Changing Tasks … Changing Evidence: A Comparative Study of Two Speaking Proficiency Tests

India C. Plough[1], Fabiana MacMillan[1], and Stephen P. O'Connell[2]
[1]University of Michigan and [2]University of Maryland

1. Introduction

In 2005 the Testing and Certification Division of the English Language Institute at the University of Michigan began a major review of the Examination for the Certificate of Proficiency in English (ECPE) Speaking Test. As part of a larger test validation project, the study reported here compares examinee performance on two speaking tests: The former speaking test (henceforth referred to as the Old test) and the current, revised speaking test (henceforth referred to as the New test). After a brief description of the ECPE, the theoretical and empirical foundation for the format and design of the New test is explained. This is followed by the presentation of the study proper in the Methodology, Results, and Discussion sections.

The ECPE is a test of advanced English language proficiency, reflecting skills and content typically used in university or professional contexts. Receptive and productive skills (reading, listening, speaking, writing) are evaluated through a combination of integrated and discrete tasks. The ECPE is aimed at the highest level (C2, Proficient: Mastery) of the six-level Common European Framework of Reference (CEFR) (Council of Europe, 2001).

The purpose of the Old test was to assess an examinee's general ability to interact and communicate effectively in spoken English at a university or with speakers in the international business community. It can be classified as an interview between a single examiner and the examinee. The test takes about 10–15 minutes and consists of two parts: a warm-up and a second part that builds on topics that emerge from a discussion of a picture. Examples include a picture of a casino that might lead to a discussion of gambling or a picture of a family picnic that might lead to a discussion of inter-generational communication. The examiner rates the candidate immediately after the interview. The features that are assessed include: Fluency and Intelligibility; Grammar and Vocabulary; Functional Language Use and Sociolinguistic Proficiency; and Listening Comprehension. Candidates were rated on a 4-point scale from Good/Very Good Speaker (4) to Limited Speaker (1).

1.1. Construct

The revision process began with a thorough review of current research and a simultaneous review of the characteristics of the test taking population and the uses of the test. These reviews resulted in

[*] We gratefully acknowledge the assistance of the people who participated in this study, including the test-takers, center administrators and teachers at the Hellenic American University, Athens, Greece. We are also very grateful to Üte Roemer for her helpful insights on the analysis of lexical complexity and to Matthew Brook O'Donnell for designing a tool to manage our lexical data electronically. An earlier version of this paper was presented at the Second Language Research Forum at the University of Maryland, College Park, in 2010. We thank members of the audience for valuable comments. Finally, thanks are also due to Jayanti Banerjee and Natalie Nordby Chen for their committed collaboration. All shortcomings remain our own.

redefining the construct to be assessed. The New test is based on an interactionalist perspective of second language performance, which draws on applications of systemic functional theory to second language acquisition research and assessment. As Swain notes, within this framework, “performance is jointly constructed and distributed across the participants. Dialogues construct cognitive and strategic processes which in turn construct student performance, information which may be invaluable in validating inferences drawn from test scores” (Swain, 2001, p. 275).

The New test measures the ability to understand the linguistic (phonology, syntax, vocabulary), pragmatic (appropriacy, implicature), and sociolinguistic (situational, topical, cultural) information included in an interaction; to use this knowledge to engage in an extended interaction; and to produce extended samples of realistic spoken language spontaneously (Buck, 2001; Fulcher, 2003). Candidates should be able to “understand virtually everything heard; summarize information and reconstruct arguments in a coherent presentation; express [themselves] spontaneously, very fluently and precisely, differentiating and conveying finer shades of meaning even in more complex situations; hold their own in formal discussion of complex issues, putting forward an articulate and persuasive argument” (Council of Europe, 2001: 24, 78). This construct was then used to inform test format and task design.

In a change from the Old test's one-on-one (examiner: examinee) format, the New test uses a paired examinee format. Among the first researchers to argue against the one-on-one format were Van Lier (1989), Lazaraton (1996), and Young & Milanovic (1992). These researchers maintained that the power relations between the examiner and the examinee are inherently unequal and that this asymmetry distorts the language that is produced. Additionally, the unequal power relations seriously restrict the different language functions that can be probed in an assessment.

