The SPM*-based Speaking Test at Toyo Eiwa University:
A look into the contextualized scoring system

SPM を基準としたスピーキング・テスト：
学習者のコンテクストに適した評価システムの開発

*SPM stands for “sentences per minute.”

Stephen Soresi, Suzuki Takashi
スピーキング・ソレイシィ、鈴木 卓

本学国際社会学科国際コミュニケーション専攻では、学生の英語コミュニケーション能力向上のため、またそのような教育・学習効果測定の一環として、2007年度より対面面接型英語能力試験（スピーキングテスト）を実施している。このテストは、SPM In-class Speech Test (Soresi, 2004, 2008) を原型として、本学学生の進度等の特性に応じて調整を加えた(Contextualized)もので、質問に対する自由応答の発話率（Speech Rate）を測定する。この試験の定期的な実施により、英語コミュニケーション能力の中でも、とくに談話を継続・展開し、表現内容を発展させる能力の向上が期待できる。本稿では、本テストの開発者であるS.ソレイシィと本学における改良の一部を担当した鈴木が、まずテストの開発理念と特色について、テスト理論や第二言語習得理論の諸概念を用い説明する。次に、実際に英語における発話率を測定する際の具体的規則を、機能言語学・談話研究等における知見をふまえつつ提案する。

キーワード: 外国語スピーキング・テスト、発話率、コンテクストに適した評価法

Second/Foreign Language Speaking Test, Speech Rate, Contextualized Assessment
Introduction

One of the youngest sub-fields of applied linguistics is speech testing (Alderson, Clapham & Steel, 1997). However, new speaking tests are becoming more and more mainstream (Soreni, 2006). Nothing emphasizes this point more than the mandatory inclusion of a speaking component on the new version of TOEFL in 2006. Taken by over 600,000 people worldwide, TOEFL’s reform sent ripples around the language testing world.

The testing organization which runs the TOEIC began offering a stand-alone speaking and writing test in December 2006 called TOEIC Speaking and Writing. Cambridge’s renowned assessment center has teamed with Japan’s STEP test (英语) organization to actively market the BULATS tests which feature a speaking component. Global English Language education brand names like Berlitz have produced a test which includes direct measures of speaking skills. Their parent company, Benesse, has spread the test, called G-TEC, to China, Taiwan, South Korea and of course Japan. The Chinese government is considering tests of productive skills (i.e. speaking & writing) as a basic requirement for graduation from university (Honna, 2006). A shift from testing receptive to productive skills is taking root in Asia.

Most educational institutions have been slow to catch up with this trend. Many continue to use “paper-based tests,” many of which test intake of target language forms and the knowledge of grammar or the receptive skills of reading and listening (Brown, 2000). They usually claim to indirectly measure communication skills and are thus unproblematically used for high stakes assessment such as entrance exams or to attain course credits or to graduate. Very few institutions have any assessment vehicles in place to directly test speech.

An ambitious assessment project is underway at Toyo Eiwa University that involves nine teachers who are implementing a new speaking test based on the “SPM Interview Test.” The testing is not affiliated with any institution and thus bears none of their traditional restraints, such as registration, venue, inappropriate content or level, or delayed results turnaround. In one in-class test, 139 students experienced the speaking test as part of their end of the semester exam. More regular testing is planned.

The speech prompts for this context can address both course content (e.g. “What do you think about non-native speakers’ English?” or “Tell me about the relationship between sports and international understanding”) and personal background topics (e.g., “Tell me
about your hometown” or “What’s your major?”).

The test asks students to orally make as many sentences as they can in thirty seconds about a topic, and then a rater counts the number of comprehensible sentences or clauses (See “Matching Scoring System with Scoring Aims”, pp. 19-21 of this paper, for discussion). The score is mainly based on this quantitative assessment (i.e. speech rate) with an overall qualitative assessment possible for both inferior speech (i.e. a series of utterances that do not address the topic) and superior speech (i.e. speech which is cohesive, well-developed and eloquent). It is important to note that this test focuses on a macro-linguistic, not micro-linguistic speech abilities. In other words, the test measures, rewards and thus encourages students to develop discourse-level speech abilities. In this paper, we illustrate one critical aspect of the speech test, its scoring system.

