The Use of Semi-scripted Speech in a Listening Placement Test for University Students

Martyn Clark
University of Maryland

This paper describes the feasibility of using semi-scripted spoken lectures as stimulus materials in a test of academic listening. The context for this study was the development of a revised test of academic listening designed to place enrolled university students into one of two levels of a language support course for non-native speakers. Because academic listening often involves listening to monologic speech such as lectures (Ferris & Tagg, 1996a), and because 'authentic' spoken language is qualitatively different to scripted speech (Biber et al., 2004), the revised test uses semi-scripted spoken mini-lectures as stimulus passages rather than relying on scripted material. Test questions were developed using only the informational elements that four *model comprehenders*, proficient English listeners (both native and non-native), were able to retain from a single hearing of the passages. Test data from 222 students were analysed using a Rasch methodology. Results show that this test development method did result in testable content that was appropriately targeted at the population of interest, though several aspects of the process could be improved. The paper concludes with some recommendations for using semi-scripted language in academic listening tests.

Key words: listening comprehension, academic lectures, test development, placement testing

Introduction

Despite increasing calls for learner autonomy and student-centered instructional practices at all educational levels, the reality is that many students in North American universities spend a great deal of their instructional time

Martyn Clark, Center for Advanced Study of Language (CASL), University of Maryland, 7005 52nd Avenue, College Park, MD 20745, USA. Email: mclark19@umd.edu.

listening to the professor or instructor (Ferris & Tagg, 1996a). Recent research suggests that extensive use of lectures is still common in undergraduate instruction, and is especially prevalent in science, technology, engineering, and math (STEM) courses, with more than half of survey respondents saying that they use lectures extensively (Housewright, Schonfeld, & Wulfson, 2013; Hurtado, Eagan, Pryor, Whang, & Tran, 2012), a number which has shown little change compared to a decade earlier. For students studying at the college level in a language other than their mother tongue, this means that the ability to comprehend orally presented material is a skill vital to academic success (Ferris & Tagg, 1996b). Therefore, it is important to identify those L2 English speaking students who could benefit from language support classes in the early part of their collegiate careers. This paper describes the revision of a test to assess listening comprehension and place students into such support courses.

The testing context

The English Language Institute (ELI) is an intensive English program at a large public university in the Pacific, and is charged with providing language support to those matriculated students who have not yet demonstrated sufficient language ability to be exempted from such support. The ELI offers courses in reading, writing, and listening, each at two levels. Prior to the start of the semester, potential ELI students take a placement battery to determine whether or not they need instruction in any of the skill areas covered by ELI courses and, if so, the appropriate level of that instruction.

The goal of ELI courses is to improve the students' ability to understand academic materials. In the listening skill area, the relevant support courses are ELI 70 (intermediate) and ELI 80 (advanced). Students placed in ELI 70 will generally be required to take ELI 80 the following term. The Academic Listening Test (ALT) is one of two assessments used to make placement decisions into the listening courses, the other being a dictation test. The original ALT was created in the late 1980s and had remained virtually unchanged since that time. It is a 40-item multiple-choice containing four sections. Two of the sections involve attending to a spoken message for content, in the form of short talks and one extended (8 minute) lecture. The remaining sections target sub-skills of listening, specifically the ability to: (a) infer unknown words from context (vocabulary), and (b) recognise appropriate discourse structuring devices (discourse structure). The language on the test is entirely scripted, though the extended lecture is delivered in a relaxed style, and includes some scripted false starts, hesitations, and fillers.

In preparation for creating a revised version of the ALT, a Rasch analysis of several years of test data from the original test was conducted (Clark, 2004). While model fit was generally acceptable, there were some misfitting items in the vocabulary and discourse structure sections, and results suggested that it might be prudent to replace those sections with items targeting more global comprehension. To this end, the goal of the ALT revision was not to simply update the current sections of the test with new items, but rather to redesign the test to focus exclusively on lecture comprehension.

Defining lecture language for the ALT

A major decision for the revised ALT was whether or not to continue the practice of using scripted material for the test. As the language testing field has matured, the notion of assessing the construct of listening has evolved from one of simple dictation tasks to one with much more emphasis on the realism and social context of the listening tasks (Taylor & Geranpayeh, 2011). Large corpora of academic spoken and written language have been developed and analysed to inform academic test development projects (e.g., Biber et al., 2004). Despite this, even though many listening tests are informed by authentic language and the idea of using authentic listening material for assessment is not novel, unscripted spoken input is not widespread in many tests used for university admissions. As Wagner (2013, pp. 7-8) notes, '[a] review of the spoken texts used in the listening section of some of the high stakes English proficiency tests (i.e., the IELTS, TOEFL, and Pearson Test of English [PTE]) suggests that virtually all of the texts are indeed scripted, written, and read aloud'. To be useful for the ALT, the test materials should ideally mimic the characteristics of authentic spoken lectures. One primary feature of lecture discourse is that the presentation of information is interwoven with commentary and reflection (Coulthard & Montgomery, 1981). In fact, the effect of various "discourse markers" on lecture comprehension has been the target of much research (e.g., Chaudron & Richards, 1986; Dunkel & Davis, 1994; Jung, 2003) in university L2 listening.

Though discourse markers are known to be a feature of academic lectures, it is unclear whether there is any such thing as a lecture genre in the strictest sense, as a genre requires, in part, that the speech act be highly structured and conventionalised (Bhatia, 1993). MacDonald, Badger, and White (2000) suggest that 'the degree of variability in the academic lecture appears to be such that it is actually hard to identify the contours of the genre very precisely' (p. 257). In an analysis of lecture introductions, Thompson (1994) was able to identify two basic moves, *Setting up Lecture Framework* and *Putting the Topic in Context*, but wasn't able to identify any default pattern across a series of lectures, concluding that 'there is no typical sequencing pattern, but rather a largely unpredictable mix of a small set of Functions and Sub-Functions' (p. 181). Building on this work, Lee (2009) found that differences in the lexico-grammatical features of rhetorical moves used in lecture introductions seem to be influenced by class size, but noted that the sequence of moves themselves is rather unpredictable. Likewise, though Cheng (2012) was able to identify 15 strategies used in lecture closings and though certain strategies were used more frequently than others, none of the strategies were obligatory.