The issue of the effects of examiner discourse on examinee performance, which in turn affects examiner ratings, has been investigated in a series of studies on the ACTFL OPI by Ross (1992), Ross and Berwick (1992) and Berwick and Ross (1996). Focusing specifically on control (e.g., initiation of topics) and accommodation by examiners, results indicate that an examiner’s use of elements of control mainly serves an administrative function, regardless of examinee proficiency; however, a majority of final ratings could be predicted based on an examiner’s degree of accommodation. Brown (2003) has found that differences in the ways examiners provide feedback, structure talk, and formulate questions can influence examinee performance, which in turn affects the interviewer's perceptions of examinee proficiency. And most recently, Ross (2007) concluded that examiner differences (such as proclivity to backchannel) can lead to divergences in examinee performance.

There is a growing body of research supporting the paired format. In separate studies, Iwashita (1996), Nakatsuhara (2006), Brooks (2009), and Davis (2009) have investigated the effect of the proficiency of one’s speaking partner on a number of variables, including amount of talk, topic initiation, topic continuation, and assessment scores. Results indicate no significant difference in the conversational styles between same and different proficiency level pairings; additionally, it could not be concluded that one’s rating was affected by the proficiency level of one’s partner.

Further support for use of the paired format comes from the construct definition. As discussed, examinees should be able to carry the conversational burden. That is, examinees should be able to use linguistic resources to complete a given task with minimal support provided by an examiner. We maintain that assessment of this ability is best accomplished with the paired format.

1.2. Task design

The early work of Pica, Kanagy, and Falodun (1993) and Foster and Skehan (1996), the relatively more recent work of Skehan (2009), Foster and Tavakoli (2009) as well as that found in Bygate, Skehan and Swain (2001) and Fulcher (2003) was drawn on extensively in determining key task characteristics that may elicit those linguistic features from which meaningful inferences of speaking proficiency can be made.

Research indicates that tasks that demand decision-making and consensus-building while allowing for multiple outcomes are likely to generate more interaction than tasks that do not include a consensus-building element. For example, Swain (2001) examined the use of collaborative tasks in a French immersion classroom and concluded that these types of tasks ‘force’ participation by both students and result in more student output (less teacher talk). Additionally, open tasks (i.e., tasks in which relatively indeterminate or unrestricted information is provided to examinees) encourage individual expression, an important feature of learner-centered assessment (Chalhoub-Deville, 2001). Furthermore, requiring more differentiated decisions (that is, decisions must include the advantages and disadvantages of a particular choice) seems to increase linguistic complexity (Robinson 2001, 2005). Finally, the work of Tavakoli and Foster (2008) indicates that those tasks that provide clear, well-structured information allow for greater fluency.

Using a single written prompt, the content and structure of the New test is designed to allow examinees to demonstrate the full range of their speaking ability while performing a multi-stage, semi-structured task. Examinees work in pairs to complete a decision-making task. There are two examiners present during the entire test, which lasts between 25 and 35 minutes and consists of five Stages that require candidates to interact with each other and with one examiner (examiner 1 in Stages 1–3, and examiner 2 in stages 4–5). Throughout the majority of the test (Stages 2–4), the participation of the examiners is minimal; it does not extend beyond giving directions. The linguistic demands become increasingly more complex from Stage 1 to Stage 5 as the tasks become more difficult. This scaffolding provides candidates with opportunities to produce the range of language to be assessed by the ECPE.

1.3. Assessment

An examinee's linguistic ability is evaluated independently by the two examiners and separately from the other examinee’s ability. Immediately after the test, the two examiners reach a consensus on the scores to give to each candidate. Features assessed include: Discourse and Interaction (Development, Functional Range, and Listening Comprehension); Linguistic Resources (Range and Accuracy of Grammar and Vocabulary); and Delivery and Intelligibility. The scoring rubric consists of five bands from Expert Proficiency (A Level) to Limited Proficiency (E Level).

2. Current study

2.1. Purpose and research questions

As part of a larger test validation project, the purpose of this study was to examine the language (data) elicited by the new test design and format. A comparative methodology was employed to analyze test taker performance on the Old test and their performance on the New test. Specific research questions were:

1) Is there a significant difference in the complexity of language produced by an examinee on the Old ECPE Speaking Test and the New ECPE Speaking Test?

