehension.** The third study was conducted by Templeton in 1977, more than ten years after the second. He used two five-minute passages and examinees first listened to the original passage and then again with every fifteenth word bleeped out. He reported that “The KR20 coefficient was .96, and the alternate forms coefficient was .95” (p. 295). Considering how low reliability usually is for any test—let alone listening tests—this was very high indeed and probably cloze testing’s greatest strength. The correlation coefficient between test scores and teacher ratings of students’ ability was  $r=0.91$ , and after making a correction for the imperfect reliability, that coefficient went up to  $r=0.98$ , which is approaching a perfect correlation. Templeton named two other advantages to the cloze, namely, it is quick and easy to administer, and it is reasonably easy to write. Testers merely choose a listening text and create numerical deletions; the items discriminate well no matter the difficulty of the passage, and it seems to assess competence rather than performance. The problem with this method is that it is possible to complete the task without fully understanding the passage, and it becomes more of a memory test and less of a linguistic problem-solving assessment. It may be capable of measuring aspects of listening related to working memory, but it fails to capture other important facets of listening comprehension. Using a summary of the passage, rather than the same passage twice, may avoid this issue.

**Testing Listening Comprehension in Japanese University Entrance Examinations.** The final study was conducted by Buck in 1988, more than thirty years after the first. His goal was to find listening assessment techniques that would be appropriate for inclusion in Japanese

university entrance exams. He criticized so-called noise tests for their washback effect. A noise test is like an aural cloze—students listen to a passage with words replaced by white noise. He felt like students would focus on listening to passages with words removed in order to study for such a test, and this would be bad for their linguistic development. Following a suggestion from Alastair Pollit, Buck wrote a summary of the test passage, which he “mutilated.” Buck used a method of trial-and-error to find the proper configuration for the gaps in the summary. The examinees read the summary of the English passage in Japanese and filled in the blanks. On the repeated occasions of administering such a test to more than 400 Japanese ELLs, Buck found reliability coefficients higher than .80, and its correlation with other measures of listening were higher than for measures of reading. Buck concluded that such a test has a firm theoretical foundation with encouraging results from trial administrations with sufficient face validity. He also felt that the washback would be beneficial because students would have to understand the passage in its entirety to complete the task.

The criticism Buck encountered with his study is that a summary cloze is no longer a test of reduced redundancy as cloze tests are supposed to be. A spoken passage has a certain amount of redundancy built in, but a summary is likely to have little to no redundancy, and the gaps may not be removing redundant information, but core information. This did not concern Buck, however, because “that doesn’t seem very important if the method produces good results” (p. 28), which it did. Furthermore, removing core information is the key to summary clozes; a test cannot get at meaning if only articles and their ilk are removed. He further praised summary clozes as an alternative to normal clozes because they avoid the issue of students listening for a single word and not actually trying to understand what they are hearing.

What all three studies lacked was any discussion of the effect of background knowledge on examinees' performance. One of the biggest dangers in a cloze is favoring those with more background, but by making the test a summary of something the examinees have all listened to, everyone taking the test is given the background knowledge necessary to complete the task *if they understand the passage*. True, those with more background knowledge will understand the passage more easily, but those with the target proficiency should still understand it. And if they do not understand the passage, they have no hope of completing the task effectively even though they may have the same background knowledge in their native languages because their skill is not up to par.

There are essentially two ways to assess listening. The first is to create a task so authentic that it is indistinguishable from the real world. That way, any non-listening factors are acceptable because they are part of the communicative task. However, while this is admirable, it should not put the assessment of the core of listening—the on-line processing of oral information—at risk (Buck, 1997). The other way to assess listening is to create a task that isolates and examines the underlying competencies. This can be the more challenging way because the goal is to prompt authentic thinking (Buck, 1997).

All of the preceding information, then, is the justification for creating an Aural Gapped Listening Summary assessment tool, and we must ask

1. To what extent does Aural Gapped Listening Summary exhibit the qualities of usefulness (reliability, construct validity, authenticity, interactiveness, impact, and practicality) at BYU?
2. To what extent does an Aural Gapped Listening Summary correlate with more traditional measures of academic listening?

3. To what extent does an Aural Gapped Listening Summary discriminate between people whose English proficiency would not hinder them at an English-speaking university and those who would benefit from further EAP instruction?

These questions are important because they tell us whether we are measuring what we want to measure, and if the test is worth the energy required to write and administer it.

### Chapter 3

#### Developing the Aural Gapped Listening Summary

The AGLS was inspired by the description of gapped summaries in Alderson (2000). Gapped summaries were originally conceived for the assessment of reading, and most applications of the technique to listening assessment have involved the examinees reading a summary of the passage that they listened to. The problem is that this may assess reading instead of listening (Buck, 2001). My adaptation attempts to avoid assessing reading by using an aural summary in order to isolate listening. My summary was also written in a very different way from how the summaries are written for assessing reading (see below).

The AGLS consisted of a three-minute lecture excerpt and a summary of the same passage. The passage discussed the cognitive demands of reading out loud. The summary was created by extracting phrases containing core information and leaving those phrases mostly intact, modifying where necessary to maintain cohesion. The wording was not changed as much as might be expected in a summary (and what is usually done in summary closes) because the examinees would need to remember the words to go in the blanks. By making the task a summary of a passage that examinees first listened to, background knowledge was at least partially controlled for. The passage—what would be expected from a standard general education class—was chosen because the topic was obscure, yet accessible; it did not require any field-specific knowledge to understand it. This gives this pilot study some degree of content validity.

The passage was not a written text read out loud. It was an authentic listening passage, from a podcast found through iTunes U<sup>®</sup>, originally created by a native speaker of English for native



speakers of English. It was rerecorded in a soundproof recording studio so that background noise and other factors of sound quality would not affect comprehension, and also so that the same voice would be used in both the passage and the summary. The passage and the summary were recorded at a rate of speech of about 150 words per minute, which was identified by Williams (1998) as the most comfortable for native speaker comprehension. It also featured the planned, yet somewhat conversational style typical of academic lectures.

The first line of the summary was intact, but thereafter every eighth word was replaced by low-volume static using Audacity (Mazzoni, 2009). Every eighth word was decided upon after pretesting several different configurations of gaps because every tenth word for this passage ended up being mostly function words (articles, helping verbs, and, etc.) and every sixth word was too difficult for the examinees to understand, and every twelfth word was too easy. Cobb's (2009) cloze generator at [http://www.lex tutor.ca/cloze/n\\_word/](http://www.lex tutor.ca/cloze/n_word/) was used to gap the passage with these different configurations. It is significant to note that upon reading the different configurations, it was easy to see that gapping every eighth word would be appropriate. Buck (1988) also reported intuitively knowing which gap configuration would work best even before pilot testing. In the case of this study, gapping every eighth word removed a good balance of function and content words with no more than two blanks per thought-group. This seemed to provide enough information to be able to fill in the gaps without being too easy. The gap configuration could change depending on where the counting off of blanks begins and on the different listening passages used. There were twenty-one blanks in all.<sup>1</sup>

Also during the pretesting period, it became apparent that the examinees required instruction and practice for how to take the AGLS. The examinees complained that the AGLS would be

---

<sup>1</sup> See Appendix A for a transcript of the lecture and Appendix B for a summary of the lecture with blanks.

hard even for native speakers, so I determined to test a small sampling of native speakers so that I could be sure not to demand a higher standard of performance from the examinees than could be expected from the average native speaker. The pretesting period identified a few technical mistakes, as well, that were made in the creation of the AGLS, and I was able to rectify them before piloting the AGLS.

The recording of the summary was split up into sixteen sound files by phrase to ensure maximum comprehension without overwhelming students with too much to do at one time. The phrases were all logical thought-groups with one intonation contour. For example, the sentence, “Reading out loud, and being able to continue reading through paragraphs and pages, / without much stopping or pausing, / and speaking clearly, while reading unfamiliar, unmemorized texts / happens to kick several major areas of your brain into action,” was split in four pieces heard in isolation based on the intonation contours.

Anything would be an acceptable answer that made semantic and syntactic sense.<sup>2</sup> For example, in the sentence “\_\_\_\_\_ brain has to process this visual input \_\_\_\_\_ speech output,” acceptable answers for the first blank included *the*, *your*, *our*, and acceptable entries in the second blank were *into*, *and*, *as*, *while at the same time doing the*, and *with*.

---

<sup>2</sup> See Appendix D for a table of acceptable and unacceptable responses.

## **Chapter 4**

### **Pilot Testing**

This chapter contains the description of the quantitative research conducted for this project. The study is designed to determine the usefulness of an Aural Gapped Listening Summary (AGLS) as a measure of academic listening proficiency at Brigham Young University. To this end, several types of evidence were collected to argue for this assessment technique's validity and reliability. The test was piloted with students of a wide range of linguistic abilities. Test data were collected and analyzed, along with survey data. Discussions in this chapter will deal with the demographics of test subjects, procedures, and data analysis of the two listening tests, followed by a brief discussion of the survey portion of the study.