Part of this variation may be due to the fact that although lectures tend to be largely monologic, they are still interactive. For example, Fahmy and Bilton (1990) discovered that over several weeks' worth of EAP lectures, many of the elaborations produced by the speakers were spontaneous and not in response to specific student questions. This seems to be due to the effect of a live audience, in that similar characteristics have also been noted in medical conference monologues (Webber, 2005). Strodt-Lopez (1991) found that asides are often used to establish greater global coherence by temporarily 'stepping back' (p. 132) from the immediate discourse to gain greater perspective. In fact, corpus analyses reveal that much instructional language in classroom teaching and study groups are marked for features of general face-to-face interaction (Biber, Conrad, Reppen, Byrd, & Helt, 2002). All of this suggests that the spontaneous nature of spoken language contributes greatly to spoken language's features and to its variability, even for semi-planned speech events such as lectures. As Lee (2009) notes, '[w]hile lecturers may plan and most likely utilize their notes in the delivery of lectures, these communicative events are nevertheless performed in real time' (p. 43).

The research points to a number of differences between scripted and unscripted speech. For example, a study asking the same subject to produce both written and spoken texts found that the spoken texts were 'more involved and more fragmented' (Redeker, 1984, p. 49) than their written counterparts. In other words, a written text read aloud is different from a spontaneously produced spoken text. Even when only looking at four grammatical features, several differences were found between written grammar and its spoken counterpart (Carter & McCarthy, 1995). Materials developed for language learning differ greatly from natural speech (Porter & Roberts, 1987), and Flowerdew and Miller (1997) found that even EAP specific listening materials are deficient in their portrayal of natural language. Taken together, these results would suggest that the act of merely reading a text aloud does not imbue it with the features of authentic spoken input. In fact, Coulthard and Montgomery (1981, p. 35) note that formal papers are difficult to follow when read aloud precisely because they lack the features of spoken discourse.

In an attempt to address this problem in an assessment situation, Hansen and Jensen (1994) describe an approach in which introductory lectures were taped, scanned for portions that could potentially be used as stand-alone lecturettes, and reproduced in the studio with the original lecturer reciting only the chosen portion. For the revised ALT, we chose to modify that general approach, and rather than try to reproduce portions of longer lectures, we chose to directly record spontaneously delivered mini-lectures in their entirety. In these mini-lectures, like most live university lectures, the speaker has planned the general focus and scope of the content, and perhaps even presented the material multiple times previously, but has not planned every specific word, phrase, and sentence to use in delivering that content. Thus, on a continuum from completely unplanned and unscripted extemporaneous speech to completely planned and scripted speech, the mini-lectures might be considered planned but not completely scripted, or semi-scripted.

Measuring lecture comprehension

After electing to use mini-lectures as input for the revised ALT, we turned to the question of how to measure mini-lecture comprehension. As Lynch (1998) notes, the listening process itself is 'unseen and inaccessible' (p. 6). In language processing, comprehension is the creation of an orderly mental structure from potentially chaotic input (Kintsch, 1998, p.5). Once a mental representation of the discourse has been constructed from the input, that representation can be used to answer questions about the input, provide a summary of the input, and so on (Kintsch, 1998, p. 163). For language assessment, Buck (2001, p. 144) defines the listening construct as the ability:

- To process extended samples of realistic spoken language, automatically and in real time,
- To understand the linguistic information that is unequivocally included in the text, and
- To make whatever inferences are unambiguously implicated by the content of the passage.

This is a useful definition, but problems still remain. What are inferences that are unambiguously implicated in the text? Language is notoriously underspecified. Gee (1999), for example, contends that there are over 100 interpretations of the sentence *Lung cancer death rates are clearly associated with an increase in smoking*. Although the end point of the comprehension process is a single mental representation of the text, Kintsch (1998) differentiates between the two components on which it is based, the textbase and the situation model.

The textbase is essentially the elements in the text itself and is 'what would be obtained if a patient linguist or psychologist were to translate the text into a propositional network and then integrate this network cycle by cycle…but without adding anything that is not explicitly specified in the text' (Kintsch, 1998, p. 103). As a result, the textbase is often 'impoverished and often even incoherent' (Kintsch, 1998, p. 103). The more complete situation model is derived from the interaction between the textbase and the comprehender's knowledge base.

Research by Tauroza and Allison (1994) noted that students could generally follow the logical argument of a lecture with the exception of its final evaluation section and, though their overall comprehension was much higher, native speakers had trouble with the same section of the lecture (Tauroza & Allison, 1990), suggesting that some comprehension problems are not unique to L2 listeners. Therefore, if we want to assess comprehension of a given passage, it seems that we need to have some idea of what is reasonably comprehensible from that passage – the inferences that are commonly made, as it were. This suggests that creating items to measure lecture comprehension should not begin with an exhaustive analysis of the text itself, but with reference to how that passage is generally comprehended by representative listeners. Thus, we decided to base ALT test items on information that was proven to be comprehensible by a sample of listeners.

Research questions

The goal of ALT test revision was to develop a listening comprehension test for use as a placement test into language support classes at a university. Test development uses semi-scripted rather than scripted passages, and test items are designed to target only those elements of the passages that were shown to be comprehensible. Though there is no single test of significance that will prove or disprove this approach to test development, there are several questions that, if answered positively, would provide support for this approach and the usefulness of revised ALT.

The rationale for not scripting materials is to obtain the kinds of features found in naturally occurring speech. In the Michigan Corpus of Academic Spoken English (MICASE) (Simpson, Briggs, Ovens, & Swales, 2002), for example, Swales and Malczewski (2001) found that the most frequent new episode signaling flags seem to be *OK* and *let*. Frequent use of words like *point* and *thing* are also common, depending on the speaker (Swales, 2001). Spontaneous elaborations, not directly in response to student questions, are also found in academic lectures (Fahmy & Bilton, 1990). To the extent that these types of phenomena are observed in the mini-lectures, the mini-lectures can be said to mimic features of real lectures.

1. *Do semi-scripted mini-lectures capture the features of natural academic language?*

Because test items are only written in response to what is actually produced during the live recordings, there is less control over the content than is possible with scripted material. For this reason, there is a danger that the mini-lectures may exhibit a more rambling delivery than would be the case with scripted material which could, in turn, lead to a low information content in terms of testable material. This is akin to the notion of sufficiency proposed by Norris, Brown, Hudson, and Yoshioka (1998) who note that the authenticity of a language sample does not necessarily guarantee its sufficiency for assessment purposes (p. 61).