2) Is there a significant difference in the range of vocabulary produced by an examinee on the Old ECPE Speaking Test and the New ECPE Speaking Test?

3) Is there a significant difference in the range of linguistic functions produced by an examinee on the Old ECPE Speaking Test and the New ECPE Speaking Test?

3. Methodology

3.1. Test instruments

Table 1 summarizes the key task characteristics of the Old and New tests, which have already been discussed in the Introduction.

Table 1: Test instruments

	Old Test	**New Test**
Format	One: One	Two examinees and two examiners
Level	Advanced	Advanced
Duration	10–15 minutes	20–30 minutes
Design	Interview	Five Stages 1. Introductions 2. Summarizing and Recommending 3. Consensus-Reaching 4. Presenting and Convincing 5. Justifying and Defending

3.2. Participants

Within a two-week period, the same individuals sat for both the Old test and the New test. Data come from 23 New tests and 39 Old tests. Because not all of the examinees gave the researchers permission to use their tests, the total number of individuals is 39. Descriptive information, including age and gender is provided in Table 2. As shown, candidates ranged in age from 15 to 50 years old, with a median age of 22; 6 males and 33 females participated. The first language of all participants is Greek. Effort was made to recruit volunteers from a range of proficiency levels in order to determine how well the test discriminates between different levels, particularly at the cut score level (i.e., C on the CEFR).

Table 2: Participants

Participants	**Age Range**	**Gender**	**L1**

n = 39	15–50
	(median = 22)
		Male = 6
		Female = 33
			Greek

3.3. Data collection

Before the New test became operational, and one week before the last administration of the Old test, candidates were asked to take a “practice” test, which was actually a version of the New test. All administrations were video-recorded. Unfortunately, it was not possible to alternate the order in which examinees took the tests; that is, all examinees took the New test and then 1–1.5 weeks later they took the Old test.

3.4. Coding protocol

All tests were transcribed and each transcription was checked by one of the three researchers. All transcripts of both the Old and New tests were coded for analysis of speech units (AS-units), which were used to normalize the data in order to investigate linguistic complexity through measures of

Clause per AS-Unit and Sub-Clause per AS-Unit (research question 1), and for linguistic function (research question 3). Standardized type/token ratios were used to measure range of vocabulary (research question 2). The definitions of all features were agreed upon by the three researchers, each of whom coded for all features. These coded transcripts were then exchanged so that each transcript was checked by at least one other researcher. Detailed explanations of the definitions adopted are not presented here. However, selected features are covered in the Results section.

3.5. AS-units

Following Foster, Tonkyn, and Wigglesworth 2000, analysis of speech units were defined as “a single speaker’s utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either” (p. 365) and, as mentioned, were used to normalize the data. The AS-unit is syntactically-based but also integrates the use of intonation and pausing in determining AS-unit boundaries. Each transcript was coded for number of clauses, sub-clauses, false starts, repetitions, idiosyncratic fillers, unintelligible speech, and incoherent speech per AS-unit. Discussion is limited to clauses and sub-clauses, examples of which are shown in (1) and (2) below. AS-unit boundaries are marked by an upright slash (|), clause boundaries within an AS unit are marked by double colons (::), and sub-clausal units are marked by an upright slash followed by an ‘s’ (|s).

(1) Clausal unit:
S1: *I am a student* **|** *I have decided* **::** *to be a teacher* **|**
[2 AS-units, 3 clauses]

(2) Sub-clausal unit:
S1: *yes in primary school* |s
[1 AS-unit, 1 sub-clausal unit]

3.6. Vocabulary range

The vocabulary range in the transcribed speech of each examinee in both the Old and New tests was analyzed using the WordSmith 5.0 software package (Scott, 2008). Prior to generating word lists for individual examinees in each test, all instances of unintelligible and incoherent speech as well as first language use were eliminated from all transcripts. A standardized basis of 1,000 tokens (total number of running words) was used to normalize speech sample size, and values were obtained for the ratio of types (number of different word forms) and tokens produced by each examinee in the Old and New tests. The resulting Standardized Type/Token Ratios (STTRs) consist of average type/token ratios based on consecutive 1,000-word portions of each speech sample.