Receptive to Productive Testing Background and Context

The new wave of institutional standardized speaking tests (ISST) may be seen as an admission that “paper-based tests” can not properly measure oral production abilities. Cloze tests, for example, which require students to insert proper words into gaps in a text have been criticized for their reliance on the exam for the language input, thus failing to encourage any spontaneous production by the test-taker (Brown, 2000). While rather easy to implement and score, these tests send a message to students that they need knowledge of the language conventions rather than an ability to use it. The validity of claims of “equivalence” between indirect and direct tests of productive abilities and related research based on psychometrics has also been called into question (Spolsky, 1989, 1990).

It is very important to understand the teaching context and to explain how this testing may impact learning, as well as to consider the properly contextualized ELT solutions. Despite six years of formal English language education, first year students at university have difficulty producing English in spoken form. They may be considered “false beginners” (Grundy, 1994). In other words, they possess some declarative knowledge, but exhibit almost no procedural knowledge of the English language. They profess recognition of many English language forms and words, but very little ability to recall them, especially in spontaneous speech.

To more vividly demonstrate this, four spontaneous speech samples from in-coming
Freshmen students are presented below. These four women were placed in the top 15% of their peers according to their scores on a standardized achievement test administered in April, 2007. These interviews were conducted about one week after that testing, the first day of class at Toyo Eiwa University. All of the members of their class were given the in-class speaking test to assess their starting level of speech skill. The four samples selected from a group of 20 in their class are somewhat representative of the speech of the group. They were all asked “Tell me about your hometown” and given 30 seconds to answer.

**STUDENT A: STEP test equivalent/ 2nd Grade**

00:00

My hometown is near by sea.
And... n.....n....
The our town has very few shop.
N...., but ..... many restaurant.
00:30

**STUDENT B: STEP test equivalent/Pre 2nd Grade**

00:00

U...m...
My hometown is Yama.²
And eh?
Yama is nn ?
Eh ? wakannai [=I don't know], eh ?
Eh, eh, eh, eh, there is many people in Yama.
There are, a, there are many shop.
XXX [=unintelligible] ?, shop.
Ah, moh, dame dah [=I can't do this any more].

---

1 The speech samples and STEP test equivalents are shown here with permission of the students.
2 All names have been altered to protect the identity of the speakers.
Wakannai yo [= I don't know].
Nanio ieba iino [= what can I say]?
00:30"

STUDENT C: STEP test equivalent/Pre 2nd Grade

00:00
Uhm...
...(extensive silence)
My hometown is... eh..ta...Kawa, Fuji.
....
Fuji is u...u...u..... a famous...
Ahh...
00:30

STUDENT D: STEP test equivalent/Pre 2nd Grade

00:00
My hometown is Eno.
Ah... Eno is very urban.
And...many green Icho, Icho [=Ginkgo] trees.
And famous food is Sashimi and Ramen.
Famous...ahh...
00:30

The question for teachers is where would you go from here? We consider these students' speech samples typical of graduates of Japanese schools that produce students who have passed numerous paper-based tests. They have covered basic grammar points explicitly over and over. Obviously they lack the ability to perfectly apply that grammatical knowledge in spontaneous speech. However, by discussing grammar here, we would be ignoring the "elephant in the room."
We must focus on a much more profound and obvious “error”. It is an error that can be clearly seen on the discourse level: these students’ speech samples reveal an utter inability to sustain and develop their speech. The demand to do so is explicit on many ISST. It is also an implicit demand in a significant portion of international communication (Soresi 2005).

Obviously another common issue involved schema and the question of what to say, which would later be combined with the issue of how to express it. The specific solutions for overcoming these issues are not our main focus here. However, we hope for such discussion and to question whether these learners really benefit from ELT approaches which offer more English language exposure and ask them to input some more target English language forms.

Contextualizing English language teaching or methodology to these learners needs is a very valid discussion, but not the focus here. Contextualizing English language speech testing is our focus. Specifically for this essay, we would like to share one aspect of an in-class testing solution, the test scoring system. How tests score helps determine how students prepare, learn and ultimately perform.

