Matthew Barbee
SLS 490 Testing in a Language Program
University of Hawaii at Manoa
22 March 2013

Assessment Article Commentary:
Lee, H., & Winke P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing, 30*(1), 99-123.

HyeSun Lee and Paula Winke (2013), in their article, *The Differences among Three-, Four-, and Five-option-item Formats in the Context of a High-stakes English-language Listening Test,* which appeared in *Language Testing* just a couple of months ago present an empirical study that investigates the claims that the number of multiple-choice options, three-, four, or five-, matters in L2 testing contexts. Specifically, the authors are concerned with the titular context—a high-stakes English language listening test given to Korean high school students.

The article's abstract is approximately 206 words long and made up of 8 sentences. It is an example of what Swales (2004) calls a *summary abstract.* Divided into five moves, the abstract roughly follows the organization of the article itself. Each move of the abstract uniquely functions to (a) provide introductory information on the topic and context of the study, (b) layout the methodology of the study, (c) present the results of the study, (d) present a summary of possible interpretations of those results, and (e) to provide a general implication to all test developers when determining the optimal number of options on multiple-choice language tests. In short, the abstract is clear, concise, and gives anyone skimming the *Language Testing* journal a good idea of what they will find in the article.

For the remainder of this assessment article commentary, we will take the assignment questions as headings and address them directly.

Is the review of the literature sufficient?

Did the framework provide and adequate framework for viewing the study?

Swales (2004) describes three possible moves that usually occur within the introductions to research articles. He said that they should (a) establish a territory, (b) establish a niche within that territory, and (c) occupy that niche. Lee and Winke (2013) use their literature review to strategically do just that. First, they make broad sweeps about the popular use of multiple-choice tests in L2 language testing. They then focus in on literature that questions how many options are optimal and critique the few studies written on the subject. They finally point out errors or holes in that existing literature. Overall, the authors do a nice job illustrating the field and how they are uniquely presenting new information.

Is the study or discussion based on any particular theory?

What research theories are assumed?

No particular research theories are assumed in this study. It does however look at several theories concerning the number of options on multiple choice language tests and how those choices may affect student outcomes. The authors systematically present each of those theories and dispute all of them or claim that further research must be done on topic. This article is their answer to that claim.

What are the research questions?

Are there stated hypotheses?

The research questions are as follows:

1. Does L2-listening-test performance vary depending on whether the test has three-, four-, or five-option multiple-choice items? Specifically, does the number of options affect the average level of test difficulty and item discrimination indices?

2. Do L2 listening tests with three-, four-, and five-option items consistently measure test takers' listening skills? Additionally, do L2 listening tests with three different numbers of options assess the same skills?
3. Does the time it takes test takers to respond to multiple-choice items differ in relation to the number of options the items have?
4. What are test takers' opinions regarding three-, four-, and five-option-item formats on the high-stakes, CSAT English listening test?

The authors do not hypothesize on the outcome of their study.

What methodology is used?

This study was based on 264 Korean private school 10th graders from Seoul, who had been learning English in an EFL classroom setting for the past 7 years. The materials chosen to conduct research were three original versions of the CSAT English listening test containing 17 5-option items. Lee and Winke (2013) used the option deletion method, deleting the least attractive option, to develop the adapted 4-option and 3-option versions of the listening test. In total, 9 tests were used (the three original 5-option versions, three adapted 4-option, and three adapted 3-option tests) to answer the research questions. At the end of this quantitative research, a survey asking the test-takers preference on test format and options per item was administered to draw qualitative analysis.

What types of data analysis are used? Are the procedures clear?

How does the study address reliability (consistency) and validity?

The 264 Korean students were randomly divided into three groups. For the purposes of the study, several materials were developed. Each group was administered a version of a 3-option, 4-option, and 5-option test each. These tests were administered in three sessions a week

apart, followed by one survey session. The 17 listening items were available for the students to view prior to listening to the audio files. The timeframe for each test was 20 minutes, with 15-second pauses between items. The procedures are easily understandable and clear.

In addressing how the study checks reliability, we have created the following chart to illustrate how checks for reliability were conducted during each part of the procedure.

Procedure	Reliability Check
1. CSAT tests used	Equivalent-forms reliability
2. Sequence of testing	Test-retest reliability
3. Test reliability	Cronbach's alpha coefficient
4. Survey	Interrater reliability check

If there is a "Results" section, does it address the full study?

How clear are the findings?