3.7. Functions

Transcripts of both Old and New tests were also coded for functions. Given the relative lack of previous work in which linguistic functions had been measured to any great extent, a grounded methodology was adopted; that is, a functional taxonomy was developed from the data. In order to reach consensus (and ensure coding reliability), functions were defined primarily based on syntax and lexicon. So, for example, in order for an utterance to be labeled as "comparing," an explicit marker of comparison, such as the word "more" or the comparative form of an adjective, had to present, or for an utterance to be labeled as "providing opinion," an explicit marker of opinion such as "I think” or “in my opinion” had to be present. Examples are provided in (3) and (4):

(3) Comparing
S1: *Denver eh has ah* ***more*** *facilities to offer*

(4) Providing opinion
S2: ***I think*** *that’s very important*

During coding, it became clear that examinees expressed functions in support of overarching functions, so to speak. In the end, there were instances in which functions were embedded to five levels. An example is provided in (5).

(5) Multiple functions in a single utterance (lexical indicators of functions shown in bold)

(Level 1) Defending
(Level 2) Question: Interactive move (request floor)
Oh can I add something?
(Level 2) Speculation
if *she organized person*
(Level 3) Opinion
I think *that's very important*
(Level 4) Explanation
because *ah ah once she is going to design a site maybe um we need to the picture that ah shows this ah ah this site we we want to be easy and um very approached to the others*
(Level 5) Emphasis
so I think that her organized skills are ***very*** *important that she has this*

Note that functions in a single utterance were not double-counted. That is, the "if" in Level 2 was coded as an indicator of the function of speculating. The use of "maybe" in the first line of Level 4, Explanation, was not coded as yet another instance of speculation. Similarly, as the use of "very" in Level 5 was coded as an indicator of emphasis, the use of "very" in Level 3, Opinion, was not coded as an additional instance of emphasizing. Finally, the use of "I think" in Level 3 was coded as indicator of expressing opinion; therefore, the use of "I think" in Level 5 was not also coded as expressing opinion.

4. Results

Quantitative analyses conducted to address the first two research questions are presented first. A paired sample t-test was conducted to compare measures of linguistic complexity in the Old test and the New test. Results are presented in Tables 3–6 below. As shown in Table 3, the mean of clauses per AS-unit produced on the New test and the mean of clauses per AS-unit produced on the Old test are equivalent. Similarly the mean of sub-clauses per AS-unit produced on the New test and the mean of sub-clauses produced on the Old test are roughly equivalent.

Table 3: Paired Samples Descriptive Statistics for Clauses per AS-Units

		Mean	n	SD	Std. Error Mean
Pair 1	**Clause/AS** **New**	1.60	39	0.34	0.06
	Clause/AS **Old**	1.61	39	0.25	0.04
Pair 2	**Sub-Clause/AS** **New**	0.26	39	0.08	0.01
	Sub-Clause/AS **Old**	0.24	39	0.10	0.02

As shown in Table 4, there is no significant difference between the mean of clauses per AS-unit on the New test and the production of this feature on the old test. Similarly, there is not a significant difference between the mean of sub-clauses per AS-unit on the New test and the production of this feature on the Old test. Thus, the response to the first research question – Is there a significant difference in the complexity of language produced by a candidate on the Old ECPE Speaking Test and the New ECPE Speaking Test? – is no.

Table 4: Paired Samples T-Test for Clauses per AS-Units

	Paired Differences							
	Mean	**SD**	**Std. Error Mean**	**95% Confidence Interval of the Difference**		**t**	**df**	**Sig.** **(2-tailed)**
				Lower	**Upper**			
Pair 1 **Clause/AS** **New-Old**	-0.01	0.36	0.06	-0.13	0.10	-0.22	38	.826
Pair 2 **Sub-Clause/AS** **New-Old**	0.03	0.11	0.02	-0.01	0.07	1.65	38	.108

Turning now to lexical diversity, as shown in Table 5, the mean of the standardized type/token ratios (STTRs) in the speech produced by examinees on the New test is only slightly higher than the mean of the STTRs observed on the Old test.