**Proper Testing & Scoring for Our Teaching Context**

Even if we recognize the need to directly test speech, in-class testing is a formidable challenge. It is much easier to use paper-based tests. Do-it-yourself interview tests can directly measure students’ speech, but they carry risks. Even though we allot a few minutes to interview each student, we may go over time. We may lose track of the scoring scale that we planned to use, or struggle with objectivity in other ways. The struggle to score speech objectively is formidable, even for trained raters (Alderson, 1991; Derwing & Rossiter, 2004).

A testing solution must be very practical, objective, and meaningful. It can not be overstated that **there is no perfect test.** At Toyo Eiwa University we have adopted the SPM In-class Interview Test as a contextualized testing solution to meet our students’ needs. The test we implemented bears relevance to our particular educational context. It has been designed to address specific speech and larger language learning issues among our university students. In the limited space here, we would like to focus on the SPM-based
in-class test scoring. Test scoring determines the success of this test in several ways. Let's make it clear what we are aiming for in our test design and its scoring system.

First, since this is an in-class speaking test, we must maximize practicality. This is perhaps the most crucial testing feature for in-class speaking tests. Smooth scoring is one key to overall practicality. Considerable resources are used by ISST to record, then have a network of expert certified raters score speech according to a complex scale or rubric (Fulcher 2003). While widely used by ISST, that scoring system is not practical for in-class testing in most teaching contexts.

If scoring involves training to use complex rating scales and/or recording and then hours for tape reviewing and scoring, regular speech testing will be nearly impossible for the average English language instructor. Furthermore, scoring should be done on the spot, so that scoring turnaround and test performance feedback can be timely. If not, students will forget what they did right or wrong, or lose interest and teachers can not deliver final grades on time.

Next, our purpose for testing speech directly is to cause some constructive learning impact. The contents of the learning impact will be strongly influenced by how we score. After all, students will want to improve their scores. Therefore, it is important to establish a practical, meaningful and stable scoring system for our speaking test, which impacts the learners and learning in our context.

Finally, scoring needs to be applicable to the students' levels. Many students in Japanese universities have spontaneous speech skills which are so under-developed that they could not produce any speech spontaneously under ISST assessment conditions. ISST do not and can not customize their testing. SPM testing does and this article will show one example of a contextualized solution, the scoring.

Matching Scoring System with Scoring Aims

This paper offers a detailed definition of the output unit we use for in-class speech testing in the form of guidelines, parameters and examples for counting output units. The scoring system claims both practicality and positive learning impact, and a potentially larger "washback" in our teaching context. In the following sections we will carefully describe the SPM test's unique, yet useful scoring system.
SPM stands for "sentences per minute." What's new about this speaking test approach is that a **target speech rate range** serves as the main standard for performance. We've found that this leads to positive learning results among a wide range of students. The target speech rate range focuses students' cognitive load on output amount and rate. Speech rate is the frequency of output units relative to a time unit. Thus the number of comprehensible output units made in 30 seconds constitutes the students' raw scores. How utterances are segmented into output units is important to our scoring system. How we segment the utterances will serve as a model for how we want students to speak and develop spontaneous speech abilities.

Speech rate has typically been measured by counting either syllables per second/minute (Derwing & Rossiter, 2004; Kowal, Wiese, & O'Connell, 1983) or words per minute (Bell, 2003; Pimsleur, Hancock, & Furey, 1977; Tauroza & Allison, 1990). For our purpose, measuring speech rate as a form of both in-class testing and peer/self assessment, however, using either of these small units would be impractical because of the difficulty involved in counting: students would not be able to count syllables or words while listening to their classmate's speech, and still less while producing and monitoring their own speech simultaneously.