2. *Do semi-scripted mini-lectures produce sufficient testable content?*

Because the purpose of a placement test is to spread students out, it is important that the test do this. The test should be capable of identifying at least two proficiency levels, as there are two instructional levels into which the students are to be placed.

3. *Does the revised ALT succeed in spreading the students out sufficiently to make placement decisions?*

As the goal of the ALT is placement on the basis of listening comprehension and not academic knowledge, it is important that the passages on the ALT be appropriate for the ELI population. The passages should be interesting but not overly technical.

4. *Are the level and content of the semi-scripted mini-lectures appropriate for the ELI population?*

In addition to positive results for these specific questions, for the ALT to be a useful instrument, the overall test needs to show acceptable psychometric properties.

Methodology

Test development

Test development consisted of four main stages: (a) recording the mini-lectures (b) obtaining summaries of those mini-lectures by *model comprehenders* (c) creating items from the mini-lecture summaries, and (d) obtaining intuitive pretest predictions of item difficulty. Each of these stages will be described in turn below with the procedures and materials for each stage outlined.

Stage one: Preparing mini-lectures

The goal for the ALT listening passages was to create speech samples that have some of the features of spoken academic lectures but are shorter in length and more appropriate for a lay audience. These mini-lectures would provide the auditory input for the ALT.

Participants for mini-lecture development. Participants were recruited from the university population to act as lecturers. No specific requirement for being a professor or instructor was set, though participants were expected to have experience in teaching or public speaking. In addition, participants included non-native-speakers of English, reflecting the actual makeup of the university's instructional faculty. A total of five people, including the researcher, participated as lecturers. All of the participants were PhD students in various departments of the university and all had experience teaching undergraduate classes. One speaker was female and the remaining four were male. The female speaker was a native speaker of Hindi; all male speakers were native speakers of English.

Procedure for developing mini-lectures. The general procedure for developing and recording the mini-lectures was as follows:

1. Identify a potentially interesting topic.
2. Using a college textbook and other sources, sketch out a brief description of what the mini-lecture will cover.
3. Practice delivering a lecture from the notes taken to ensure a smooth delivery.
4. In a recording studio, spontaneously deliver the mini-lecture based on the notes.

Participants were encouraged to record mini-lectures on several topics to build up a bank of passages. The onus for identifying topics and preparing notes was on the lecturers. This was done to ensure that all of the mini-lectures were on topics familiar to the speakers and would result in a natural delivery. Lecturers were given considerable leeway in choosing topics with the caveat that they should be pitched at an introductory level and not assume any specialised knowledge. Lecturers were encouraged, however, to choose a particular idea or problem to focus on during the mini-lecture, such as explaining a basic concept in their field in layman's terms. It was hoped that this minimal direction would help to give each mini-lecture a clear focus.

Procedure for recording mini-lectures. The recording sessions were conducted by a trained audio technician in a sound attenuated studio with the researcher present to act as a live audience. The lecturers did not use a recording script, but had access to any notes or outlines that they had previously prepared. Lectures were recorded as single continuous takes directly into digital audio format. In two instances, the lecturer stopped in the middle of their presentation and requested retakes, which were then recorded.

The five lecturers recorded a total of 13 mini-lectures of varying topics and lengths, shown in Table 1. After the recording session, the mini-lectures were screened by the researcher. Three of the recorded mini-lectures were identified as problematic by the researcher due to length or content and were not developed further. The 10 remaining mini-lectures were used in the subsequent summary writing session.

Table 1. List of mini-lectures.

Topic	Speaker	Length
India's Three Language Policy	Female A	8:20
Marketing, Branding, and Advertising	Male A	3:56
Fear	Male B	6:49
Freedom	Male C	3:22
English Schools in India	Female A	7:41
Public Relations	Male A	4:27
Research Paper Assignment Instructions	Male C	4:48
Extraterrestrial Life	Male D	7:04
Checks & Balances in the Constitution	Male C	4:53
Academy Award Voting	Male D	8:50
Academic Honesty*	Male A	4:05
Arranged Marriages in India*	Female A	10:04
Importance of Résumés*	Male A	5:32

Note. *Eliminated

Stage two: Summary writing session

Before writing items for the mini-lectures, it was necessary to determine to what information in any given passage was generally comprehensible after just one hearing. If test items are based on a close reading of a detailed transcript, then there is a danger that the items will reflect the kinds of inferences and analysis only possible after such a close reading. In order to be a test of listening comprehension, it was important that items developed for the ALT were truly reflective of comprehension based on listening and not reading. To this end, it was necessary to enlist the help of *model comprehenders*, people whose comprehension would form a baseline for the information typically available after a single hearing of a given mini-lecture.

Participants for summary writing session. Model comprehenders were recruited from the general university population, with the stipulation that they exceed the ELI's exemption requirements in terms of English ability. A total of four people participated as model comprehenders, three male and one female. Three of the participants were PhD students in Second Language Acquisition and the fourth had recently completed an MA in Second Language Studies. One of the male participant's was a native speaker of Japanese and the female participant was a native speaker of German. The remaining two participants were native speakers of English. One of the model comprehenders also acted as a lecturer and did not participate in the summarising for the mini-lectures that he had given. All participants had experience teaching in the ELI.

Summary writing protocol. Model comprehenders were asked to listen once to each of the mini-lectures, taking notes as they listened. Once they finished, they were asked to write a summary based on their notes and any other details that they could recall. The instructions were to recall as much of the information as possible. They were not to listen to each lecture more than once. Because listening to several mini-lectures can be quite tiring and time consuming, it was not practical to do in one sitting. Instead, participants for this part of the project each received a CD with all of the mini-lectures, each mini-lecture identified with a number preceding it, and were asked to complete the task at their leisure. Though participants were asked to keep track of any irregularities (i.e., interruptions) that occurred during their individual recalls, none were reported. Both the raw lecture notes as well as the summaries were collected from each participant.

Mini-lecture ratings. Because all participants in this stage of the project had experience teaching in the ELI, their evaluation of the interest level, difficulty, and appropriacy for a test of listening of each of the mini-lectures was elicited using a 5-point Likert scale. This was done to identify the most promising lectures for which to create items. The ratings given to the passages are shown in Table 2. Based on these ratings, the two mini-lectures with the lowest interest and appropriacy scores were dropped. This left a total of eight lectures for which items were produced.

Table 2. Lecture ratings sorted by Appropriacy.