#### **Test Subjects**

The examinees of this study were 42 students from levels 4 and 5 at the English Language Center and 49 fully matriculated university students from BYU who had achieved a score of at least 580 on the paper-based TOEFL or 85 on the internet based test. In addition to the nonnative English speakers, there were 20 native speakers of English, undergraduate students at BYU randomly recruited using flyers around campus, who provided a baseline for the cut score. The examinees were of various ages ranging from 18 to 42 with a mean of 26, median of 25, and mode of 24, and they spoke fifteen different languages with the majority speaking Spanish followed by Korean, Portuguese, Chinese, and others (see Table 4.1) Seventy-one percent were female and the balance were male.

Table 4.1  
*Examinees' Native Languages*

<b>Native Language</b>	<b>Speakers</b>
Spanish	38
Korean	15
Portuguese	12
Chinese	6
Russian	6
French	3
German	2
Japanese	2
Bambara	1
Farsi	1
Malagasy	1
Nepalese	1
Norwegian	1
Turkish	1
Ukrainian	1

### **Administration**

The test was administered six times over a period of four semesters. Students sat at computers at one of two different testing centers. They were allowed to take notes as they listened to the lecture in order to increase ecological validity and authenticity. Lynch (1998) says we must make allowances for an ELL's shorter short-term memory by letting students listen more than once, but this study investigated ways to discriminate between those who no longer require any sort of second language listening instruction to be successful at the university and those who could benefit from more training, so no such allowances were made for memory. Because students usually only listen to a lecture once, letting students listen only once gives the test more authenticity. Although students could only listen once, they could choose to listen to each phrasal unit of the summary up to three times, and in whatever order they chose.

The AGLS was inserted into the middle of the traditional listening exam so the students would be warmed up and also not feel pressed for time when they completed it. Before students began listening to the passage, they went through a tutorial that taught them about the AGLS, and they completed three sample questions. Then they were given the answers to the sample questions and allowed to read the justification for the answers (see Figure 4.3).

When they were ready to begin the actual test, students pressed play. Once they came to the summary, a counter counted down each time they clicked play so they would know what they had listened to and how many times. They listened to one phrase at a time, not necessarily in order, and then they typed the word or words that they thought fit best in the blank box next to

The screenshot shows a software window titled "English Credit Exam: Listening". The main content area is titled "LISTENING: Part C Practice. Follow the directions below to prepare for Part C." and includes a "Practice Questions" section with navigation buttons (left arrow, MENU, right arrow). The interface is divided into sections for listening practice and answer explanation.

**Section A:** "First, click this button to hear an example passage." followed by a button labeled "A. 1". Below it, the instruction says: "Next, click this button (A) to hear the first sentence of the example passage."

**Section B:** "Click this next button (B) to hear the next sentence with words removed. Type the word(s) to fill in the blanks provided:" followed by a button labeled "B. 1". Below this, a listening passage is provided: "You heard, 'Thousands \_\_\_\_\_ new airplanes were coming off assembly lines \_\_\_\_\_ needed to be delivered to military bases \_\_\_\_\_.' Two words would work best in the first gap: *of* and *more*. Either word makes sense grammatically and according to the meaning of the passage. For the second gap, you have two sentences being joined together, so you would need a coordinating conjunction like *and*, *or*, *nor*, *for*, *yet*, *but*, or *so*. For this sentence, *and* makes the most sense. On the third gap, it would make the most sense to tell where the military bases are. So, something like *worldwide*, *nationwide*, *around the world*, *in Europe*, *overseas*, etc. would make the most sense." To the right of the passage are three input fields labeled "1.", "2.", and "3.".

**Section C:** "Click the next button (C) to hear the next sentence with words removed. Type the word(s) to fill in the blank provided:" followed by a button labeled "C. 2". To the right is an input field labeled "4.".

**Section D:** A button labeled "D. 2" followed by two input fields labeled "5." and "6.".


Figure 4.3 AGLS Practice with Answers and Explanation

English Credit Exam: Listening

**LISTENING: Part C.** First, listen to the passage about reading out loud. You will listen to the passage only once, but you may take notes. Make sure you answer the question about the passage. Then listen to a summary of that passage one sentence at a time (A through N). You may listen to each sentence three times. Every eighth word of the passage has been removed and replaced by noise. You will need to complete the sentence with the appropriate word(s). The first sentence that you listen to will not have any words deleted.

Questions 26-47

← MENU →

1 

26. Why is reading aloud such a challenging activity?

a. Because you have to scan slightly ahead.  
 b. Because you have to read while you speak.  
 c. Because it requires several areas of your brain.  
 d. all of the above

A. 3

B. 3 27.  28.

C. 3 29.  30.

D. 3 31.  32.

E. 3 33.  34.

F. 3 35.

G. 3 36.  37.

H. 3 38.

I. 3 39.

J. 3 40.  41.

K. 3 42.  43.

L. 3 44.  45.

M. 3 46.

N. 3 47.

Figure 4.4 AGLS Instructions and Blanks

the play button. For those phrases with two blanks, there were two boxes side by side—the first answer went in the first box and the second answer in the second (See Figure 4.4).

### Survey Questions

At the end of the test, students were asked to respond to two questions. Survey data can provide important information about research that cannot be obtained from more objective methods. It can be extremely varied and has more to do with observations and opinions. The survey questions were asked to address the issue of face validity. In order to satisfy this motive, the two questions that students were asked are as follows:

+ In your classes, do you feel you understand your professors . . . ?

0% of the time    
  25% of the time    
  50% of the time    
  75% of the time    
  100% of the time

+ The aural gapped listening summary is where you filled in the blanks as you listened to a summary of a lecture. How well do you feel it measured your listening ability?

very bad    
  bad    
  okay    
  good    
  very good

The first question, which provides evidence of concurrent validity, was asked to ascertain whether student perceptions of their own listening ability coincided with what the AGLS told testers as well as with the more traditional measure of academic listening. If a test says that a student should do well in a lecture hall, but that student feels like he or she is always a step behind the lecture, then there could be detrimental consequences for the student. Similarly, if a student is barred from entering a university because it has been inaccurately determined that he or she does not have the linguistic ability necessary to study, it is not fair to the student. By asking the students how they feel about their language abilities, one can evaluate how well confidence coincides with a passing score and self-doubt with a failing score.

The second question has to do with the face validity of the AGLS. Even if a test is superbly reliable and valid, if the users of the test scores, namely students and administrators, do not *feel* that it is valid, they will not put much confidence in the results.

## Chapter 5

### Quantitative Evidence for Usefulness

This chapter will disclose the results of the statistical procedures used in the analysis of the usefulness of the AGLS as a measure of academic listening proficiency at BYU. Among other things, the AGLS was compared to the traditional listening test administered at BYU as part of the ESL credit exam using a number of statistical procedures.

#### Traditional Test

The traditional computerized listening test was comprised of thirty-nine multiple-choice questions. Portions of the test were written by me under the direction of Dr. Diane Strong-Krause, who wrote the rest of the test. Dr. Strong-Krause is an assessment specialist. Her dissertation for a Ph.D. from BYU dealt with automated assessment, and she is coordinator of ESL testing at BYU. She teaches undergraduate and graduate courses in language assessment and has chaired many MA committees for theses and projects about language assessment.

The test covered lecture comprehension, vocabulary, vocabulary in context, listening for details, listening for main ideas, and discourse features. One section required students to listen to an isolated statement and choose an appropriate response among three options. For example, one statement said, “Did he borrow his sister’s car?” and the appropriate response was, “Yes, she let him use her car.” Other questions were standard, short-lecture listening passages followed by multiple-choice comprehension questions. One passage described some artwork and students had to choose which painting was being described among four pieces that were painted by the same artist. Students listened to each passage twice and were allowed to both preview questions and take notes. Dr. Strong-Krause and I thoroughly analyzed and rewrote each question for construct



validity and reliability over a period of two years of testing and retesting. The Cronbach's Alpha coefficient for this test was  $r=0.895$  and Dr. Strong-Krause judged it to be a relatively reliable and valid measure of academic listening proficiency.

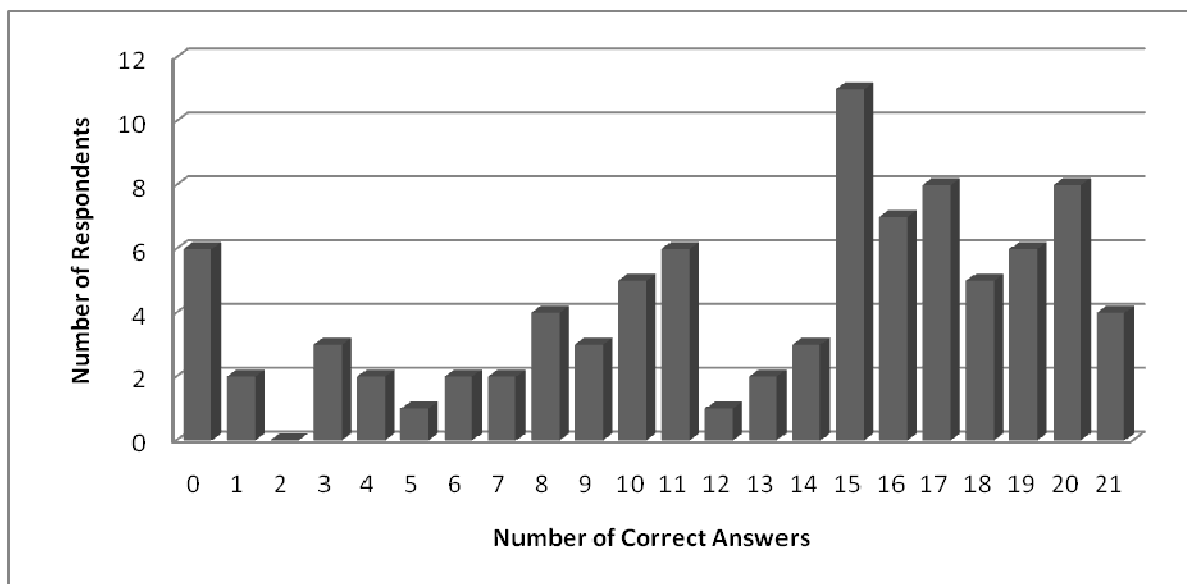
### **Measures of Central Tendency and Dispersion**

As shown in Figure 5.1, the scores for the AGLS were not normally distributed (based on a Lilliefors Test for Normality), while the scores for the traditional test were normally distributed to a greater extent, but still not sufficiently to employ parametric statistical procedures, as illustrated in the histogram in Figure 5.2. Since the data were not normally distributed, all of the statistical procedures used in this project will be nonparametric because, as Buck (1994) stresses, researchers are obliged to select procedures that fit their data.