In this test, output units are tallied on the spot, not by reviewing a recording or transcript of the speech. This testing methodology, which allows for 'on the spot' calculation of results, has several advantages. First, it is more practical and takes less time. Secondly, raters can count output units while factoring in **prosody**, a vital aspect of speech. Thirdly, it allows for self-study and peer practice. In other words, since we aim our testing at **positively impacting learning**, scoring must be clearly understood by learners. By making scoring practical and clear to learners, they are able to prepare for the tests by actually speaking and then judging their own speech or their classmates. Finally, the testing’s clear standards take the form of informational feedback and can lead to some autonomous learning (Deci & Ryan, 2002).

For these reasons, we need to use larger units for measuring speech rate, such as clauses or sentences. Using those syntactic units has an additional pedagogical advantage: students will be constantly reminded of the necessity to construct their speech based on 'subject-verb' sequences. The importance of this effect cannot be overemphasized when Japanese university students' procedural knowledge of English grammar can be so inadequate that some have difficulty constructing grammatically acceptable sentences (Ono et al., 2005).
Traditionally, sentences have often been considered the most basic units of language. Greenbaum and Quirk (1990), for example, describe sentences as 'those language units which we must regard as primary, in comprising a minimum sense of completeness and unity[.]' (Greenbaum & Quirk, 1990, p. 12) On the other hand, those who view language from more functional rather than formal perspectives and attach more importance to speech in their discussion of language tend to regard clauses, rather than sentences, as the basic units (Butt, Fahey, Feez, Spinks, & Yallop, 2000; Halliday & Matthiessen, 2004). They consider sentences as a pattern of language that occurs in written texts and dismiss the concept of sentences being the basic units of speech by arguing that '[s]poken language is not divided into sentences, although we often think and talk about it in those terms ...' (Butt et al., 2000, p. 30, Emphasis in the original). Thus, it is clauses that are considered 'the ...fundamental grammatical unit[s]' (Halliday & Matthiessen, 2004, p. 16).

Of the two syntactic units, clauses and sentences, both of which are large enough for students to identify and count on the spot, we find clauses to be better suited for our purpose of assessment at Toyo Eiwa University. As our aim is to develop a practical test with positive learning impact, we measure speech rate using clauses as the smallest output unit. Counting clauses in speech is easier for students than counting syllables or words, and from our experience, can be done reasonably accurately by students of varying abilities in a Japanese university English course. However, from the viewpoints of test validity and fairness, there are some cases in which specific rules for counting clauses need to be articulated. Also in light of recent research on spoken discourse, we believe the definition of a 'clause' should be expanded to include not only syntactic, but also some other standards for our purpose. The following sections introduce the tentative rules we have developed for these reasons and some explanations of the rationale behind each of them.

How to identify an 'output unit' (in the SPM-based Speaking Test at Toyo Eiwa University)

As stated earlier, we base our assessment of speech rate on output units which generally correspond to syntactic 'clauses'. We also claim that identifying those output units can be done with reasonable accuracy by students and teachers. To understand this we must look beyond syntax and consider the vital role of prosody and pragmatics in speech. Recent research on conversation suggests that spoken language can be best understood as consisting
of interactional units which can be described in not only by syntactic, but also prosodic and pragmatic terms (Ford & Thompson, 1996; Furo, 2001; Ikeda, 2004). Ford and Thompson (1996), for example, report that speakers’ change in spoken English correlates with ‘Complex Transition Relevance Places’ where syntactic, prosodic and pragmatic completion points overlap. This suggests that English users commonly use such linguistic features to define units of spoken English. We have tried to incorporate such features into the following guidelines for counting ‘output units’ in order to reflect in our assessment system how English is actually spoken.

**Rule 1**

In general, a syntactically complete clause is counted as one output unit.

**Rule 1-1**

An independent clause (underlined in the following examples) is generally counted as one output unit.

**Examples** (_______ indicates the boundary of a segment to be counted as an output unit. AUT indicates an authentic example taken from a sample test.)

1. I have two sisters and a brother. _______ > 1 output unit
2. My father is a teacher, and my mother is a designer. _______ 2 output units
3. My sister likes Hip-hop, and actually I do, too. So we get along. _______ > 3 output units
4. (AUT) Prompt: What do you think of (Prime Minister) Shinzo Abe?
   Reply: I don’t like Mr. Abe. And I don’t watch news. _______ > 2 output units

**Rule 1-2**

A dependent clause (underlined in the following examples) is generally counted as one output unit.