In the results section of the article, the authors address answers to all four of their research questions while providing detailed explanation as to why they employed certain approaches to their research. Lee and Wilke (2013) were able to conclude that the number of options in a listening test did in fact affect the mean score, however it did not seem to affect item discrimination. In the end, it was concluded that a three-option test would be optimal. That said, their results were not so cut-and-dry. For clarity, they turned to qualitative measure to balance their results. The researchers also found that "despite the psychological advantages of a three-option format, the developers of the CSAT may be reluctant to adopt it" (Lee & Wilke, 2013, p. 119). Overall, they showed that "There may *not* be a universal, easy answer to the optimal number of multiple-choice options. Test developers and researchers have to consider the

statistical, affective, and contextual factors when determining the optimal number of options for their test" (p. 119).

Are there suggestions for further research?

Somewhat disappointing and later discussed a little further in this paper, the authors only make one actual suggestion for further research. As an extension of their self-prescribed limitation that their study has a relatively small sample size, they plead for more studies with more items or longer tests.

It should be noted that while the article does have a section labeled, "Need for Further Research," the actual purpose of that section was to clear the way, for this study to occupy the niche that this section pointed out.

Are the findings new, important, useful?

As stated by the authors, "the question might not be how many options are optimal universally, but rather how many options are optimal in a given testing context" (Lee & Wilke, 2013, p. 103).

It is made clear in their literature review that this is a new question that they are asking. In addition, Lee and Winke (2013) are also interested in test purposes, score uses, and qualitative information, which they say should enlighten test-makers choosing the number of items they do. This along with presenting the opinions and preferences of test takers in regards to the number of options on tests is presented seems to us as new and useful information that could result in the better tests.

What are the limitations (an potential flaws) of the study/discussion?

How appropriate are the generalizations or recommendations?

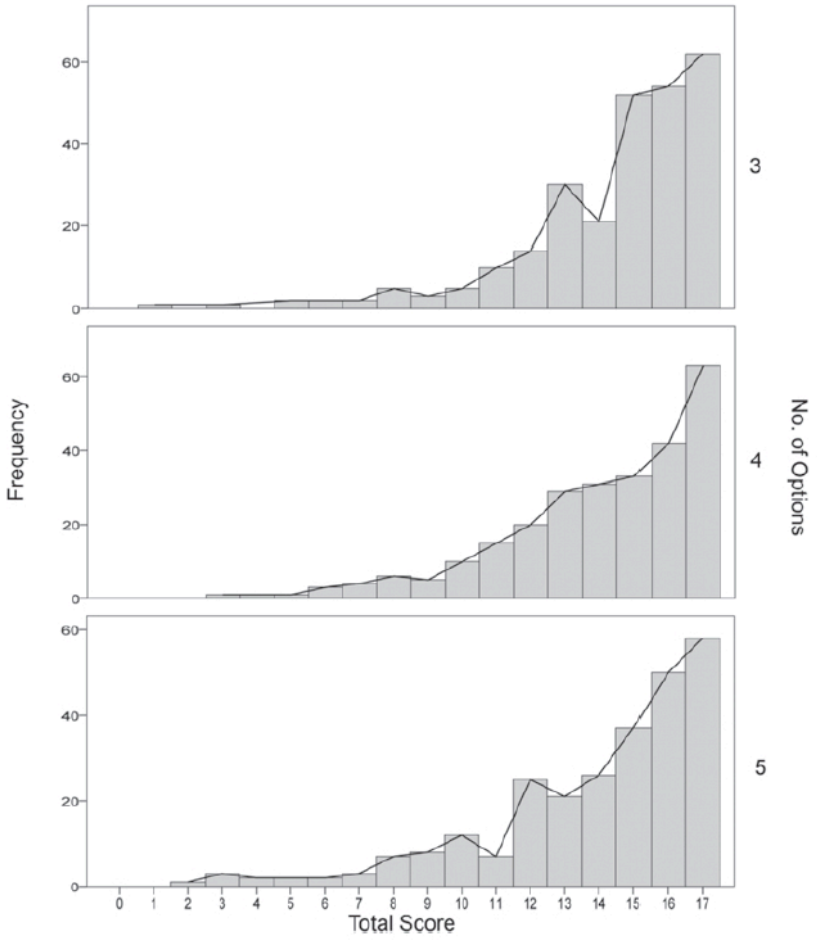
Are the results over-interpreted?