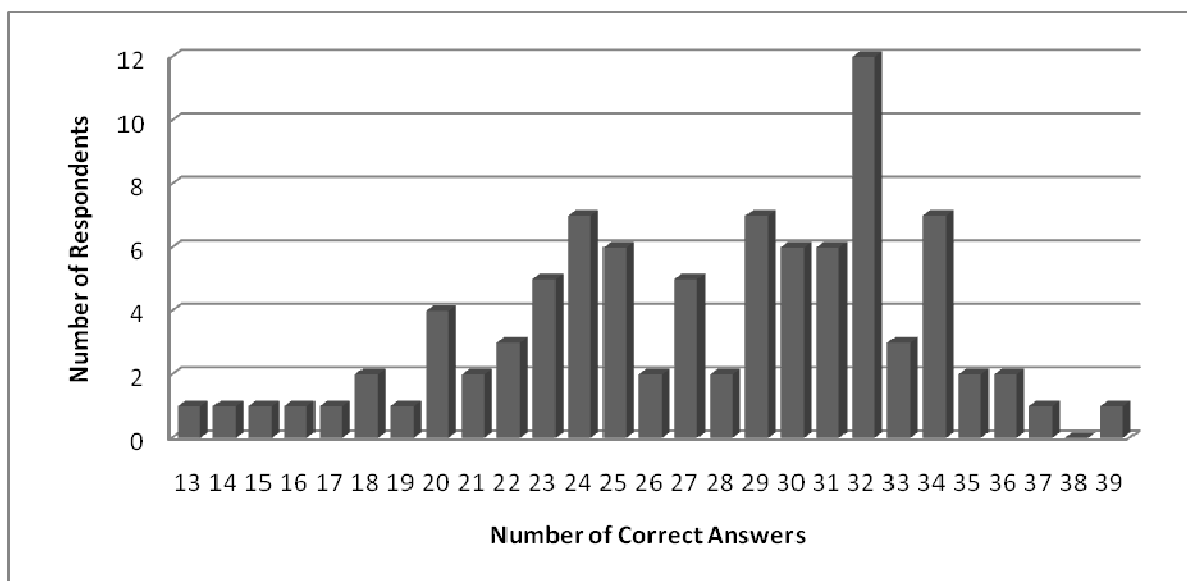
The mean, median, and mode of the AGLS were 61%, 71%, and 71% respectively with a standard deviation of 29%, whereas the mean, median, and mode of the traditional test were 70%, 74%, and 82% with a standard deviation of 14%. The measures of central tendency are skewed for the AGLS because there were a number of examinees who did not provide any words to fill any of the blanks, thus the large number of zeros in the score totals as seen in Figure 5.1. If the scores of those examinees who declined to answer are excluded, the mean, median, and mode of the AGLS are 67%, 71%, and 71% with a standard deviation of 24%, which is still quite large compared to the traditional.

### **Reliability**

The AGLS was also analyzed for reliability using Cronbach's Alpha and using the split-half method to provide evidence of reliability and task validity because a test cannot have any degree of validity without also being reliable (Brown and Hudson, 2002). The AGLS was determined to be 92.23% percent reliable using Cronbach's Alpha test of internal consistency. The split-half



*Figure 5.1 Nonnative English Speakers' Aural Gapped Listening Summary Totals*



*Figure 5.2 Nonnative English Speakers' Traditional Test Totals*

correlation was  $r=0.8625$ , but with the Spearman-Brown Prophecy correction, the correlation became  $r=0.9262$ , which is comparable to the Alpha value. These are excellent reliability coefficients for any test, but especially for a test administered on such a small scale (Hughes, 2003).

## Comparison of Matriculated and EAP Students

The cut scores for both tests were determined by averaging the scores of the twenty native speakers and subtracting their standard error. The average native speaker's score was used because the aim of the test is to discriminate between students whose language would not hinder them in English-speaking universities and those whose language skills are insufficient. The standard error was subtracted because the "real" native speaker average could have been that low and we want to give students on the cusp the benefit of the doubt. The cut scores came out to be 86% and 79% for the AGLS and traditional tests respectively. Of the 49 matriculated students, 31 passed the AGLS; none of the EAP students passed. Forty students passed the traditional test, two of whom were from the EAP group. Thirteen students who passed the traditional test failed the AGLS, and four of the students who passed the AGLS failed the traditional exam. However, those students' scores were well within a standard deviation of the cut score on the test that they failed. See Figure 5.3 for an illustration of the students who passed and failed each test.

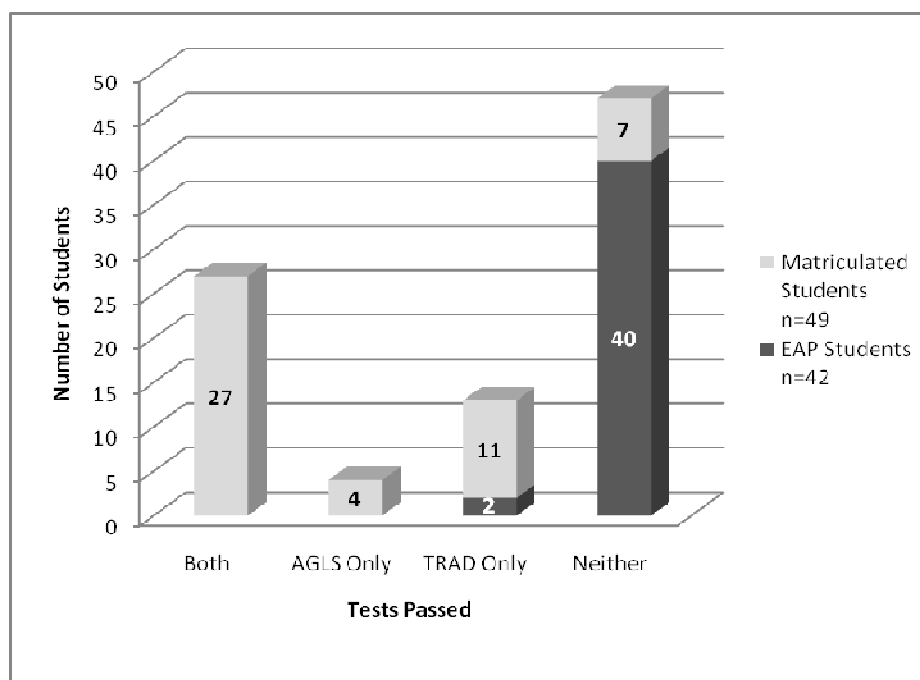


Figure 5.3 Tally of Students who Passed Both, One, or Neither of the Tests

Of the seven matriculated students who passed neither test, three reported that they understood 100% of what their professors said, three said they understood 75%, and one said that he did not understand anything. There could be no follow-up interview with these students because identifying data was not collected, so it is uncertain whether the test is underestimating their abilities or they are overestimating their abilities.

A Mann-Whitney U-test, used to compare the two groups, showed  $W=2428$ ,  $p<.00001$ . Figure 5.5 shows a comparison of histograms between the two groups. As expected, the matriculated students, who had passed the TOEFL, generally did better than the students still preparing for the TOEFL.