**Examples (Adverbial clauses)**

5. I went to McDonald’s because I was hungry. _______ > 2 output units

---

3 A note on transcription: a comma indicates a continuing (non-final) intonation, while a period indicates a final intonation.
6. When I was 6, my parents bought me a piano. > 2 output units

7. I went to Paris last summer. I went there because my cousin lives there. > 3 output units

Adverbial clauses describe time, place, reason, or purpose for example. They are typically initiated by words or phrases such as 'when', 'where' 'because' or 'so that'.

Regarding example 7, one may wonder whether the second part ('I went there') should be assigned the same status as the other two parts, given the fact that it doesn't offer any new information. Although it may be semantically redundant, we treat this part as one output unit just like others since we are aiming students' speech abilities toward international communication and mutual intelligibility among various English varieties. Repeating a phrase may not be aesthetically pleasing but could contribute to better understanding in the context of international communication. That being said, interviewers in our tests are asked to mark down for blatant “stacking” of sentences that add no new information to the discourse or do not address the topic whatsoever. Raters reported only encountering two “borderline” stacking cases in about 140 interviews for a high stakes test in July, 2007. In 5 years of testing Japanese students, such stacking has not been a factor in testing so far, but needs to be watched as the stakes for tests rise.

Examples (Adjectival clauses)

8. I have a cousin who lives in Canada. > 2 output units

9. I have a ring my mother gave me. > 2 output units

An adjectival clause modifies a noun which precedes it. It is typically led by a relative pronoun, although our average students seldom able to compose relative pronouns in their speech, probably because of the structural complexity.

Rule 1-3

A clause with an elliptic subject (underlined in the following examples) is generally counted as one output unit. However, it may be counted together with another output unit depending on its prosodic features.
Examples
10. **I went home.** (pause) **And took a long shower.** > 2 output units
11. **I’ll go and see it.** (no pause before ‘and’) > 1 output unit

In actual testing, students and raters are expected to identify such prosodic boundaries rather intuitively based on subjective auditory perception. However, such perceptual judgments are known to be based on a range of features such as pauses and pitch ‘reset’ (Schuetze-Coburn, Shapley, & Weber, 1991), which can be also defined acoustically. Also, it has been reported that trained analysts can make consistent and reliable judgments about intonational boundaries (Ford & Thompson, 1996, p. 146). For these reasons, we believe it is reasonable to incorporate prosodic features in our assessment guidelines even though they are based on subjective perception.

**Contextualized Exception to Clause Rule (Basic nominal clauses)**
12. **I think (that) he was lying.** > 1 output unit
13. **He said (that) we must pay.** > 1 output unit
14. **I don’t know if she’s sick.** > 1 output unit

A nominal clause usually functions as a subject, object, or complement, without any of which a clause may be considered incomplete propositionally, and in some cases also structurally. Therefore, it could be argued that the nominal clause should be counted as part of the larger (main) clause which embeds it, rather than counted on its own. This is one item in our test scoring that might be subject to flux.

At this point in our students’ development, the nominal clauses made are almost always very simplistic ones. Specifically, students’ explicit use seems limited to two or three forms: ‘I think + S+V’, or ‘She/He said S+V’, or occasionally ‘He/She seems +S+V’. The most prevalent form by far is “I think +S+V” or in many cases a mistaken for example as “I think difficult.” in which they omit the subject verb before “difficult.” Here are more examples from Freshmen tested April 2007, “(It) seems (that) (the) teacher is very good” “I think (that) he looks like (a) bulldog”. The level of syntactic simplicity they exhibit in spontaneous speech has shown us that the nominal clauses which our learners make should be counted with the main clause as one output unit. They very rarely if ever use wh-words such as “I can’t decide what I should get for my sister’s 20th birthday.” This will be one area to watch as they progress.
In some cases, we might see how two separate output units could be counted by raters in our speaking test. For example, “My biggest problem at school this year by far is... whether I can get up early in every morning.” Even our most advanced students do not make these complex sentences at this point.