Topic	Interest	Difficulty	Appropriacy
India's Three-Language Policy	4.25	1.50	4.75
Checks & Balances in the Constitution	3.33	2.00	4.67
Extraterrestrial Life	4.50	2.75	4.50
English Schools in India	4.00	1.50	4.50
Academy Award Voting	3.75	3.25	3.75
Freedom	4.00	1.67	3.67
Marketing, Branding, and Advertising	3.25	3.25	3.50
Public Relations	2.50	3.25	2.75
Research Paper Assignment Instructions*	1.33	2.00	2.33
Fear*	3.00	3.00	2.00

Note: *Eliminated. Ratings are on 5-point scale; higher number indicates more interest, difficulty, appropriacy.

Stage three: Item writing

Once all of the summary protocols were received, they were used as the basis for item writing. Any information that appeared on all of the protocols for a given passage, either in the notes or in the summary, was considered appropriate for item development. That is, items were only created that ask for that information that was identified by the model comprehenders after one listening. The item writing was done primarily by the researcher. Once items for each lecture were written, they were given to a second person who was instructed to try to answer them without the benefit of hearing the lectures themselves. Through this process, several items were identified as being too obvious and were revised or eliminated.

An example will help to illustrate the process. Below is part of the "India's Three Language Policy" mini-lecture. In this excerpt, potential test information has been highlighted.

> Well, uhm, it was decided that the Indian language which would be would become the official language of the, of Independent India, would be Hindi. **And that was because, ah, 39% of the Indian population spoke Hindi, and this is the greatest number of people that speak one language**. And also, of course, it was a a political decision because, ah, Hindi's also the first language of, ahm, a few states which have provided most of the political leaders of India, and these are the people who was involved in writing the constitution, so there was that political aspect which influenced it as well.

This information was present in all four model comprehender summaries:

- Model Comprehender #1: Hindi represented about 39% of the population's language and was chosen because the new leaders were widely represented by Hindi speakers.

- Model Comprehender #2: The most common language is Hindi and 39% of the population speak it.
- Model Comprehender #3: Thus, they chose Hindi as the official language, since Hindi was the most wide-spread of all the Indian language – about 39% of the Indian population at the time spoke it.
- Model Comprehender #4: Hindi was chosen for two reasons: first, Hindi speakers constitute the largest language group (39% of the population), so this was thought to be useful for quite many people, and second, Hindi was the language of the important political leaders who were involved in establishing the Indian government.

This information was developed into the following item:

D4 Approximately what percent of the Indian population spoke Hindi?

a. 14 percent.

b. 39 percent.

c. 63 percent.

d. 89 percent.

Although one criticism of discrete-point items is that they tend to mix trivial details and main ideas, the fact that all model comprehenders identified this particular information in their summaries indicates that it is salient from a single listening to this mini-lecture.

Stage four: Intuitive predictions of item difficulty

To ensure that items and mini-lectures were at an appropriate level of difficulty for the ELI population, intuitive predictions of difficulty were solicited. The intuitive judgments were also aimed at locating the items relative to the relevant ELI population, i.e., students in intermediate (ELI 70) and advanced (ELI 80) listening classes. These judgments are considered intuitive in that the participants were given no specific features or criteria to consider in their assessment of perceived difficulty other than their own experience with ELI students.

Participants. Four people participated in the pretest judging of item difficulty. All four were current or former instructors in the ELI, including the Lead and Co-Lead Teacher for the Listening and Speaking skill area and the former Lead Teacher for the Reading Skill area. All of the participants were female. Three were native speakers of Japanese and the fourth was a native speaker of

Korean. Together, this group represented many semesters worth of experience interacting with ELI students.

Procedure. Participants listened to each lecture once, and took notes on a sheet provided by the researcher. After listening, they indicated the overall difficulty of the passage on a graphical rating scale that included labels indicating listening ability below, at, or above ELI course levels. This type of scale has previously been successfully employed in rating sessions (Myford, 2002). Participants then turned the page and answered individual test items, indicating each individual item's relative difficulty on a separate graphical rating scale for each item. No specific time limit was given for this part of the task. Once all of the participants had rated the items, they were encouraged to discuss their choices as a group.

Though the graphical rating scale allows for fine gradations, upon inspection of the ratings and review of the rating session recording, it became clear that the participants were only making broad distinctions (e.g., ELI 80-Level, between ELI 70 and 80 level) between items rather than fine-grained ones. For this reason, the rating scale was recalculated to reflect a 7-point scale (ranging from *Well below ELI 70* to *Well above ELI 80*). These revised ratings were analysed using the FACETS computer program (Linacre, 1994) which performs a Rasch analysis for rated data.

Pilot testing

Participants. To collect actual test data, the test was piloted during regularly scheduled ELIPT administrations. The total testing time for the entire ELIPT is four hours, with the ALT portion being administered approximately 90 minutes into the process. A total of 222 students took the pilot test over four separate test administrations. Because data was collected during a regular test administration, it was not possible to collect individual biodata from each student. The ELI does routinely collect general background information during registration, though students do not always provide it. By definition, students taking the ELIPT have TOEFL scores between 500 and 600 on the paper-based test (PBT) and 173 and 250 on the computer-based test (CBT). For students who reported TOEFL scores, the mean for the PBT was 554 and the mean for the CBT was 216. The mean for the TOEFL Listening section scores were 54.15 out of a possible 68 (PBT) and 21.44 out of a possible 30 (CBT). ELI students for these administrations came from a variety of linguistic backgrounds as summarised in Table 3. Although this was a convenience sample, it is also the population to which the results should be generalised in that it is representative all of the test takers for that semester.

Table 3. Language background of test-takers.

First Language	Percent
Japanese	31.0%
Chinese	16.8%
Korean	15.1%
Vietnamese	4.3%
Thai	3.9%
Filipino	3.0%
Other	25.9%

Note. Other includes: Arabic, Chamorro, French, German, Indonesian, Kurdish, Marshallese, Mongolian, Norwegian, Palauan, Pohnpeian, Polish, Samoan, Sinhala, Spanish, Tamil, Tetun, Ukranian, Yapese.

Materials. The final version of the ALT consisted of five listening passages described in the preceding section, with a total 35 items. Although a longer version with more passages and items was originally planned, it was shortened to fit the time requirements imposed by the ELI's testing schedule. The mini-lectures that were identified as most appropriate by the ELI teachers were chosen for the test, taking into account the topic, speaker, and length.

Results and discussion

Do semi-scripted mini-lectures capture the features of natural academic language?