The Cronbach's Alpha reliability coefficient for the matriculated students was  $r=0.7842$  and the coefficient for the EAP students' tests was  $r=0.8596$ . The difference in reliability is slight, but it could have been caused by a number of factors. The matriculated students could take the test whenever they wanted during testing center hours, and they were also taking the full credit exam (including reading and writing portions) to test out of elective ESL classes, whereas the

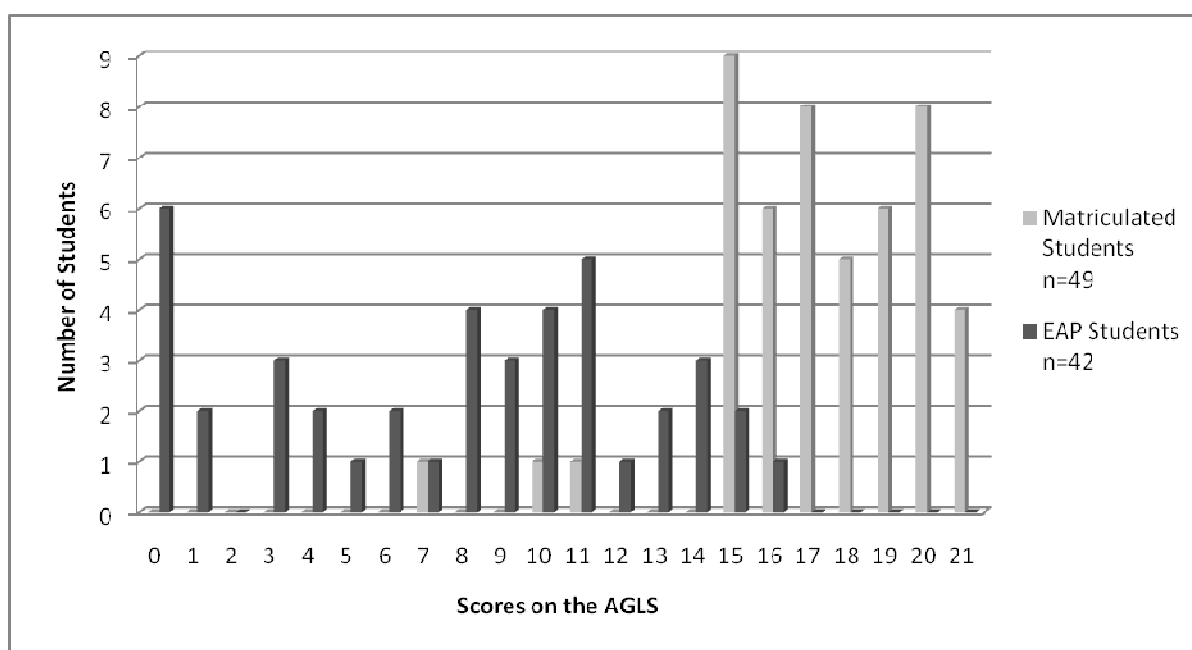


Figure 5.4 Differences between Groups

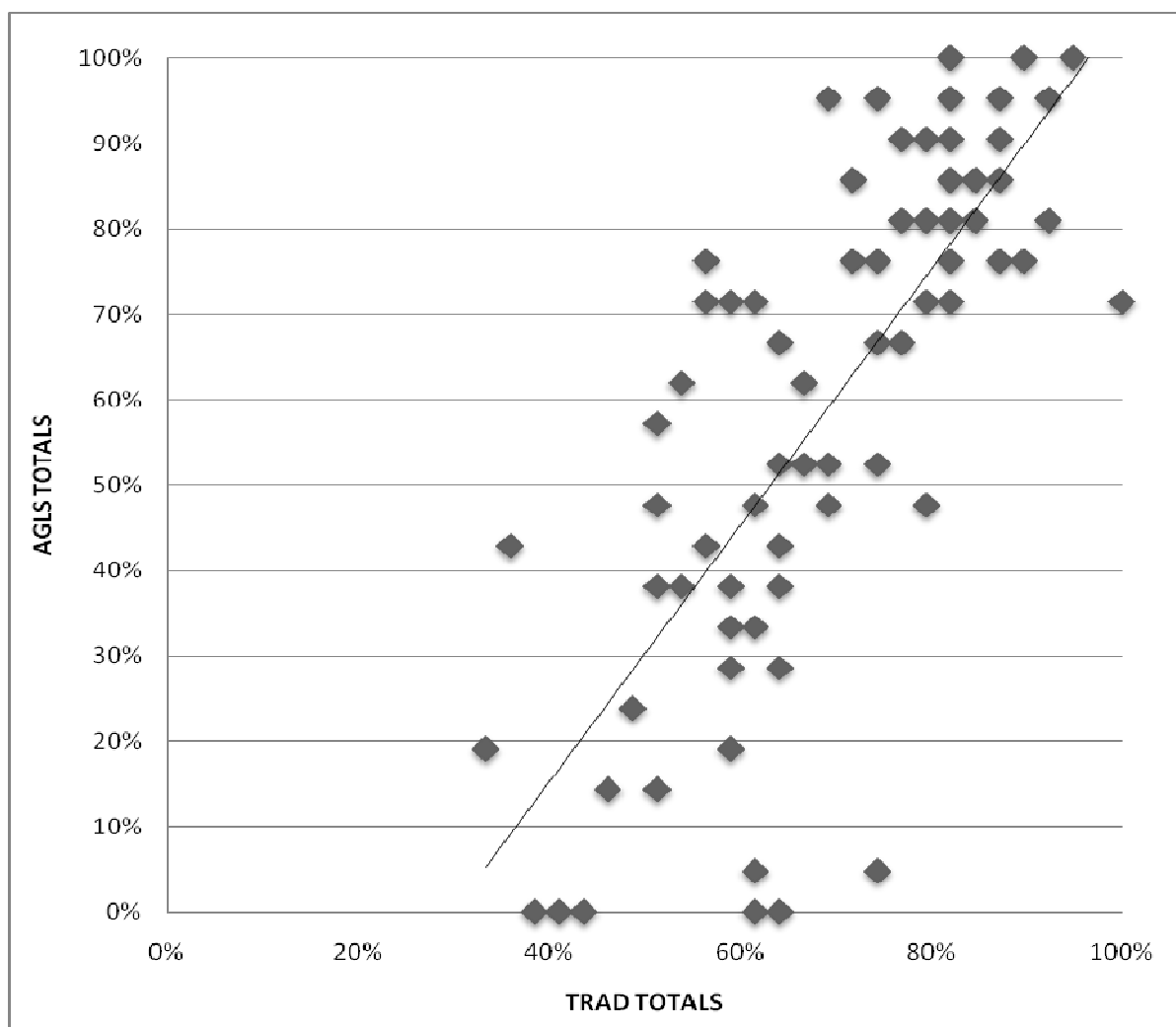
EAP students had to come and go at specific times, were watched over by a test proctor, only took the listening portion of the exam, and had no stake in the exam. All of these factors could affect reliability for the two groups of test takers.

### **Correlation of the AGLS to the Traditional Test**

The percentages of accurately answered items for the AGLS were correlated with a traditional listening test using a Spearman Rho correlation test, which provides evidence of concurrent and convergent validity. The correlation was  $r=0.7731$ ,  $p<.00001$  (see Figure 5.4). The results were also compared using a Wilcoxon Signed Rank test, which is the nonparametric version of a matched-pairs t-test, to assess whether there was a significant difference between rankings on the two exams. The comparison of the tests using the Wilcoxon Signed Rank test yielded  $V=855$ ,  $p<.00001$ . A correlation shows the extent to which one measurement can predict the outcome of another, and a correlation of  $r=0.7731$  means that if a student does well on one test, it is reasonable to expect the student to do well on the other test. However, the results of the Wilcoxon Signed Rank test mean that students generally got higher percentages on the traditional test than on the AGLS.

### **Variability Across Gender, Native Language, and Age**

To analyze the variability of scores across gender, a Mann-Whitney U-test was used. Kruskal Wallis tests were used to analyze the variability of scores across first language and age. These tests were chosen because they are the nonparametric alternatives to a between-groups t-test and ANOVA, respectively (Sheskin, 2000). These supplied evidence of population validity, fairness, and generalizability. There is some evidence to suggest that men and women process language differently, and it could be that these different modes of processing give one sex the upper hand



*Figure 5.5 Correlation Between AGLS and a Traditional Test*

in the AGLS (Putrevu, 2001). Further, first language and age could have a significant effect on how well students are able to cope with this cognitively demanding task, and we do not want an assessment technique with a group bias.

A comparison of gender and age exposed no differences ( $W=352$ ,  $p=0.6025$ ;  $c^2=2.6131$ ,  $df=3$ ,  $p=0.4552$  respectively). The comparison between the largest three language groups did show a significant difference for first language ( $c^2=10.569$ ,  $df=2$ ,  $p=0.0051$ ). Further inspection revealed that Asian students got an average of 62% on the traditional test and 33% on the AGLS. Indo-European students scored an average of 74% on the traditional test, and 72% on the AGLS.

The Asian students as a group did much worse on the AGLS than on the traditional test, whereas those students speaking Indo-European languages did about the same on both tests. All but one of the students who failed to supply any answers for the AGLS were Asian. However, even with those students' scores removed from the analysis, the Asian students still did much worse on the AGLS than on the traditional test. More research is needed to find the cause of the phenomenon.

### **Discrimination**

The next step in analyzing the quantitative data was to determine the degree of discrimination. The item discrimination (percent correct for the pass group minus the percent correct for the fail group) for each item on the AGLS ranged from .59 to .94, with the average being .80, which exceeds the levels Kehoe (1995) identified as indicative of a good test. Moreover, the results of a Mann-Whitney u-test comparing rankings on the traditional test of the group that passed the AGLS and the group that failed the AGLS (determined by averaging the native speakers' scores minus their standard error) showed  $W=198$ ,  $p=0.0019$  (Sheskin, 2000). It can be concluded that there is a highly significant difference in rankings on the traditional listening exam for the high and low proficiency groups as determined by the AGLS. This means that the AGLS separates high and low proficiency very efficiently.