Table 5: Paired Samples Descriptive Statistics for Standardized Type/Token Ratios

	Mean	n	SD	Std. Error Mean
Pair 1				
Old STTR	28.82	39	2.40	.384
New STTR	29.05	39	3.42	.548

As shown in Table 6, there is no significant difference between the mean of the standardized type/token ratio on the New test and the mean of this ratio on the Old test. Thus, the response to the second research question — Is there a significant difference in the range of vocabulary produced by a candidate on the Old ECPE Speaking Test and the New ECPE Speaking Test? — is also no.

Table 6: Paired Samples T-Test for Standardized Type/Token Ratios

	Paired Differences							
	Mean	SD	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
Pair 1 **Old STTR** **New STTR**	-.232	3.533	.566	-1.377	.913	-.410	38	.684

Things become a bit more interesting when we examine linguistic functions. Keeping in mind that the data have not been normalized for time on task, it is possible to compare the relative frequencies of the production of linguistic functions produced on each test. The commentary presented here focuses only on those functions that were produced 100 or more times. As shown in Table 7 the range of different functions is greater in the New test (34 versus 21). It should be pointed out that this difference actually becomes greater if the functions are not conflated. For example, Interactive Moves have been combined into a single category; when separated, there are only two different kinds of interactive moves in the Old test (offer floor, initiate dialogue) in contrast to five different kinds in the New test (take floor, offer floor, request floor, request information, offer opinion). If all functions are similarly separated, the totals become 27 for the Old test and 50 for the New, almost twice as many. Thus, the totals listed in Table 7 are quite conservative. The response to the third research question — Is there a significant difference in the range of linguistic functions produced by a candidate on the Old ECPE Speaking Test and the New ECPE Speaking Test? — is yes.

Table 7: Different Functions Produced on Old and New Test

Old Test	21 Different Functions
New Test	34 Different Functions

If one extends the qualitative comparison to include a closer examination of the specific functions produced on each test, one could argue that, overall, the kinds of functions produced on the New test are more challenging than those produced on the Old test. As seen in Table 8, those functions that were used relatively frequently in the Old test include: explanation (334, 21%), general information (145, 9%), opinion (326, 20%), personal information (293, 18%), and hypothesis/speculation (131, 8%). The language elicited reflects the nature of the test, which is primarily an interview composed of questions asking for an opinion or a speculation.

*Table 8: Function Use by Stage in the Old Speaking Test**

Function	**Stage 1**	**Stage 2**	**Stage 3**	**Total (Frequency)**
Comparison	16	2	50	68 (4%)
Contrast	13	0	41	54 (3%)
Description	0	37	9	46
Emphasis	7	0	15	22
Evaluative comment	27	2	22	51
Explanation	98	11	225	334 (21%)
Example	5	1	59	65
General information	35	2	108	145 (9%)
Hypothesis/Speculation	9	26	96	131 (8%)
Interactive move (offer floor, initiate dialogue)	1	0	3	4
New supporting information	0	0	9	9
Opinion	51	12	263	326 (20%)
Organization of discourse	6	0	16	22
Offering solution	0	0	3	3
Paraphrasing	1	0	0	1
Personal information	176	1	116	293 (18%)
Preference	0	0	1	1
Providing clarification of information	0	0	1	1
Request (clarification of information/ question/ language; repetition)	12	0	25	37
Recommending	0	0	1	1
Softener	1	0	5	6
Total				1620

*margin for error on totals +/- 4

As shown in the function totals for the New test in Table 9, with the exception of providing general information and hypothesis/speculation, those functions that were produced relatively frequently in the Old test are also used to a great extent in the New test: explanation (284, 11%), opinion (351, 14%), and personal information (138, 5%). This indicates that the format and design of the New test also provides test-takers with opportunities to express and explain their opinions and to share personal information. However, unlike the Old test, the New test elicits a broader range of functions that may be considered more challenging, such as Negotiation (273; 11%), Presenting (105, 4%), and Summarizing (163, 6%). Also of note is the difference in test-taker use of discourse markers of organization. These markers were used more frequently in the New test (119, 5%) than in the Old test (22, 1%). In terms of test validation, this seems to be another indicator that the test design and format of the New test ‘pushes’ test takers to organize their speech in a way that the interview format of the Old test does not. Finally, it should be noted that the use of these functions indeed corresponds with the stage that was designed to elicit a particular function: 159 of the total 163 instances of summarizing are produced in Stage 2 (Summarizing and Recommending), 271 of the total 273 instances of negotiating are produced in Stage 3 (Negotiation); and, finally, all instances (105) of presenting are produced in Stage 4 (Presenting).

