Assessing Students’ Listening Comprehension of Different University Spoken Registers

Tingting Kang

Applied Linguistics Program

Northern Arizona University

Abstract

The awareness of varied registers has given rise to the general field of English for Academic Purposes (EAP), which focuses on teaching language that is used in university registers. This paper aimed to design an achievement test for formative purposes to measure advanced language learners' (*N* = 25) mastery in different university spoken registers through listening. The intended consequences of the test are to provide stakeholders the opportunity to evaluate instruction and learning processes, examine the course objectives, reinforce administration goals, and in the long run, make the program administration more effective. The test results showed that most of the students passed the test. Following Bachman and Palmer's (2010c) four claims in Assessment Use Argument (AUA), various evidence has been provided to support the validity argument of this test: (a) consequences were beneficial; (b) specific decisions made on basis of scores; (c) results could be interpreted as indicator of construct; (d) records were consistent.

Keywords: EAP, listening comprehension, university register

Assessing Students' Listening Comprehension of Different University Registers

Background

To ensure students' academic success, it is not enough that they have vocabulary and grammatical knowledge; they should also be equipped with the knowledge of register differences (Biber & Conrad, 2009). Similarly, the awareness of varied registers has given rise to the general field of English for Academic Purposes (EAP).

The purpose of this test was to measure advanced language learners' listening comprehension of different university spoken registers and further help stakeholders evaluate instruction and learning processes, examine the course objectives, reinforce administration goals, and make the program administration more effective.

Target language use (TLU) task was defined as "a specific language use tasks which test takers are likely to perform in specific setting" (p. 60), which fell inside the TLU domain (Bachman & Palmer, 2010). In terms of the spoken language in the university settings, it includes various university spoken registers, such as university office hours, lectures, class discussion, study group, student presentation, and service encounters (Biber, Conrad, Reppen, Byrd, Helt, Clark, Cortes, Csomay, & Urzua, 2004). The four TLU tasks that served as basis for my test purpose were university office hours, lectures, university encounter services, and university student conversations.

According to Buck (2001), listening comprehension involves two types of knowledge: linguistic knowledge and non-linguistic knowledge. In this listening comprehension test, the three subconstructs of listening comprehension included two aspects of linguistic knowledge (i.e., vocabulary and cognitive understanding) and one aspect of non-linguistic knowledge (i.e., situational context) were examined.

Hypotheses

My first hypothesis was that main idea and detail questions would be easier than the situational context and vocabulary questions, and the hardest questions should be about vocabulary. Second, since the three subconstructs were hypothesized to measure the same construct, listening comprehension of different university registers, their internal consistency should be very high. Third, students' scores on this test should match with their listening comprehension abilities in classroom performance.

Methods

The participants were PIE Level 5a (*n* = 14) and 6 (*n* = 11) students, enrolled in Fall 2013 Listening and Speaking courses. The listening comprehension test was administered by the test developer during participants Listening and Speaking class for about 30 minutes.

The Table of Specification (Miller, Linn, & Gronlund, 2009c) for this test is in Appendix A. It is a two way chart that relates the subconstructs (i.e., situational context, cognitive understand, and vocabulary) and the test tasks (i.e., university office hours (listening 1), lectures (listening 2), service encounters (listening 3), and student conversations (listening 4)). Each task was worth 25% of the overall score. There were 10 questions per listening passage, which included (a) three situational context questions (i.e., participants, setting, and topic) (b) four vocabulary questions, and (c) three cognitive questions (i.e., one main idea question and two detail questions). The subconstructs accounted for 30% (situational context), 30% (cognitive understanding), and 40% (vocabulary) of the total score. The questions assigned for each subconstruct were participants (Question 3, 10, 17, and 24), setting (Question 4, 11, 18, and 25), topic (Question 6, 13, 20, and 27), main idea (Question 5, 12, 19, and 26), detail (Question 1, 2, 8, 9, 15, 16, 22, and 23), and vocabulary (Question 7, 14, 21, and 28). Each question was worth one point.