Rule 2

A word or phrase without a verb or verbal is generally not counted as one output unit. However, the following are exceptions to this rule.

Rule 2-1

A word or phrase without a verb or verbal is counted as one output unit when it is clear from the context that the speaker intended to produce a clause/sentence.

Examples

15. I live in Urayasu, Chiba. [Urayasu famous for Disneyland] > 2 output units
16. I don’t like Japan. [Too many people in small space] > 2 output units

While we admit that students should eventually be made aware of these grammatical ‘mistakes’, we nonetheless count these elements as ‘output units’ in consideration of the lack of expressiveness of our students and Japanese university students in general. In other words, we count “comprehensible” clauses and sentences. This approach considers the seriousness of errors or ‘error gravity’ (James, 1998). If the error prevents comprehensibility, it is not counted. On the other hand, we consistently and clearly penalize what is perhaps the most grave and pervasive speech “error” in Japanese high school graduates; a lack of speech, lack of discourse extension and development. This standard is closer to the “real world” standards which they will face. With our students’ development in mind, in particular, we want to reward any and all comprehensive or cohesive speech output units.

Rule 2-2

A word or phrase without a verb or verbal is counted as one output unit when it is conventionally used as an independent unit.
Examples

17. My sister likes Jazz. Me too. > 2 output units
18. So we met in Paris by chance. Small world. > 2 output units

Utterances such as those above are not lacking anything from the communicative point of view even though they do not contain a verb or verbal. Since it is the students’ communicative ability that we aim to assess, rather than the ability to produce grammatically fuller utterances, we treat these phrases as a full output unit.

Rule 2-3

A word or phrase without a verb or verbal is counted as one output unit when it constitutes a pragmatically complete element (e.g. an answer to a question).

Examples

19. Prompt: What’s your favorite food?
   Reply: **Onigiri**. I especially like Onigiri with fried shrimp. > 2 output units
20. (AUT) **Mr. Abe has many problems. For example, the pension problem and so on.** > 2 output units

While the word ‘Onigiri’ in example 19 is simply a single word, it is nonetheless a complete answer to the preceding question. In other words, it constitutes a pragmatically complete element in this ‘question-answer’ sequence, and therefore should be counted as a full output unit. Likewise, the segment initiated with ‘for example’ in example 20 is also pragmatically complete since in authentic conversations, examples are often introduced by simply listing nouns. Thus, this part should also be counted as one output unit.

Rule 3

A syntactically complete clause is NOT counted as one output unit in the following cases.
Rule 3-1

A syntactically complete clause is not counted as one output unit when it is judged to have a 'non-final' intonation and repaired directly afterwards.

Examples

21. So, I go. (continuing intonation) I went... I went there > 1 output unit
22. Mr. Abe is... many... He has many problem(s). > 1 output unit
23. I went to LA. (final intonation). No, I went to Seattle, first. > 2 output units

As with Rule 1-3, this intonational judgment is made perceptually but will be based on such features as pauses and pitch ‘reset’. It is important to note that these recasts seem to increase as SPM rises in students, pointing to the assertion that fluency actually leads to more accuracy (Nation, 1991).

Rule 3-2

A syntactically complete clause is not counted as one output unit when it is a filler and produced repeatedly.

24. Let me see. I have three telephones. > 2 output units
25. Let me see. let me see. I have three telephones. > 2 output units
26. Let me see. I have three telephones, and let me see. I have two PHS. > 3 output units

Fillers can serve a variety of interactional functions even though they may not carry much referential meaning (Schiffrin, 1994). Therefore, in principle, they should be given the same value as other semantically fuller clauses. However, some fillers such as ‘you know’, for example, are also known to occur repeatedly as a ‘speech habit’ (Ostman, 1981). When a test taker produces a particular filler expression repeatedly as a habit, counting all of such occurrences as full output units would not seem reasonable from the viewpoints of test validity and fairness. We thus only count the first occurrence of such a filler and discount the following occurrences.
Reflections

As can be clearly seen from actual speech samples taken from in-coming Freshmen, most high school graduates come to university with vast latent or implicit knowledge of English language, but extremely limited explicit knowledge exhibited in speech. The ELT context in Japanese universities dictates a contextual approach. Contextualized or "situated" ELT has overtaken the idea of universal "methods" and arisen as the new way to look at ELT in scholars who publish work in journals from TESOL Quarterly to Applied Linguistics to World Englishes (Kumaravadivelu, 2003, 2006). The SPM-based speech test we have adopted is a contextualized English language testing approach for average Japanese university students.