*Table 9: Function Use by Stage in the New Speaking Test**

Function	**Stage 1**	**Stage 2**	**Stage 3**	**Stage 4**	**Stage 5**	**Total (Frequency)**
Acknowledgement (issue, opinion)	0	0	9	0	2	11
Agreement	0	10	63	0	16	89
Comparison	10	51	37	4	6	108 (4%)
Concession	0	3	10	1	11	25
Contrast	12	45	35	2	17	111 (4%)
Confirmation (of information, task)	0	14	1	3	0	18
Convincing	0	0	2	29	0	31
Disagreement	0	4	15	0	17	36
Emphasis	2	13	27	18	17	77
Evaluative comment	25	28	16	10	4	83
Explanation	44	74	74	46	46	284 (11%)
Example	5	4	8	3	4	24
General information	24	0	2	0	1	27
Hedge/Softener	0	6	31	6	19	62
Hypothesis/Speculation	9	27	14	6	25	81
Interactive move (offer [to take] floor, request floor, opinion, information)	2	21	39	4	2	68
Justification	0	0	0	0	44	44
Negotiation	0	2	271	0	0	273 (11%)
New supporting information	0	1	4	10	16	31
Opinion	17	135	109	31	59	351 (14%)
Organization of discourse	2	63	12	39	3	119 (5%)
Offering solution	0	2	4	1	12	19
Paraphrasing	1	10	4	0	1	16
Personal information	137	0	0	1	0	138 (5%)
Preference	0	3	0	1	0	4
Presenting	0	0	0	105	0	105 (4%)
Providing (clarification of information; content; task)	0	8	10	0	1	19
Request (clarification of information/ task/ question; repetition; language)	18	42	9	6	11	86
Recommending	0	39	0	2	0	41
Repetition	0	1	1	0	0	2
Resolution	0	0	0	0	2	2
Rhetorical question	0	0	2	0	0	2
Sentence completion	0	5	4	0	2	11
Summarizing	0	159	4	0	0	163 (6%)
Total						2561

*margin for error on totals +/- 4

While a thorough examination of the stage-by-stage totals is not provided here, a brief look at one example highlights the difference in the two tests in terms of the effects of scaffolding to elicit qualitatively different functions. In the New test, personal information is produced almost exclusively (137 of 138 instances) in Stage 1 (Introductions and Small Talk). In contrast, in the Old test, even though the majority of instances (176 of 293) are also produced at the beginning of the test (Stage 1: Warm Up), providing personal information is also elicited to a great extent in Stage 3, where another 116 instances are observed. That is, it appears that the Old test does not allow test takers the opportunity to move beyond what may be considered 'easier' communicative functions (i.e., providing personal information).

As explained in the Methodology, functions were categorized in terms of number of levels of embedding. Table 10 presents the totals of each of the different levels produced on the Old test and on the New test.

Table 10: Embedded Functions Old and New Tests

	Old Test		**New Test**	
	Instances	**% of Total**	**Instances**	**% of Total**
1-level	273	29.6	348	25.9
2-level	584	63.3	785	58.4
3-level	61	6.6	191	14.2
4-level	4	0.4	16	1.2
5-level	0	0	4	0.3

As can be seen, one- through four-level functions were produced on the Old test. In contrast, one- through five-level functions were produced on the New test. Additionally, the frequency of utterances that contained embedded functions is higher in the New test. Three-level and four-level functions comprise 15.4% of the total instances of embedded functions on the New test, whereas they make up only 7% of the functions in the Old test. Initial analysis indicates that the differences in levels of embedding are indicative of topic development and elaboration.

5. Discussion

The Discussion is organized around two main questions:

1) First, what, if anything, does the comparison of the language produced by examinees in the two tests tell us about our testing instruments employed to elicit language?

2) Second, what, if anything, does the comparison of the language produced by examinees in the two tests tell us about the measures employed to analyze that language?