### **Survey Data**

The results of the comparisons with survey data were less informative than the comparisons with the Traditional Test. The correlation coefficient between students' self-perceptions and their score on the traditional measure of listening was  $r=0.1841$ , and the correlation between their perceptions and the AGLS was  $r=0.3421$ . When the matriculated students' and EAP students' data was parceled out, the correlations for the matriculated students were very

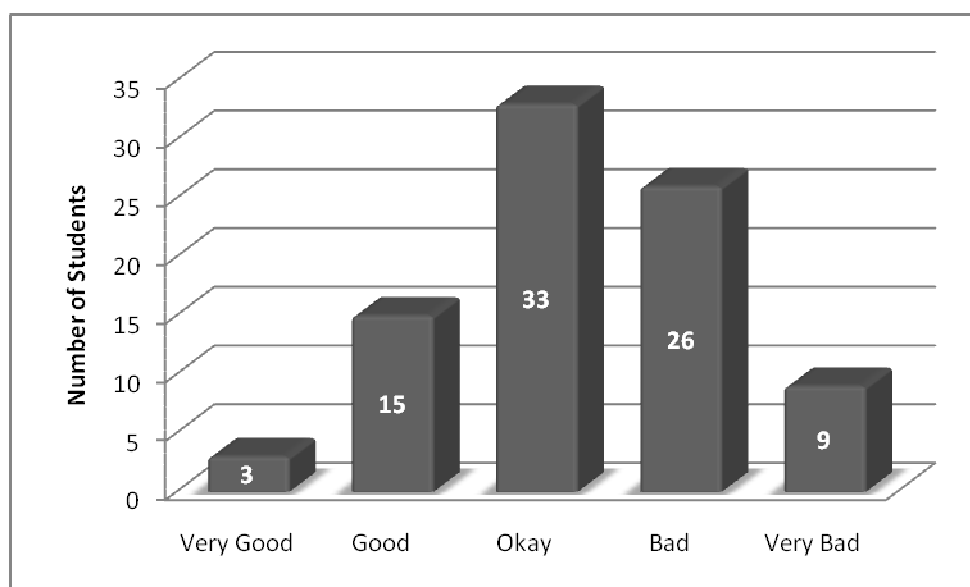
Table 5.1

*Correlation of Students' Self Perceptions and Scores on the Traditional Test and the AGLS*

<b>Group</b>	<b>Traditional Test + Self Perceptions</b>	<b>AGLS + Self Perceptions</b>
All	r=0.1841	r=0.3421
Matriculated	r=0.4094	r=0.4567
EAP	r=0.1817	r=0.4622

similar, whereas the correlations for the EAP students nearly doubled from the traditional test to the AGLS (see Table 5.1).

As shown in Figure 5.6, 4% of the students thought that the AGLS was a very good measure of their listening abilities, while 18% thought it was good, 38% thought it was okay, 30% thought it was bad, and 10% concluded that it was a very bad measure of their listening abilities. In-depth qualitative research would be helpful in determining if students rated the test poorly because they were afraid they had done poorly, or for some other reason. It is significant to note that all of the Asian examinees rated the AGLS bad or very bad. This may partially explain the animosity Buck (1988) reports receiving from his Japanese test subjects during his study of summary clozes.



*Figure 5.6 Student Opinions of AGLS*



## Chapter 6

### Conclusion

Evidence has been collected to evaluate the potential usefulness of an Aural Gapped Listening Summary at Brigham Young University by answering these questions:]

1. What is the academic listening construct?
2. To what extent does Aural Gapped Listening Summary exhibit the qualities of usefulness (reliability, construct validity, authenticity, interactiveness, impact, and practicality) at BYU?
3. To what extent does an Aural Gapped Listening Summary correlate with more traditional measures of academic listening?
4. To what extent does an Aural Gapped Listening Summary discriminate between people whose English proficiency would not hinder them at an English-speaking university and those who would benefit from further EAP instruction?

#### **Discussion of the AGLS's Usefulness**

A test is only as good as how useful it can be—useful to students, teachers, and administrators. Thus a test must display a degree of reliability, construct validity, and authenticity, as well as acceptable levels of interactiveness and positive impact while maintaining practicality.

**Reliability.** The Table of Acceptable Responses (see Appendix D) was used to ensure rater reliability on the AGLS (the traditional test, being multiple choice, did not require such a measure), and while the estimates of internal consistency showed that the traditional test was

very reliable ( $r=0.895$ ), the AGLS was even more reliable. The reliability coefficient for the AGLS was quite high at  $r=.926$ . This percentage is even higher than that of the traditional test and is a good score for any examination. It is probably so reliable because lucky guesses are nearly impossible. While the test-retest method of estimating reliability was not employed because of practicality issues, the other three types of reliability estimates all show the scores on the AGLS to be very consistent.

**Validity.** Reliability is a key factor determining the validity of a test. It is impossible to make strong inferences regarding the results of a test if its level of reliability is extremely low. In this respect, the high level of reliability of the AGLS contributes greatly to its validity level. In addition to reliability, a number of other forms of evidence for validity were also collected. First of all, the construct of academic listening was defined as collocational competence and the ability to hold linguistic information in working memory long enough to manipulate it in order to restore missing information. As shown in Chapter 2 (Review of Literature), there is evidence that the AGLS involves these competencies. The proposed interpretation of the AGLS was that successful examinees have these competencies and therefore need no further EAP instruction in order to succeed in academic contexts, and it would be recommended to those who fail that they take elective ESL classes.

Mathematically, the AGLS displayed a certain amount of concurrent validity. A correlation of .7731 between the two listening tests is not a perfect correlation, so one would suspect that they measure similar things (providing some evidence for concurrent validity), but the fact that the correlation is not extremely high could indicate that the two tests may be measuring slightly different constructs. Oller (1983) and Abraham and Chapelle (1992) found that cloze tests are highly correlated with measures of whole-language proficiency. Similarly, Buck (1988) found a

sound theoretical basis for summary clozes, as did Keshavarz and Salimi (2007) among others, which gives the AGLSs some construct validity. It is possible that the cloze format approaches more global academic listening comprehension abilities than the traditional test, but further research is needed in this area to make any sound conclusions.

The AGLS demonstrated excellent discriminant validity; it is very good at separating low and high level learners. Face validity does present some problems. As a non-standard assessment technique, the AGLS seems to be met with a variety of reactions from skepticism to animosity. It is especially alarming that the face validity seems to be much lower for Asians than for those who are native speakers of Indo-European languages. However, that may be overcome through familiarizing students with the test format and making them aware of the merits of the AGLS.

There may be some degree of construct underrepresentation in this study because it has not yet been determined to what extent, if at all, the subject matter of the selected passage influences performance on the AGLS. This is because this testing method was only applied to one passage due to time constraints. It is important to note that this is a weakness in the study, not the AGLS itself. Also, there were some types of evidence not collected because of practicality issues, including population validity, predictive validity, and temporal validity.

**Authenticity.** As for authenticity, or what was historically known as ecological validity, it can be argued that no test could ever be completely authentic because it is a test and not a real world task, but tests are a part of everyday life, especially in academia. And with regard to the authenticity of AGLSs specifically, it's easy to imagine students missing a word or just not understanding and having to use all the knowledge at their disposal to make inferences and interpretations. Other efforts to build authenticity included allowing examinees to take notes,

controlling the rate of speech to match the average person's, and using a listening passage made by a native speaker for native speakers.

**Interactiveness.** Because the literature is quiet on the topic of interactiveness, and the originators of the idea, Bachman and Palmer (1996), are not specific as to how to measure it, it is difficult to judge the quality of interactiveness of the AGLS. As far as linguistic knowledge is concerned, examinees need a high level of collocational competence, a good vocabulary, and a second language working memory approaching that of their native language. All of this is in addition to what examinees need in order to comprehend any academic lecture. The involvement of topical knowledge is limited because the examinees cannot choose their topic and the construction of responses is tightly constrained. It was hoped that the topic of the passage would be interesting, so as to avoid the affective barrier of boredom, but interests of examinees vary so widely that it is not very practical to address all of them. The most that can be done is to find topics that could encourage interest from the largest number of people (which is not unlike what must be done in any college class). Also, some test-takers may have felt uncomfortable with the difficulty of the AGLS and their inability to fill in the blanks. All in all, the linguistic interactiveness of the AGLS is satisfactory, but the interaction with topical knowledge and affective schemata may leave something wanting. It would be possible to increase the interaction with topical knowledge by having a number of AGLSs to choose from, but that introduces reliability and practicality issues.

**Impact.** In the event that students actually study before taking the credit exam at BYU, the AGLS can only encourage them to listen to real, academic lectures and improve their vocabulary. The impact that failing this test would have on examinees is that they would not receive credit for the elective ESL classes offered at BYU without actually taking those classes.

This might mean that they would take the elective classes in order to get the credit anyway, which could only help them to improve their academic listening skills. It might also mean that they would study another language to fulfill that requirement for a Bachelor of Arts degree, or they could switch to a Bachelor of Science and take math instead. The stakes of this test are relatively small; only in extreme circumstances might it have a truly negative impact. The impact of passing this test would be receiving credit for ESL classes without actually taking them, thus fulfilling the requirements for a Bachelor of Arts, and possibility added confidence in their listening abilities as they go through their academic careers.

This test could impact BYU as an institution in a variety of ways, such as a change in enrollment for the elective ESL classes. It might also impact those ESL classes themselves by setting a standard for what it means to pass those classes, as the current exam does now.