That has implications not only for our speech test scoring system and its definition of the output units, but also for the test design based on a target speech rate, as well as the very idea that we should go out of our way to implement an ambitious interview test. The reaction to the new interview test from students, teachers and administrators can be described as cautiously excited. Major concerns and even a threat to any speaking tests involve practicality first and foremost. We feel we have overcome this with a proper test design. However we feel it is necessary to address issues like test reliability and fairness by explaining our scoring system and defining the output unit used for the target speech rate range.

Our definition does not aim for some universal description of spontaneous speech, although we draw from scholars' description of "clauses" as the basic unit of such speech. Obviously, scoring by counting output units which may be considered comprehensible "sentences"/"clauses" risks some subjectivity. If we took ten minutes of spontaneous speech, for example, and asked 100 English language teachers to count how many sentences or clauses are uttered, we would rarely ever get the exact same total. There can be a reasonable margin of error expected.

In the current design we aim to reduce the margin of error and increase inter-rater reliability for the test by outlining what constitutes a "clause" (or "sentence" depending on the students' understanding of syntax) prior to the test. We also ask several questions (no less than four) to give students more chances to speak and reduce the margin of error. With these measures, we would expect multiple raters' score totals to be quite close, especially with less experienced, lower level speakers. In other words, our students' current
amount and rate of speech production is low enough to make counting output units relatively
straight-forward. Once speech becomes more voluminous and proficient, speech count
admittedly becomes less straight-forward and may require more rater training.

In conclusion we offer a before and after sampling of one student’s spontaneous speech
in reply to two similar prompts.

STUDENT E: STEP test equivalent/Pre 2nd Grade

APRIL (BEFORE the course began)
00:00
1. My hometown is near Ida.
2. My house ...n...my house,
3. Ah...there are many trees, xxx (unintelligible)s.
4. So, air is very clean.
5. My house is ....u.....m,

00:30

One error is the article omission in line three. The much larger problem is an inability
to develop or extend her speech. She fails to do so despite six years of ELT based mostly on
input of target language forms and their grammar and syntax.

JULY (AFTER a semester of English course incorporating the SPM-based test)
1. My home town is near Ida.
2. And my station name is Koida.
3. People don’t know my cit, city.
4. And there are many forest ah...tree, trees.
5. And the air is very clean.
6. So I love my town.
7. And ...my, my town is very...

As Nation (1991) convincingly argued, this learners’ accuracy actually increased with
fluency gains. She both has autonomously extended and developed her speech and fixed
her previous rudimentary grammar mistake (article omission) in line 5. Such development
makes a very strong case for a shift in ELT toward more output-based tasks, lessons and courses, which maximize student speaking time per student, especially in Japan where there is no domestic speech community.

Our SPM-based test, scoring system and definition of an output unit are carefully designed to meet a specific issue: too many Japanese universities are graduating students who can not sustain the most basic spontaneous speech in English. We hope that a truly contextualized speech test, like the SPM-based test at Toyo Eiwa University, can help students and educators realize this issue and lead to more constructive learning results among students of various backgrounds and abilities.

References


DISCOURSE: A COMPARISON OF ACOUSTIC AND AUDITORY ANALYSES.

—— (2005). Multicultural English Proficiency: A New Model for Global Considerations,
—— (2006). The New Wave of Speaking Tests. Invited Presentation at KOTESOL,
Sookmyung Women's University, Seoul, South Korea.
T. Stewart (Ed.), Classroom Practice Series on Speaking Skills. Arlington, Virginia:
Teachers of English to Speakers of Other Languages, Inc.

Linguistics, 10 (2).


(1), 90-105.