At the end of this section the larger issue of test validity is addressed and a summary is provided of what, if anything, the preliminary results of this study reveal about the nature of the evidence required to make valid interpretations of the behavior identified (or predetermined) as key.

The results of the analyses of syntactic complexity and of the lexical diversity indicate that there is no significant difference in examinee performance on the Old and the New tests. In response to our first question regarding our testing instrument, this finding is relatively uninformative. That is, based on an analysis of examinee performance, one might be led to conclude that there is no significant difference in test format, that the New test was unsuccessful in eliciting the language to be measured as defined in the construct. However, given the notable and indisputable differences in the design and format of the two tests—and what research has shown us regarding the effects of task on learner performance—it would be premature to draw this conclusion. Additionally, the analysis of the communicative functions used by examinees in the Old test and the New test clearly indicate that test design and format have an impact on the quality and quantity of functions that are elicited.

Examining the measures employed for syntactic complexity and lexical diversity may help interpret the findings of non-significance in these two areas. Throughout a recent special issue of *Applied Linguistics* (Norris & Ortega, 2009) the authors cautioned researchers of the limitations of using measures of fluency, complexity, and accuracy in isolation. The use of general measures alone has proven insufficient in that they do not provide a complete representation of learner production. For example, Robinson, Cadierno and Shirai (2009) re-visited a study investigating the effects of task complexity in which general measures of syntactic complexity were used to examine learner data. The initial study reported no significant effect; however, re-examining the data using specific measures revealed that there indeed is more complex use of tense-aspect morphology with an increase in task complexity. Results of the current study certainly support the arguments for supplemental measures of syntactic complexity.

It has also been suggested (Arnaud, 1984; Richards, 1987; Vermeer, 2000) that the type-token ratio is a limited resource when it comes to assessing lexical diversity. In fact, in the aforementioned issue of *Applied Linguistics,* Skehan (2009) calls for the necessity to incorporate text-external measures of lexical sophistication into current measures of complexity, accuracy, and fluency measures of language production. Although the current study addressed limitations of type-token ratios by normalizing speech samples, results do not provide information about the frequency of individual words in general language use as informed by spoken corpora. Further analysis is needed to investigate the level of sophistication of vocabulary produced by examinees in the New test, which might provide a clearer picture of the possible effects of test design on this specific aspect of examinees' performance.

At this point, limitations of the current study must be noted. The primary caveats center around methodological issues. In terms of data analysis, time on task must be normalized so that a more thorough quantitative comparison of the linguistic functions produced on the Old test and on the New test can be performed. In terms of data coding, while the transcriptions were checked by at least two researchers, the length of pauses in examinee speech (essential for the AS-unit measure) and the number of levels of particular functions must still be confirmed. Additionally, in terms of data collection, the order of test administration was the same for all examinees; and, the context was exclusively a testing situation.

With respect to the participants of the study, one could argue that the fact that the L1 is the same and that the majority of participants are female is an advantage in terms of controlling for these variables. Recall that an attempt was made to recruit a range of proficiency levels. However, the majority of participants actually performed in a very narrow band at the lower end of the 'high intermediate' (C-level in CEFR) range. This certainly has implications on a scoring rubric as fine distinctions between levels in the linguistic areas to be scored cannot be determined if participants are all approximately at the same level. Within the context of the current study, the fact that participants were relatively homogenous in terms of proficiency level may have played a role in the results of syntactic complexity and lexical range. That is, it could be that these features are more informative or 'sensitive' to proficiency level rather than effects of test format and design, and that they actually are critical indicators of differences among learners at different levels of proficiency (Norris and Ortega, 2009). At the same time, however, it is not necessary for a measure "to vary across subjects [in order to be valid], but it must adequately represent its underlying construct" (Pallotti, 2009, p. 591). It should be noted that grammatical and lexical range and accuracy are indeed central to the construct of the New ECPE Speaking test and also integrated into the scoring rubric. Several final caveats relate to the need for further analyses of the data. Measures of accuracy and fluency must be completed; and, all measures have to be examined with respect to learner proficiency. Not only must the stage-by-stage function totals be examined more thoroughly, but it may also be informative to examine production of functions by individual learners. It appears that the production of specific functions is not idiosyncratic of one or two examinees, but this must be confirmed empirically.