**Practicality.** Compared to the effort involved in writing the traditional exam, the AGLS was very easy to write, and it used the same technology in the administration while taking less than a quarter of the time. Even though the AGLS requires a human rater, unlike the machine-rated traditional test, in order to avoid testing spelling (thereby introducing construct irrelevance), it is quick and easy to do so. The human rater aspect of the AGLS may mean it requires slightly more resources than the traditional test, but the added reliability, discriminative powers, and the sound theoretical basis of the test may mean that it is worth it.

### **Further Discussion**

It is true that the AGLS does not have a perfect correlation with traditional measures of listening, but the traditional exam and the AGLS may test different aspects of listening. The traditional exam was designed to assess vocabulary, vocabulary in context, listening for details, and passage comprehension, whereas the AGLS was designed to test collocational competence,

working memory, and phonological sequencing (though it may be argued that each test assesses some or all of what the other was designed to assess). All three of these elements are excellent indicators of linguistic abilities (see Ellis, 1996; Gass et al., 2006; Keshavarz & Salimi 2007); therefore, passing the AGLS could mean at least one of three things. First, the examinee demonstrated collocational competence, which correlates very strongly with exceptionally high proficiency because they have adequate the knowledge of and experience with the language as to be able to predict formulaic sequences. Second, if the students could not predict what should have gone in the blanks, they remembered what it was from the original passage. This is significant because the retention of unfamiliar material is an excellent predictor of language proficiency, which means that the plain act of remembering words from a passage heard a few minutes earlier indicates a “superior” language learner (Skehan, 1982). A person’s working memory is severely limited in their second language, and the closer their second language working memory is to their native language working memory, the more proficient the learner is. Lastly, if students could neither predict nor remember what belonged in the blanks, then they had to hold each phrase in their minds using the phonological loop while analyzing the phrases to formulate probable options based on their comprehension of the passage and choose the best alternative. The capacity of the phonological short-term memory indicates an aptitude for vocabulary learning, which is a significant barrier in academia when the average person’s vocabulary is estimated anywhere from 17,000 base words (Goulden, Nation, & Read, 1990) to more than 30,000 words (Crystal, 1987), and the number just increases with each new subject studied in college. Clearly, the 2,000 most frequent words (Nation & Waring, 1997) and an additional 570 academic words (Coxhead, 2000) are inadequate, so an aptitude for vocabulary learning is essential in academic studies. In the end, it does not matter whether students use one

or two or all three of the strategies listed to complete the task because each strategy reveals that they have a high level of proficiency and an aptitude for learning.

Anderson (1972) and Dunkel, et al. (1993) asserted that, in order for there to be any evidence of listening comprehension, the assessment task must require a transformation of the listened-to information at a deep, structural, and semantic level. The AGLS requires examinees to comprehend “mutilated” information (gapped passages), analyze the structure, and restore the passage in a semantically and syntactically correct way based on what they’ve heard. It is a reconstruction process that provides good evidence of understanding.

### **Recommendations**

While this method may not be a panacea for listening test limitations, it can certainly be used as a quick method to triangulate data about a student’s listening ability. For a low-stakes test like the BYU ESL credit exam, it is not unreasonable to think the AGLS could replace the listening portion with further validation.

Some may raise the concern that there have been so few attempts at similar assessments with so few people making use of them for a reason. Perhaps the few test writers who have tried assessments similar to the AGLS found that they were not up to their standards or that student misgivings toward it too great to merit further use. However, the lack of popularity of the AGLS should not be a strong call for alarm. Language education professionals seem to feel the need to reinvent the wheel with each new generation (Richards & Rodgers, 2001). It’s a reasonable impulse—there are so many ideas out there to be vetted. But this habit means that sometimes good ideas fall by the wayside. Ever since the Audio-Lingual Method, “rote memorization” has become a dirty word in education and few teachers ask their students to learn anything by heart even though it has been proven to be a very beneficial practice, such as in learning new

vocabulary (Ding, 2007). Reduced redundancy tests and their ilk are just among those other ideas that have grown a little dusty but are no less useful.

### **Limitations**

A limitation of the test is the inability to access the amount of typographical errors. For example, if a student answered “read,” it is not possible to know whether he or she meant [rid] or [rɛd]. Or if the student typed “or” where “of” would have been appropriate, it is impossible to know if it was a typographical error or a lack of knowledge of the English language. This limitation could be overcome by having students audio-record their responses. However, that would make the test more time-consuming to grade, and it also introduces complications regarding pronunciation.

The biggest limitation to the AGLS was its face validity. Most students did not take it seriously. Because the test was difficult and unlike anything they were used to, at least one student became angry due to his frustration. But this frustration would be reduced and face validity would pose less of a limitation if this technique were to become a common practice because students would have opportunities to learn the technique and get used to it. More practice was not provided during this project because of practicality issues. Providing practice would have required assembling the test subjects at least twice. It was difficult to find test subjects at all, and the attrition rate should they have been required to return for a second administration would likely have been formidable. However, as this study demonstrates, the technique seems to be valuable and could become a common practice, and as more students are exposed to it, face validity will become less of an issue. One might hesitate to employ it because of its difficulty, but as Brown (1995) states,



We can learn rather little about the processes of comprehension when they flow comfortably. . . . We have an opportunity of learning rather more where understanding is difficult to come by, where interpretation is only partially achieved, or where an attempt to communicate results in misunderstanding. (p. 42)

While the analysis of this project is inconclusive, it should not be dismissed out of hand. This assessment technique shows promise in terms of usefulness at BYU. Since this project was not a comprehensive validation of AGLS, there are certainly steps that could still be taken to further evaluate the AGLS. These steps will be discussed in the next section (Suggestions for Future Research).

### **Suggestions for Future Research**

To fully validate the AGLS, it is vital to collect evidence of criterion-rated validity, population validity, predictive validity, and temporal validity. TOEFL listening scores could provide more evidence for criterion-related validity. A larger sample size and broader range of topics and passages are critical to be able to generalize results to the greater population. Students' fields of study could give some advantageous background knowledge even considering that all students hear the same passage before listening to the summary, so different topical passages should be included in future versions of the test. Maybe the most important research to conduct in the same vein, however, is to see if this technique could be used with low-level learners, if the passages were level-appropriate in the context of placement or diagnostic testing.

Qualitative data regarding students' opinions of the AGLS and thought processes while taking the AGLS should yield interesting information. Furthermore, longitudinal data collection including testing students as they enter school, then tracking the progress of the students who pass and fail the AGLS and comparing GPAs between the two groups would supply support for

predictive and temporal validity. It would also be good to compare the ELLs' GPAs to native English speakers. Both of these actions would determine if there's a significant difference between the pass and fail groups of ELLs and native speakers. Triangulating the results with some qualitative data from exit surveys in which students estimate how well they understood their professors and textbooks could provide support for a degree of predictive validity. Unfortunately, because an MA takes two years and this proposition would take at least four, it was not possible to do it for this study.

Researchers could also experiment with different spacing of gaps in the passage. It would be very intriguing to see if gapping every eighth word always produces similar results. Based on the characteristics of speech, there is some likelihood that there is a configuration that would most often be appropriate. Buck (1997) described speech as clause-like units of about seven words, usually two seconds long with a single intonation curve, strung together with coordinating conjunctions.

Something that is not specifically part of Bachman and Palmer's usefulness criteria but does argue for the AGLSs usefulness is the knowledge gleaned from wrong answers. While incorrect responses were scored as zero no matter how close they were to the correct answer, many of the incorrect responses yielded interesting information. The words entered into the blanks were frequently common collocates of the word before or after the blank—they just were not appropriate for that particular context. For example, many students put *very* in the blank before *well*. Other wrong answers were words that were semantically appropriate while failing to fit into the sentence syntactically. Still others were just words from the passage, seemingly chosen at random. It would be interesting to analyze this in light of students' interlanguage. An in-depth

analysis of incorrect responses could reveal potential for the AGLS to be used as a diagnostic test (Yamauchi, 1990).

In order for any validation of the AGLS to be comprehensive, studies will have to be done comparing the AGLS to tests of working memory, collocational competence, and phonological sequencing. As of right now, it is not certain that the AGLS does in fact test these three things; it is conjecture with a sound theoretical basis. An analysis of examinee responses may reveal which of the three, if any, of these things are being assessed. Furthermore, if tests of collocational competence, working memory, and phonological sequencing were also given concurrently with the AGLS, simple correlations between these tests and the AGLS could be informative.

## **Conclusion**

I have tried to address the concerns that Vandergrift (1997, 2006, 2007, 2010) and others raise repeatedly about the validity of listening assessments and supply a solution. This effort involved a thorough exploration of academic listening, the constructs involved in listening comprehension, and of listening assessment. I also studied the concept of test usefulness (Bachman and Palmer, 1996). Based on my review of listening assessment, I created the AGLS to measure academic listening comprehension. The exam was then piloted through BYU's English Language Center and the credit exam for matriculated students. Finally, the results of the pilot were analyzed to determine whether future investigation was merited. Further investigation is merited.

While it remains debatable that the AGLS could perform the same function as BYU's credit exam for listening proficiency with equal or greater validity, it has clearly performed with higher reliability in much less time and with greater discrimination.