One concern that we had involved the creation of materials, i.e., how the researchers went about eliminating options to create the 4-option and 3-option versions of the test seemed problematic. Perhaps random elimination of distractors would have fit the reality of how tests are created. In making tests, test-makers neither have the luxury to prescreen a test and statistically, using distractor analysis, remove the least frequently chosen distractors to eliminate, which was also claimed by these researchers, nor can they afford to be so meticulous in getting a 28-member board of expert test-makers and experienced test-takers to subjectively eliminate their least plausible distractors, the method of elimination chosen by these researchers. Perhaps a method that better replicates the reality of test-making could have been used to eliminate options. One possibility would be to use random deletion, i.e., one of the four distractors from each item of the original 5-option test could have been randomly chosen for deletion, perhaps applying a reduced Latin Square technique. This would ensure that neither all the least plausible distractors, making for a harder test, nor the most plausible distractors, making for an easier test, were eliminated in the case of each item.

Another problem dealt with the authors' third research question, "Does the time it takes test takers to respond to multiple-choice items differ in relation to the number of options the items have?" To answer this question, only a sampling of the 264 participants was used (4.5% of the students who took the 3-option test, 4.9% of the 4-option test takers, and 4.5% of the 5-option test takers). These participants timed how long it took them to answer each item. While the results showed that the 3-option test's items required less time, it is possible that this result

isn't generalizable across all participants. Perhaps it was a limitation of resources, but why not allow all participants to time themselves or create a testing system that times students' responses for them. Problems can be foreseen no matter which of these methods is chosen.

Further, the skewed results of the three versions of the listening test also seemed problematic. The histograms for the three versions of the test, as provided by Lee and Winke (2013) are as follows:



While Lee and Winke (2013) claim that "standard deviation, skewness, and kurtosis values did not indicate that the distributions of test scores were different from one another in each group," those distributions were however negatively skewed, and greatly so. In dealing with this issue, the authors' used the Alpha Test, which is utilized in the UX_1 statistic. While we don't claim to

understand the methods used here, it did make us wonder about the researchers' overall methodology. That is, it becomes obvious that while negatively skewed, the listening test was never intended to be treated differently than or isolated from the whole of the College Scholastic Ability Test (CSAT), which consists overall of several sections, one being a 70-minute English language test. The listening test makes up only 20 minutes of this one section of the overall test. In isolation, we argue that it is not the intention of the test-makers to discriminate students' English ability using only the listening test. This negative skewness may find greater normal distribution when analyzed with the entirety of the English portion of the CSAT. That information, however, is beyond the scope of Lee and Winke's (2013) article.

This also raises the question of test difficulty. For the population chosen to participate in this study, 264 Korean high school students from one school, the means on all three of the original 5-option listening tests were high (see the following table).

Test version	*k*	*M*	*SD*
I5	17	12.65	3.21
II5	17	14.31	3.04
III5	17	14.71	3.07

Coupled with high means, the tests also showed low mean item facility as reported by the authors. For this to be a high-stakes, norm-referenced test for all Korean high school students, this appears problematic. Perhaps the test in general, regardless of how many options there were per item, was too easy for this small population sampling. While this study's small sampling size is pointed out in the authors' section on limitations to the study, the inability to generalize

across the entire population of Korean high school students and its effects on skewness in this study were overlooked.

Lastly, we were concerned with another area of generalization where the authors may have over reached. This generalization concerns the fact that the study only addresses one of the four skills: listening. It could possibly be argued that listening tests are unique in that it is impossible for students to have referential access to the script or text while answering the questions about the said text. In all cases, listening tests, more than tests of any other skill rely on memory to answer questions. This could be problematic when considering time and L2 reading speed, e.g., longer it takes a student to read the options of a multiple choice test, the longer they will have to retain memory of the listening text. Not only was this not addressed, but also we feel that this prevents generalization across the other three skills, reading, speaking, and writing.

Overall, while we scoured the article for other problems or potential flaws, we were met with impeccable design and scrutinizingly meticulous methodology. That, coupled with the authors' literature review which found and exposed gaps in the field with scalpel-like precision, knowing exactly how they were going to fill those gaps, and our limited understanding of the some of the more advanced statistics, made it difficult to critique this article too harshly. In the end, we see this as a good example of test analysis and were able to use knowledge and skills learned in this testing class.

References

Lee, H., & Winke P. (2013). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. *Language Testing, 30*(1), 99-123.

Swales, J. M., & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (2nd ed.). Ann Arbor: University of Michigan.