A prime objective in using semi-scripted mini-lectures on the revised ALT was to try to capture the features of monologic speech in academic settings. It is clear when listening to the mini-lectures that they are very much in a spoken mode – there are many pauses and hesitations, with more pauses per minute (calculated with Praat; Boersma, & Weenink, 2006) than the scripted lecture on the original ALT (shown for comparison in Table 4 and subsequent tables). This difference is probably the primary reason that the passages are instantly recognisable as being samples of spoken language, rather than text that has been read aloud. Common discourse micro-markers (e.g., Chaudron & Richards, 1986) were also evident in the passages (see Table 5).

Table 4. Filled and unfilled pause count and duration by lecture.

	Count				Duration					
	Total		Per Minute		Mean (SD)		Max		Min	
Lecture	FP	UFP	FP	UFP	FP	UFP	FP	UFP	FP	UFP
ALT (Ref)	20	116	2.62	15.20	0.35(0.14)	0.73(0.29)	0.61	1.57	0.12	0.2
Checks	13	84	2.66	17.20	0.37(0.12)	0.48(0.33)	0.53	2.26	0.20	0.2
Marketing	15	62	3.81	15.76	0.28(0.09)	0.45(0.21)	0.41	1.24	0.12	0.2
Freedom	8	128	2.38	38.02	0.37(0.12)	0.51(0.33)	0.58	1.62	0.23	0.2
3 Languages	73	222	8.76	26.64	0.38(0.17)	0.45(0.26)	0.82	1.93	0.10	0.2
Alien Life	102	204	14.43	28.87	0.24(0.10)	0.56(0.30)	0.44	1.59	0.07	0.2
Mean	42.2	140	6.41	25.29			0.56	1.73	0.14	0.2

Note. FP = Filled Pause, UFP = Unfilled Pause

Table 5. Number of discourse markers by lecture.

Lecture	you know	well	so	now	actually	anyway	basically	I mean	OK	Total
ALT (Ref)	1	6	5	3	0	1	2	1	1	20
Checks	0	0	4	2	1	1	0	0	0	8
Marketing	0	1	4	2	2	0	0	0	0	9
Freedom	0	2	8	0	0	0	0	0	1	11
3 Languages	0	1	7	2	0	0	0	1	2	13
Alien Life	4	3	18	9	2	0	2	1	4	43
Mean	0.8	1.4	8.2	3.0	1.0	0.2	0.4	0.4	1.4	16.8

The mini-lectures also showed instances of self-correction and the use of questions to the audience (see Table 6). The use of questions has been noted as a feature of monologic speech, as speakers are continually assessing the comprehension of their audience and trying to draw them into the material by asking questions, even though answers to these questions are not expected or sought (e.g., Crawford Camiciottoli, 2008; Thompson, 1998). Finally, as seen in Table 7, the mini-lectures exhibited a high degree of personalisation common to academic lectures in terms of the use of the first person (see Biber et al., 2004) and especially the use of 'we' (see Fortanet, 2004).

Table 6. Number of self-corrections and questions by lecture.

Lecture	Self-corrections	Questions
ALT (Ref)	8	1
Checks & Balances	8	1
Marketing	15	1
Freedom	10	6
Three Language Policy	18	1
Extraterrestrial Life	6	3
Mean	11.4	2.4

Table 7. Number of personalization markers by lecture.

Lecture	we	I	you
ALT (Ref)	1	8	19
Checks & Balances	6	2	1
Marketing	9	2	1
Freedom	21	1	3
Three Language	0	7	8
Extraterrestrial Life	15	8	13
Mean	10.2	4.0	5.2

A short excerpt from the "Extraterrestrial Life" mini-lecture illustrates the type of language produced, highlighting some of the features mentioned:

> *So*, the question *we* need *to ask to think through* Fermi's Paradox of, *you know*, where are the aliens, is approximately how many extraterrestrial civilizations might there be, *ah*, which would be people who might be able to visit us, *so* how many aliens might there be out there in the Galaxy. And, *ah*, there's actually a man by the name of Dr. Frank Drake, *ah*, who in 1961, came up with something called the Drake Equation, which attempted *to, ah, give, ah, a bit of a way of thinking about* how many extraterrestrial civilizations there might be. *So* today *I'm going to* give you *a, ah, abridged version, a simplified version*, of the Drake Equation.

Although the scripted ALT lecture shared many of these features, the new passages represent a wider range, most likely due to the use of several speakers. These results suggest that asking people to produce mini-lectures based on notes results in language that contains many of the elements that have been identified in the literature as features found in authentic academic lectures. That said, it must be reiterated that the genre of academic lectures has generally defied clear specification and it is unclear to what extent those features of academic speech identified in previous corpus studies are representative of academic lectures in other institutions not included in the corpus.

Do semi-scripted mini-lectures produce sufficient testable content?

Of the 13 passages originally recorded, only eight were considered suitable for item development (albeit for a variety of different reasons). This suggests that not every passage will be a good candidate for test material. Nevertheless, the process did create enough viable passages so that options were available for the final test. By establishing ongoing recording sessions, it should be possible to develop a bank of potential mini-lectures in a reasonable amount of time.

In addition to a passage being viable input or not, it is also possible to conceptualise listening passages as being more or less efficient. An efficient passage would lend itself to a reasonable number of questions given the length of the passage. On the original ALT, the scripted "lecture" passage was 7 minutes and 38 seconds long and had 11 items associated with it. In other words, there is one test item for every 41.73 seconds of listening material. The new passages were only slightly less efficient overall, with an average of one item per 45.87 seconds of material (see Table 8) for all passages, and an average of one item per 43.79 seconds of material for the five passages included in the final test. In addition, it seems that shorter passages were slightly more efficient than longer ones, suggesting that there is a point of diminishing returns for passage length, at least with the passages recorded for this project. (Note that not all of the passages in Table 8 were included in the final version of the ALT although items were developed for them).

Table 8. Efficiency of listening passages

Topic	Length (Seconds)	No. Items	Seconds/Item
Reference Passage (Original ALT)*	459	11	41.73
Checks & Balances	293	6	48.83
Marketing, Advertising, Branding	236	6	39.33
Freedom	202	6	33.67
India's Three Language Policy	500	10	50.00
Extraterrestrial Life	424	9	47.11
Public Relations*	267	6	44.50
English Schools in India*	461	8	57.63
Average (SD)	340.43 (118.81)	7.29 (1.7)	45.87 (7.74)

*Note: Not on final test.