## REFERENCES

- Abraham, R. & Chapelle, C. (1992). The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, 76, 468-479.
- Alderson, J. (2000). *Assessing reading*. New York, NY: Press Syndicate of the University of Cambridge.
- Alderson, J., & Bachman, L. (2001). Series Editors' Preface. In G. Buck (Ed.), *Assessing listening* (pp. x-xi). New York, NY: Cambridge University Press.
- Alderson, J. & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, 35, 79-113.
- American Educational Research Association. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anderson, R. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research*, 42, 145-70.
- Angoff, W. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity*, (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum.
- Bachman, L. (1990). *Fundamental considerations in language testing*. New York, NY: Oxford University Press.
- Bachman, L., & Palmer, A. (1996). *Language testing in practice*. New York, NY: Oxford University Press.
- Baddeley, A. & Hitch, G. (1974). Working memory. In G. Bower (Ed.), *The psychology of learning and motivation*. New York, NY: Academic Press.
- Bahns, J. & Eldaw, M. (1993). Should we teach EFL students collocations? *System*, 21, 101-114.

- Birch, B. (2002). *English L2 reading: Getting to the bottom*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bodie, G., Worthington, D., Imhof, M., & Cooper, L. (2008). What would a unified field of listening look like? A proposal linking past perspectives and future endeavors. *The International Journal of Listening*, 22, 103-122.
- Breen, M. (1985). Authenticity in the language classroom. *Applied Linguistics*, 6, 60-70.
- Brennan, R. (1992). *Elements of generalizability theory*. Iowa City, IA: The American College Testing Program.
- Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, 18, 171-191.
- Brown, G. (1995). *Listening to spoken English*. London, UK: Longman.
- Brown, H. (2003). *Language assessment: Principles and classroom practices*. New York, NY: Pearson ESL.
- Brown, J. (2002). Do cloze tests work? Or, is it just an illusion? *Second Language Studies*, 21, 79-125.
- Brown, J., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32, 653-675.
- Brown, J. & Hudson, T. (2002). *Criterion-referenced language testing*. New York, NY: Cambridge University Press.
- Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal*, 10, 15-39.
- Buck, G. (1992). Listening comprehension: Construct validity and trait characteristics. *Language Learning*, 42, 313-357.

- Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing*, 11, 145-170.
- Buck, G. (1997). The testing of listening in a second language. *Encyclopedia of Language and Education*, 7, 65-74.
- Buck, G. (2001). *Assessing listening*. New York, NY: Cambridge University Press.
- Call, M. (1985). Auditory short-term memory, listening comprehension, and the input hypothesis. *TESOL Quarterly*, 19, 765-781.
- Chan, C. (2001). Phonological processing in reading Chinese among normally achieving and poor readers. *Journal of Experimental Child Psychology*, 80, 23-43.
- Chapelle, C. & Abraham, R. (1990). Cloze method: What difference does it make? *Language Testing*, 7, 121-146.
- Cobb, T. (Programmer). (2009). *Nth-word / rational deletion cloze builder*. [Web]. Retrieved from [http://www.lex tutor.ca/cloze/n\\_word/](http://www.lex tutor.ca/cloze/n_word/).
- Coombe, C., Folse, K., & Hubley, N. (2007). *A practical guide to assessing English language learners*. Ann Arbor, MI: The University of Michigan Press.
- Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34, 213-238.
- Crystal, D. (1987). How many words? *English Today*, 12, 11-14.
- Cumming, A. (Ed.). (1996). *Validation in language testing*. Bristol, PA: Multilingual Matters LTD.
- Dickens, M. & Williams, F. (1964). An experimental application of 'cloze' procedure and attitude measures to listening comprehension. *Speech Monograph*, 31, 103-108.
- Ding, Y. (2007). Text memorization and imitation: The practices of successful Chinese learners of English. *System*, 35, 271-280.

- Dunkel, P., Henning, G., & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *The Modern Language Journal*, 77, 180-191.
- Ellis, N. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, 18, 91-126.
- Ferguson, G. (1949). *On the theory of test discrimination*. Springer, NY: Psychometrika.
- Flowerdew, J. (Ed.). (1994). *Academic listening*. New York, NY: Press Syndicate of the University of Cambridge.
- Flowerdew, J. & Miller, L. (2010). Listening in a second language. In A. Wolvin (Ed.), *Listening and human communication in the 21st century*, (pp. 158-177). Malden, MA: Blackwell Publishing.
- Fontenelle, T. (1994). What on earth are collocations: An assessment of the ways in which certain words co-occur and other do not. *English Today*, 10, 42-48.
- Gardner, D., & Witherell, S. (2006, November 13). *New enrollment of foreign students in the U.S. climbs in 2005/06*. Retrieved from <http://opendoors.iienetwork.org/?p=89251>.
- Gass, S., Roots, R., & Lee, J. (2006, September 15). *Inhibition and working memory capacity in a second language*. Paper presented at the 16<sup>th</sup> European second language association conference, Antalya, Turkey.
- Gass, S. & Selinker, L. (2008). *Second language acquisition: An introductory course*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Glicksberg, D. (1963). *A study of the span of immediate memory among adult students of English as a foreign language*. Unpublished dissertation. University of Michigan, Ann Arbor, MI.

- Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, 11, 341-363.
- Grefen, D. & Straub, D. (2005). A practical guide to factorial validity using PLS-graph: Tutorial and annotated example. *Communications of the Association for Information Systems*, 16, 91-109.
- Gregory-Panopoulos, J. (1966). *An experimental application of "cloze" procedure as a diagnostic test of listening comprehension among foreign students*. Unpublished dissertation. University of Southern California, Los Angeles, CA.
- Herbst, T. (1996). What are collocations: Sandy beaches or false teeth? *English Studies*, 77, 379-393.
- Hughes, A. (2003). *Testing for language teachers*. New York, NY: Cambridge University Press.
- Imhof, M. (2004). *Zuhören und Instruction - Empirische Zugänge zur Verarbeitung mündlich vermittelter Information*. Münster: Waxmann.
- Imhof, M. (2010). What is going on in the mind of the listener? The cognitive psychology of listening. In A. Wolvin (Ed.), *Listening and Human Communication in the 21st Century* (pp. 97-126). Malden, MA: Wiley-Blackwell.
- Kehoe, J. (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, 4(10). Retrieved from <http://PAREonline.net/getvn.asp?v=4&n=10>.
- Keshavarz, M. H., & Salimi, H. (2007). Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics*, 17, 81-92.
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. New York, NY: Cambridge University Press.



- Kunnan, A. J. (Ed.). (1991). *Studies in language testing: Fairness and validation in language assessment*. New York, NY: Press Syndicate of the University of Cambridge.
- Lennon, P. (1996). Getting 'easy' verbs wrong at the advanced level. *IRAL*, 34, 23-36.
- Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory into Practice*. Hove, UK: Language Teaching Publications.
- Lewkowicz, J. (1991). Testing listening comprehension: A new approach? *Hong Kong Papers in Linguistics and Language Teaching*, 14, 25-31.
- Linn, R., Baker, E. & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20, 5-21.
- Loe, M. (1964). *Immediate memory-span in English and Chinese sentences of increasing length*. Unpublished thesis. Georgetown University, Washington, D.C.
- Lund, R. (1991). A comparison of second language listening and reading comprehension. *The Modern Language Journal*, 75, 196-204.
- Lynch, T. (1994). Training lecturers for international audiences. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 269-289). New York, NY: Cambridge University Press.
- Lynch, T. (1998). Theoretical perspectives on listening. *Annual Review of Applied Linguistics*, 18, 3-19.
- Mackey, A. & Gass, S. (2000). *Second language research: Methodology and design*. Mahwah, NJ: Lawrence Erlbaum and Associates.
- Mazzoni, D. (Programmer). (2009). Audacity. [Computer Software]. Retrieved from <http://audacity.sourceforge.net/>.

- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18, 5-11.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, 23, 13-23.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13, 241-56.
- Moon, R. (1992). Textual aspects of fixed expressions in learners' dictionaries. In P. Arnaud & H. Bejoint (Eds.), *Vocabulary and applied linguistics* (pp. 13-27). London: Macmillan.
- Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62, 229-258.
- Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmidt & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy* (pp. 6-19). New York, NY: Cambridge University Press.
- Nichols, R. (1948). Factors in listening comprehension. *Communication Monographs*, 15, 154-163.
- O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, 29, 557-582.
- Ohata, K. (2006). Auditory short-term memory in L2 listening comprehension processes. *Journal of Language and Learning*, 5, 21-28.
- Oller, J. (1983). Evidence for a general language proficiency factor: An expectancy grammar. *Issues in Language Testing Research* (pp. 3-23). Rowley, MA: Newbury House.