Results

In terms of the descriptive statistics, the average scores were 33.56 (*SD* = 5.49, *K* = 40, total), 10.88 (*SD* = 1.30, *K* = 12, situational context), 10.28 (*SD* = 1.10, *K* = 12, cognitive understanding), and 12.40 (*SD* = 5.24, *K* = 16, vocabulary). Regarding reliability (i.e., consistency of measurement (Miller, Linn, & Gronlund, 2009b)), the results of Cronbach's Alpha for item internal consistency were .86 (total), .46 (situational context), -.06 (cognitive understanding), and .96 (vocabulary). Since this was a criterion-referenced test, the results of agreement coefficient, which measured the consistency of mastery and non-mastery classifications, were between .96 and .86 and exceed the minimum requirement, .75 (Subkoviak, 1988). The scores of SEM, the margin of errors in test scores (Test Service Bulletin No. 50, 1956), ranged from 2.02 to .93.

The cutoff percentage for mastery in this test was 70%. Therefore, the cut score was 28 for the overall test, 8.4 for situational context and cognitive understanding, and 11.2 for vocabulary. Most of the students passed the test and each subconstruct. Also, students who failed the test and each subconstruct were mostly from Level 5a and very few from Level 6.

Discussion

After reviewing the item statistics, items with low *P* and *D* values were indicated as problematic items because they were hard items and failed to discriminate higher and lower proficiency students(e.g., Question 2, 4, 17, and 26). They had characteristics of being too hard, which led to guessing, or including highly possible distracters. Accordingly, revisions could be made to change some of the distracters. For example, in Question 26, the mostly selected distracter, c, could be “They want to drink coffee together.”

Even though “the interpretations and decisions that are made on the basis of assessment results can never be considered justified in any absolute sense” (Bachman & Palmer, 2010c, p. 95), the evidence provided for the validity argument in the previous section were convincing. Some evidence such as theoretical cutoff percentages, teachers’ feedback, and SEM scores on TOEFL iBT made the validity argument more comprehensible.

Overall, this test achieved its purpose, measuring advanced language learners’ listening comprehension of different university spoken registers. Also, the results confirmed some of the hypotheses. First, vocabulary questions were the hardest subconstructs, and it contained the highest number of non-masters (*n* = 8). Second, the overall internal consistency of this test was also very high (Cronbach’s Alpha = .86). Lastly, according to teachers’ feedback on this test, students’ scores matched with their listening comprehension abilities in the classroom performance. However, owning to the low internal consistency for cognitive understanding, future teachers and test developers should consider deleting this subconstruct to measure students’ listening comprehension of university registers.

References

Bachman, L., & Palmer, A. (2010). *Language assessment in practice.* New York, NY: Oxford University Press.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge, UK: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus. *ETS TOEFL Monograph Series, MS-25.* Princeton, NJ: Education Testing Service.

Buck, G. (2001). *Assessing listening.* New York, NY: Cambridge University Press.

Miller, M. D., Linn, R., & Gronlund, N. (2009). *Measurement and evaluation in teaching* (10th Edition). New York, NY: Pearson.

Test Service Bulletin No. 50. (1956). *How accurate is a test score?* NY: The Psychological Corporation.

Appendix B
Tables of Specifications

	Constructs							
	Situational Context			Cognitive Understanding		Vocabulary		
University Spoken Registers	Participants	Setting	Topic	Main idea	Detail		# of questions	%
Listening 1: Office hours Length: 3’25” Words per min: 181	3 [1]	4 [1]	6 [1]	5 [1]	1, 2 [2]	7-a, b, c, d [4]	10	25
Listening 2: Lectures Length: 3’01” Words per min: 137	10 [1]	11 [1]	13 [1]	12 [1]	8, 9 [2]	14-a, b, c, d [4]	10	25
Listening 3: Service encounters Length: 2’33” Words per min: 162	17 [1]	18 [1]	20 [1]	19 [1]	15, 16 [2]	21-a, b, c, d [4]	10	25
Listening 4: Student conversations Length: 2’57” Words per min: 193	24 [1]	25 [1]	27 [1]	26 [1]	22, 23 [2]	28-a, b, c, d [4]	10	25
# of questions	4	4	4	4	8	16	Total questions: 40	
Points per question	1	1	1	1	1	1		
Points per sub construct	12			12		16	Total points: 40	
% per subconstruct	30			30		40		100