Alternative item formats, notably summary cloze, would have provided a higher item count per unit of time measurement. This in turn would have had implications for test reliability, discussed later. In general, the use of semi-scripted material did not render the creation of a reasonable number of items considerably more difficult than using scripted material. Even if it had, some of the language in the mini-lectures would be very hard to duplicate in scripted input without sounding stilted. Of course, the true efficiency can only be determined with reference to the quality of the developed items.

Does the revised ALT succeed in spreading the students out sufficiently to make placement decisions?

Reliability and separation

A Rasch analysis of test data was performed using WINSTEPS (Linacre, 2006). Table 9 shows the Rasch summary statistics for the person measures on the revised ALT. The person reliability in Table 9 is analogous to Cronbach alpha reliability in classical test theory. The real and model estimates represent the lower and upper bound respectively. As can be seen from the table, the scores showed reasonable reliability given the relatively small number of items (k = 35) on the test.

Table 9. Summary of 222 measured persons.

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	20.4	34.9	.44	.39	1.00	.0	1.00	.0
SD	5.1	.4	.79	.06	.15	1.0	.24	1.0
Max	34.0	35.0	3.85	1.02	1.40	2.6	2.27	2.3
Min	7.0	31.0	-1.59	.37	.67	-2.8	.46	-2.7

Note. Winsteps v3.60 Table 3.1., Real RMSE=.41, Adj.SD=.67, Separation=1.65, Person Reliability=.73, Model RMSE=.40, Adj.SD=.68, Separation=1.72, Person Reliability=.75

In addition to reliability, WINSTEPS also provides a measure of separation. The higher the separation, the easier it becomes to distinguish between persons (Wright & Stone, 2004). The separation measure can be used to calculate STRATA which indicate the number of statistically distinct ability levels (Schumacker, 2004, p. 245). For the revised ALT, this results in a value of 2.53, indicating that at least two statistically distinct groups can be identified in the test results. Because the main purpose of the ALT is to determine placement into ELI 70 or ELI 80, the fact that more than two distinct ability levels can be distinguished is important. Again, more items would increase the separation, as there is more information about each test taker's position on the listening ability measure. Applying the Spearman-Brown prophecy formula for separation (Linacre, 2000), one can calculate that a test length of approximately 52 similar items would have resulted in a separation reliability sufficient to identify three distinct STRATA.

Person fit

Table 9 also gives an overall indication of the fit of the persons to the model. Two types of fit statistic exist – *infit* and *outfit*. Outfit is influenced by very unexpected responses to items, such as when persons of low ability get the most difficult items correct. Infit, on the other hand, is influenced by an unexpected pattern of responses near a person's ability estimate, that is, 'the degree of fit in the most typical observations in the matrix' (McNamara, 1996, p. 172). In WINSTEPS, two different fit statistics are available for assessing model fit – mean squares (MNSQ) and a standardised transformation of the mean-square to approximate a t-statistic (ZSTD). Infit and outfit mean square (MNSQ) have an expected value of 1.00. The standardised fit statistics (ZSTD) have an expected value of 0.0 with a standard deviation of 1.00. The values in Table 9 suggest that, for the most part, persons show a good fit to the model, with fewer than 2% of the cases showing misfit (McNamara, 1996, p. 178). The practical implication of this is that test-takers can be compared meaningfully on the metric, as higher ALT scores indicate greater listening ability.

Item separation and reliability. Table 10 shows information about the 35 ALT items. The high separation and reliability values indicate that the relative order of items on the test in terms of difficulty is consistent and would be reproduced with another sample of test takers. The average measures for infit and outfit are also within expected values. It should be noted that both person and item reliability are important for determining the reproducibility of the measures, but are not directly measures of item quality.

Table 10. Summary of 35 measured items.

	Raw			Model	Infit		Outfit	
	Score	Count	Measure	Error	MNSQ	ZSTD	MNSQ	ZSTD
Mean	129.3	221.5	.00	.16	1.00	.0	1.00	.1
SD	37.5	.6	.85	.01	.08	1.4	.12	1.4
Max	189.0	222.0	1.63	.20	1.17	3.5	1.30	3.4
Min	57.0	220.0	-1.53	.14	.89	-2.0	.76	-1.8

Note. Winsteps v3.60 Table 3.1., Real RMSE=.16, Adj.SD=.83, Separation=5.29, Item Reliability=.97, Model RMSE=.16, Adj.SD=.84, Separation=5.36, Item Reliability=.97

Item fit

The issue of fit statistics for Rasch models has been called one of the most contentious areas in the Rasch literature (Bond & Fox, 2001) and recommendations for how to best interpret model fit vary. It has been argued that because standardised residuals are sensitive to sample size, the mean square fit measures are more reliable and should be considered first, with the ZSTD only being used to "salvage" items that have poor MNSQ fit (Linacre, 2006). Other researchers have questioned this view and recommend the standardised fit index (ZSTD) (Smith, 2004). By convention, MNSQ fit values between .7 and 1.3 are considered productive for measurement (e.g., Bond & Fox, 2001), that is, although some items may show slight misfit to the model, the effect on the utility of the measure is negligible. It is also possible to approach the diagnosis of misfit from a local perspective as well and consider those items that are greater than one (Wright & Stone, 2004) or two (McNamara, 1996) standard deviations from the mean to be potentially misfitting. As ZSTD fit statistics approximate the *t*-test, items exceeding +/- 2.0 are considered misfitting when using this statistic.

Given these various perspectives, potentially misfitting items are summarised in Table 11. As can be seen, no misfitting items were found using the MNSQ statistic. Slight misfit was found in only four items when using the ZSTD criteria. These four items (A1, B1, C1, E7) also had the lowest point-measure correlation (.14, .09, .15, .17, respectively) indicating that success on these items was only weakly correlated with an increasing ability estimate. A detailed analysis of these four items can be found in Clark (2007).

Table 11. Potentially misfitting items by statistic.

	Convention		Local Fit
	MNSQ	ZSTD	MNSQ > 2 SD
Infit	None	E7, A1, C1	A1
Outfit	None	B1, A1, C1, E7	B1

Are the level and content of the semi-scripted mini-lectures appropriate for the ELI population?