- Pollitt, A., & Ahmed, A. (2009). *The Importance of being valid*. Proceedings of the Association for educational assessment-europe (pp. 257). Attard, Malta: <http://www.aea-europe.net/userfiles/p24%20Pollitt%20&%20Ahmed%20paper%20%282009%29.pdf>.
- Putrevu, S. (2000). Exploring the origins and information processing differences between men and women: implications for advertisers. *Academy of Marketing Science Review*, 10, 1-14.
- Richards, J. (1983). Listening comprehension: approach, design, procedure. *TESOL Quarterly*, 17: 219-239.
- Richards, J. & Rodgers, T. (2001). *Approaches and methods in language teaching*. New York, NY: Cambridge University Press.
- Richardson, U., Thomson, J., Scott, S., & Goswami, U. (2004). Auditory Processing Skills and Phonological Representation in Dyslexic Children. *Dyslexia*, 10, 215-233.
- Rivers, W., & Temperley, M. (1978). *A practical guide to the teaching of English as a second or foreign language*. New York, NY: Oxford University Press.
- Rubin, J. (1994). A Review of second language listening comprehension research. *The Modern Language Journal*, 78, 199-221.
- Sheskin, D. (2000). *Handbook of parametric and nonparametric statistical procedures*. London, UK: Chapman & Hall.
- Skehan, P. (1982). *Memory and motivation in language aptitude testing*. Unpublished doctoral dissertation. University of London, UK.
- Speciale, G., Ellis, N. & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25, 293-321.

- Tafaghodtari, M. & Vandergrift, L. (2008). Second and foreign language listening: Unraveling the construct. *Perceptual and Motor Skills, 107*, 99-113.
- Taylor, W. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly, 30*, 414-433.
- Templeton, H. (1977). A new technique for measuring listening comprehension. *ELT Journal, 31*, 292-299.
- Vandergrift, L. (1997). The Cinderella of communication strategies: reception strategies in interactive listening. *The Modern Language Journal, 81*, 494-505.
- Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency *The Modern Language Journal, 90*, 6-18.
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40*, 191-210.
- Vandergrift, L. (2010). Researching listening. In B. Paltridge & A. Phatik (Eds.), *Continuum companion to research methods in applied linguistics*, (pp. 160-173). New York, NY: Continuum.
- Wagner, R. & Torgesen, J. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin, 101*, 192-212.
- Wennerstrom, A. (1998). Intonation as cohesion in academic discourse. *Studies in Second Language Acquisition, 20*, 1-25.
- Williams, J. (1998). Guidelines for the use of multimedia in instruction. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, 1447-1451. Chicago, IL.
- Wray, A. (2000). Formulaic sequences in second language teaching: principles and practice. *Applied Linguistics, 21*, 463-489.

Yamauchi, S. (1990). An experiment with cloze procedure on Japanese EFL learners: On the diagnostic power of cloze procedure. *University of the Ryukyus Repository*, 35, 1-25.

Yi'an, W. (1998). What do tests of listening comprehension test? – A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, 15, 21-44.

## APPENDIX A

### Transcript of Lecture

Well today we're going to talk about something that you might feel a bit odd about doing until you get used to it. No, don't worry, it's not too weird. We're not talking about standing on your head while naked and singing show tunes or something like that. No, we're actually just talking about reading out loud. Yes, reading out loud. The thing you used to dread when called upon in grade school, and perhaps wondered why the teacher was making you do this. When, come on, I mean, why couldn't everyone just read on their own, silently? Well, although the body of research didn't exist back then, like it does now, your teachers were actually doing your brain a huge favor, beyond just testing your reading ability in general. Believe it or not, the act of reading out loud is one of the most challenging, taxing, and mentally stimulating activities that your brain can do. Period. Yes, even compared to solving those complex math or logic problems. Reading out loud, and being able to continue reading through paragraphs and pages consistently, without much stopping or pausing, and articulating clearly, and doing all the things that great orators and speakers who speak by reading unfamiliar, meaning nonmemorized, texts can do happens to kick those major areas of your brain into simultaneous action, which in turn gives your brain an excellent cross-training workout. And this makes sense if you think about it. To read pages from a book or magazine article out loud, and well, this involves your eyes following along and absorbing the words in advance before you speak them, while you're speaking what you've already read, because you're forced to scan slightly ahead. And then your brain has to process this visual input into speech output, which then kicks several motor areas into action, while processing proper inflection, analysis of how effectively your already spoken word and

inflection sounded, after you read them, and feeding back that input into upcoming speech to improve or correct tone and speed and inflection, all while you scan ahead to process upcoming words and process upcoming needed inflection and tonal changes, all while you're gauging reaction by any listeners, including yourself, all while you're trying to digest and comprehend and store the actual information that you're reading. This all happens very fast and in very real time, so really it makes perfect sense that this seemingly simple activity is actually one heck of a brain workout that utilizes numerous different areas of your brain abilities, and the research backs this up. The now-famous, brain-age professor, Dr. Ryuta Kawashima and his team at Tohoko University in Japan found that, just like doing those simple math equations quickly which we've covered in our other braincasts, reading out loud is one of the best possible mental workouts for the super-important frontal lobe of your brain. Especially as we get older.

## APPENDIX B

### Transcript of Summary with Gaps

The act of reading out loud is one of the most challenging and mentally stimulating activities that your brain can do, even compared to solving complex math or logic problems. / Reading out loud, and being able to continue [1] \_\_\_\_\_ through paragraphs and pages, / without much stopping [2] \_\_\_\_\_ pausing, / and speaking clearly, while reading unfamiliar, [3] \_\_\_\_\_ texts / happens to kick several major areas [4] \_\_\_\_\_ your brain into action. / To read pages [5] \_\_\_\_\_ a book or magazine article out loud, [6] \_\_\_\_\_ well, / this involves your eyes following along [7] \_\_\_\_\_ absorbing the words in advance before you [8] \_\_\_\_\_ them, / while you're speaking what you've already [9] \_\_\_\_\_ because you're forced to scan slightly ahead. / [10] \_\_\_\_\_ brain has to process this visual input [11] \_\_\_\_\_ speech output, / which then kicks several areas [12] \_\_\_\_\_ your brain into action, / while figuring out [13] \_\_\_\_\_ inflection and intonation. / You also have to [14] \_\_\_\_\_ how effectively your already spoken words sounded [15] \_\_\_\_\_ you said them, / and use that input [16] \_\_\_\_\_ improve upcoming speech or correct tone, speed, [17] \_\_\_\_\_ inflection, / all while you scan ahead to [18] \_\_\_\_\_ upcoming words and process upcoming needed inflection [19] \_\_\_\_\_ tonal changes. / You're also gauging the reaction [20] \_\_\_\_\_ any listeners, including yourself, / all while you're [21] \_\_\_\_\_ to understand the information that you're reading.



## APPENDIX C

### Table of Acceptable Answers

#	Answer	Acceptable Alternatives	Unacceptable Alternatives
1	reading	speaking, to read, scanning	read
2	or	and, nor, thinking	
3	unmemorized	written, new, un memorized, or new, lines of, foreign, books, or unpracticed, new, meaning nonmemorized, passages of, reading, complex, long, words and	texts, words, word, and reading, contents, read along, memories, passages, meterial, with, and, that, to you, books
4	of	in	areas, motor, mentally, practice, requires, from, process, activate, habiliy, important, of the brain, take, chllenging, or, while, leads, for of the
5	from	of, in	or, comprenhencion, book, help, on, of a, of books, aloud, carefully, quickly, is, ot
6	and	really, rather, very, pretty, to speak clearly	out, works, do, it, doing, challaging, excises, very well, evaluate, stemulate, as read, enough, is, helps, understand, better
7	and	scanning, while, reading, the text, scanning slightly ahead, the lines, scaning, then, slightly, the words, skimming	words, to, brain, of, article, into, time, mouth, by, lines, texts, ligne, with, scan, eyes
8	speak	read, say, repeat, speak, pronounce	understand, tell, reading, catch, unterdand, finish, speaking
9	read	seen	reading, concentrate, know, looking, and scan, memorize, konow, well, done, reading, correct, undertand, organized, to scan, said
10	your	the, our	because, hangend, improve, several, so, throughout, while
11	into	and, as, while at the same time doing the, to, with	before, the, into speak, flexible, information, brings
12	of	in	and put, that, make, to, up, motivate, of the brain, of memories, activate, work out, start, and, chanllenge, from, improve, or, explore, keep
13	proper	the, its, tonal, your, correct, both, reaction, pronunciation, speech, an, action, changes in, the right	what, which, [,] , read, how, for, out, of, reading aloud, with, in, it, what is

#	Answer	Acceptable Alternatives	Unacceptable Alternatives
14	analyze	judge, remember, note, think, evaluate, gauge, examine, process, check, digest, hear, measure, figure, understand, figure out, think about	know, comprehend, to, memorize, make, repeat, spoke, speak, the, listen, learn, read, comprehend
15	after	when, as, how, while, the way, and how, after	before, and, are, seldom, laud, listen, understand, digenst, to, frequent, also, correct, that, clear, perfect
16	to		into, skill, reading, conversations, speech, correct, brain, this, putting
17	and	intonation, or, pronunciation, tonality, reaction, diction	of, better, quickly, several, fast, brain, for, to, or and, the, in
18	process	read, preview, the, see, find, comprehend, understand, digest, take, search, improve, new, process, all, unknown	look, memory, listen, know, predict, improve
19	and	or, in, to predict, to, with, and hear, or, to plan your, to predict	digest, all the, processus, internal, will, braind, made, store, speech, of, read, understand, the word, your, make
20	by	of, from	or, to, read, speak, repeat, and, voice, listen, search, know, pronounce, [to; with], [read/the], a new, digest, the, processes, see, hear, to, all the, unknown
21	trying	working, trying to, processing, brain tries	thinking, processing, think, did, ability, able, do, brain, to scan, braind, mind, habilities, digest, speaking, scan, comprehension, help, haves, know, brain, are, ability, knowledge, speaking, able to