As noted in a previous section, the mini-lectures with the highest interest and appropriateness scores were selected for item development (see Table 2). In the debriefing for the rating session, the participants agreed that all of the passages and items were at an appropriate level for ELI students. This finding is borne out in the results of their intuitive difficulty ratings, as the majority of items were classified as being at the ELI 80 level or between the ELI 70 and ELI 80 level (see Table 12.) Given that the test is designed primarily to determine placement into ELI 70 or ELI 80, this is encouraging. In fact, only four of the seven possible scale categories were used, indicating that the test items were perceived to be rather similar in difficulty. The correlation between the intuitive predictions and empirical item difficulties for individual items, though positive, was not particularly strong at *r* = 0.37. This is partially due to the relatively narrow range of both item difficulties and ratings. Because most of the items were fairly well-targeted to the range of ability needed to be measured by the ALT, there was little perceptible difference between the items.

Table 12. Intuitive rating summary.

	Counts			Calibration					
Score	Used	%	Cum%	AvMea	ExMea	MnSq	Mea	SE	Label
3	18	11%	11%	-2.18	-2.09	0.8			ELI 70 Level
4	57	35%	46%	-.64	-0.65	1.0	-2.54	.30	Between 70–80
5	69	42%	88%	.88	.84	1.0	-.06	.21	ELI 80 Level
6	20	12%	100%	1.75	1.82	1.0	2.59	.27	Above ELI 80

Note. AvMea = Average Measure, ExMea = Expected Measure, MnSq = Outfit Mean Square, Mea = Measure

Conclusion

The following research questions were posed at the beginning of this paper.

1. Do semi-scripted mini-lectures capture features of natural academic language?

2. Do semi-scripted mini-lectures provide sufficient testable content?

3. Does the revised ALT succeed in spreading students out sufficiently to make placement decisions?

4. Are the level and content of the semi-scripted mini-lectures appropriate for the ELI population?

These questions can all be answered in the affirmative. The use of semi-scripted mini-lectures and model comprehenders in this study allowed for the operationalisation of listening ability that was based on natural spoken input and not on carefully scripted material read aloud. As a result of this decision, some control was lost in terms of specifying the particular features of the stimulus passages beforehand. This, in turn, affected the types of questions that could be asked based on the passage summaries. Thus, if one is interested in assessing particular aspects of academic language, such as the use of specific discourse signaling devices, this approach may not provide the desired level of control over test content. However, given that academic lectures represent a somewhat ill-defined genre and that the target for the ALT was general listening comprehension ability rather than specific sub-skills, the advantage of this process is the naturalness of the language and the saliency of the information probed by the items. If listening comprehension tests are to provide information about examinees' ability to understand "real" speech, the method employed here is a viable alternative to the use of scripted listening passages.

Although the focus of this paper is on test development rather than use, a number of use-oriented sources of information were employed in the creation of the ALT. Summarisers all had experience with ELI students, and their knowledge of the ELI population was used to cull the initial pool of lectures. Intuitive predictions were obtained from ELI content specialists, and their insights during the rating debriefing session were used to help inform test analysis and item selection. These procedures ensured the appropriateness of the items for the ELI population. The Rasch analysis suggests that the test succeeded in producing sufficient spread in the test-takers to identify at least two distinct groups for the purpose of placement, and had acceptable psychometric qualities.

Limitations and directions for future research

As with any scholarly investigation, this study has its share of limitations. The determination of mini-lecture quality and passage efficiency were based mostly on subjective criteria. Both of these determinations would have benefited from a better metric for comparison. A more comprehensive approach to test validation that included examinations of cut scores and subsequent decisions would help strengthen the overall validity evidence for the ALT.

Because this approach to item development showed promise, it would be useful to try it in other testing contexts and for other levels of ability. More research could also be usefully done on ways to improve the efficiency of the mini-lectures. Because there seemed to be a point of diminishing returns for passage length in terms of passage efficiency, it would be useful to investigate this aspect of the mini-lectures more thoroughly. The contents of mini-lectures could also be potentially improved. For this study, lecturers were given minimal guidance as they prepared for the mini-lecture recording sessions. As a result, some of the mini-lectures were deemed unsuitable for further development. Although test development projects normally involve the removal of non-performing content, perhaps a training session to highlight the features of "productive" mini-lectures could be presented to the lecturers to give them a more concrete idea of what types of mini-lectures make the best raw material for item development. It would be important, however, to make sure that the mini-lectures maintain the spontaneity that was desired in the first place.

For the multiple-choice format, the use of model comprehenders was good for the development of item stems, but the researcher essentially used instinct to create the distractors. Perhaps model non-comprehenders, students whose ability is expected to be insufficient to fully comprehend the passages, could be enlisted to help develop distractors that reflect incomplete comprehension. Though not intended as such, the intuitive item prediction session with ELI teachers provided many insights into the items and distractors. A similar session with ELI students in which they provide responses to open-ended versions of the items to provide an empirical basis for creating distractors might be equally informative, provided that test security can be maintained.

Although it may not be possible in the ELI context given the rapid score turnaround required, the exploration of alternative item formats such as summary cloze would also be useful. Test development could follow essentially the same model. Because model comprehenders have already provided summaries for the mini-lectures, it would be a relatively easy task to create an initial cloze summary from a combination of summary protocols. Of course, the cloze test would also have to be piloted to ensure that it is functioning properly. As is the case in the ELI, the multiple-choice format offers great efficiency advantages for testing large numbers of students at the same time.

Acknowledgements

This research was conducted as part of the author's doctoral dissertation and was made possible, in part, by a grant from the Educational Testing Service.

The author

Martyn Clark, Department of Second Language Studies, University of Hawai`i at Mānoa. Martyn Clark is now at the University of Maryland Center for Advanced Study of Language (CASL).

References

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings*. Harlow, Essex: Longman.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*(1), 9-48. doi: 10.2307/3588359.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004, January). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus* (RM-04-03, TOEFL-MS-25). Retrieved from https://www.ets.org/Media/Research/pdf/RM-04-03.pdf.

Boersma, P., & Weenink, D. (2006). *Praat: doing phonetics by computer (Version 4.4.13) [Computer program]*. (Available from http://www.praat.org/)

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics*, *16*(2), 141-158. doi: 10.1093/applin/16.2.141

Chaudron, C., & Richards, J. C. (1986). The effect of discourse markers on the comprehension of lectures. *Applied Linguistics*, *7*(2), 113–127.

Cheng, S. W. (2012). 'That's it for today': Academic lecture closings and the impact of class size. *English for Specific Purposes, 31,* 234—248. doi: 10.1016/j.esp.2012.05.004

Clark, M. (2004). By the numbers: The rationale for Rasch analysis in placement testing. *Second Language Studies, 22*, 61-90.

Clark, M. (2007). *Listening placement test development and analysis from a Rasch perspective*. Unpublished doctoral dissertation, University of Hawai'i at Mānoa.

Coulthard, M., & Montgomery, M. (1981). The structure of the monologue. In M. Coulthard & M. Montgomery (Eds.), *Studies in discourse analysis* (pp. 31–39). London: Routledge & Kegan Paul.

Crawford Camiciottoli, B. (2008). Interaction in academic lectures vs. written text materials: The case of questions. *Journal of Pragmatics, 40*, 1216-1231. doi: 10.1016/j.pragma.2007.08.007

Dunkel, P., & Davis, J. N. (1994). The effect of rhetorical signaling cues on the recall of English lecture information by speakers of English as a native or second language. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 55–74). Cambridge University.

Fahmy, J. J., & Bilton, L. (1990). Listening and note-taking in higher education. In A. Sarinee (Ed.), *Methodology in the nineties. Anthology Series 24* (pp. 106–126). Retrieved from ERIC database. (ED366189)

Ferris, D., & Tagg, T. (1996a). Academic oral communication needs of EAP learners: What subject-matter instructors actually require. *TESOL Quarterly, 30*(1), 31–58. doi: 10.2307/3587606

Ferris, D., & Tagg, T. (1996b). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *TESOL Quarterly, 30* (2), 297–320. doi: 10.2307/3588145

Flowerdew, J. L., & Miller, L. (1997). The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes, 16*(1), 27-46. doi: 10.1016/S0889-4906(96)00030-0

Fortanet, I. (2004). The use of 'we' in university lectures: reference and function. *English for Specific Purposes, 23*(1), 45-66. doi:10.1016/S0889-4906(03)00018-8

Gee, J. P. (1999). *An introduction to discourse analysis: Theory and method*. New York: Routledge.

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In John Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 241–268). Cambridge: Cambridge University.

Housewright, R., Schonfeld, R. C., & Wulfson, K. (2013). *Ithaka S+R US Faculty Survey*. ITHAKA.

Hurtado, S., Eagan, M. K., Pryor, J. H., Whang, H., & Tran, S. (2012). *Undergraduate teaching faculty: The 2010–2011 HERI Faculty Survey*. Los Angeles: Higher Education Research Institute, UCLA.

Jung, E. H. S. (2003). The role of discourse signaling cues in second language listening comprehension. *The Modern Language Journal*, *87*(4), 562–577. doi: 10.1111/1540-4781.00208.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Lee, J. J. (2009). Size matters: an exploratory comparison of small- and large-class university lecture introductions. *English for Specific Purposes*, *28*, 42-57. doi: 10.1016/j.esp.2008.11.001

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: Mesa.

Linacre, J. M. (2000).Predicting reliabilities and separations of different length tests. *Rasch Measurement Transactions*, *14* (3), 767. Retrieved 16 February, 2002, from http://www.rasch.org/rmt/rmt143j.htm.

Linacre, J. M. (2006). *Winsteps (Version 2.6.0)* [Computer program]. Available from http://www.winsteps.com.

Lynch, T. (1998). Theoretical perspectives on listening. *Annual Review of Applied Linguistics*, *18*, 3–19. doi: 10.1017/S0267190500003457.

MacDonald, M., Badger, R., & White, G. (2000). The real thing? Authenticity and academic listening. *English for Specific Purposes*, *19*, 253–267. doi: 10.1016/S0889-4906(98)00028-3

McNamara, T. (1996). *Measuring second language performance*. New York: Longman.

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education*, *15*(2), 187–215. doi: 10.1207/S15324818AME1502_04

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (Technical Report No. 18). University of Hawaiʻi at Mānoa.

Porter, D., & Roberts, J. (1987). Authentic listening activities. In M. H. Long & J. C. Richards (Eds.), *Methodology in TESOL: A book of readings* (pp. 177–187). New York: Newbury House.

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, *7*, 43-55. doi: 10.1080/01638538409544580

Schumacker, R. E. (2004). Rasch measurement: The dichotomous model. In E. V. Smith Jr & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 226–257). Maple Grove, MN: Journal of Applied Measurement Press.

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). *The Michigan Corpus of Academic Spoken English*. Ann Arbor, MI: The Regents of the University of Michigan.

Smith, R. M. (2004). Fit analysis in latent trait measurement models. In E. V. Jr. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 73–92). Maple Grove: MN: Journal of Applied Measurement Press.

Strodt-Lopez, B. (1991). Tying it all in: Asides in university lectures. *Applied Linguistics*, *12* (2), 117-140. doi: 10.1093/applin/12.2.117

Swales, J. M. (2001). Metatalk in American academic talk: The case of point and thing. *Journal of English Linguistics*, *29*(1), 449—460. doi: 10.117/00754240122005189

Swales, J. M., & Malczewski, B. (2001). Discourse management and new-episode flags in MICASE. In R. C. Simpson & J. M. Swales (Eds.), *Corpus linguistics in North America: Selections from the 1999 symposium* (p. 145-164). Ann Arbor, MI: Michigan University Press.

Tauroza, S., & Allison, D. (1990). Speech rates in British English. *Applied Linguistics*, *11*(1), 90–105. doi: 10.1093/applin/11.1.90

Tauroza, S., & Allison, D. (1994). Expectation-driven understanding in information systems lecture comprehension. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 35–54). Cambridge University Press.

Taylor, L., & Geranpayeh, A. (2011). Assessing listening for academic purposes: Defining and operationalizing the test construct. *Journal of English for Academic Purposes, 10,* 89–101. doi:10.1016/j.jeap.2011.03.002

Thompson, S. E. (1994). Frameworks and contexts: A genre-based approach to analysing lecture introductions. *English for Specific Purposes*, *13*(2), 171–186. doi: 10.1016/0889-4906(94)90014-0.

Wagner, E. (2013). Assessing listening. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 47—63). Oxford, UK: Wiley-Blackwell.

Webber, P. (2005). Interactive features in medical conference monologue. *English for Specific Purposes*, *24* (2), 157–181. doi: 10.1016/j.esp.2004.02.003

Wright, B. D., & Stone, M. H. (2004). *Making measures*. Chicago: The Phaneron Press.