In the light of these limitations, one can still conclude that the New test elicits that language that was identified as key in our definition of the speaking construct. Recall that the test is based on an interactionalist perspective of second language performance and that the test is intended to measure the ability to produce the discourse of, for example, compare, contrast, negotiation, and summary. Preliminary analysis of the communicative functions indicates that the paired format and multi-stage, semi-structured task of the New test, unlike the one-on-one interview format of the Old test, does indeed provide learners with the opportunity to show their ability to produce this discourse. Results also support the claims of other researchers that measures of syntactic complexity must be supplemented. We suggest that measures of linguistic functions, particularly if lexically- and grammatically-based, are promising.

References

Arnaud, Pierre J.L. (1984) The lexical richness of L2 written productions and the validity of vocabulary tests. In Culhane, Terry, Klein-Braley, Christine, and Stevenson, Douglas K. (Eds.) *Practice and problems in language testing. Occasional Papers* 29, 14-28.

Berwick, Richard and Ross, Steven. (1996) Cross-cultural pragmatics in oral proficiency interview strategies. In Milanovic, Michael and Savilles, Nick (Eds.) *Performance testing, cognition and assessment: Selected papers from the 15th LTRC.* Cambridge: Cambridge University Press, 34-54.

Brooks, Lindsay. (2009) Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing,* 26(3), 341-366.

Brown, Annie. (2003) Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1-25.

Brown, Annie and Lumley, Tom. (1997) Interviewer variability in specific-purpose language performance tests. In Huhta, Ari, Kohonen, Viljo, Kurki-Suonio, Liisa and Luoma, Sari (Eds.) *Current developments and alternatives in language assessment*. Jyväskyla: Centre for Applied Language Studies, University of Jyväskyla, 137-150.

Bygate, Martin, Skehan, Peter and Swain, Merrill. (Eds.) (2001) *Researching pedagogic tasks: second language learning, teaching and testing*. Harlow, England: Longman.

Chalhoub-Deville, Micheline. (2001) Task-based assessment: a link to second language instruction. In Bygate, Martin, Skehan, Peter and Swain, Merrill. (Eds.) *Researching pedagogic tasks: second language learning, teaching and testing*. Harlow, England: Longman, 210-228.

Council of Europe. (2001) *Common European Framework of reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Davis, Larry. (2009) The influence of interlocutor proficiency in a paired oral assessment. *Language Testing,* 26(3), 367-396.

Foster, Pauline and Skehan, Peter. (1996) The influence of planning and task type on second language performance. *Studies in Second Language Acquisition* 18, 299-323.

Foster, Pauline and Tavakoli, Parvaneh. (2009) Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning* 59(4), 866-896.

Foster, Pauline, Tonkyn, Alan and Wigglesworth, Gillian. (2000) Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21(3), 354-375.

Iwashita, Noriko. (1996) The validity of the paired interview format in oral performance assessment. *Melbourne Papers in Language Testing* 5(2), 51-66.

Nakatsuhara, Fumiyo. (2006) The impact of proficiency-level on conversational styles in paired speaking tests, *University of Cambridge ESOL: Research Notes*, 25, 15-20.

Norris, John and Ortega, Lourdes. (2009) Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4), 555-578.

Pica, Teresa, Kanagy, Ruth and Falodun, Joseph. (1993) Choosing and using communication tasks for second language instruction. In Crookes, Graham and Gass, Susan M. *Tasks and Language Learning: Integrating Theory and Practice*. Clevedon: Multilingual Matters, 9-34.

Pallotti, Gabriele. (2009) CAF: Defining, Refining and Differentiating Constructs. *Applied Linguistics* 30(4), 590-601.

Richards, Brian J. (1987) Type/token ratios: What do they really tell us? *Journal of Child Language* 14, 201-209.

Robinson, Peter. (2001) Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics* 22(1), 27-57.

Robinson, Peter. (2005) Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL* 43(1), 1-32.

Robinson, Peter, Cadierno, Teresa and Shirai, Yasuhiro. (2009) Time and motion: Measuring effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics* 30(4), 533-554.

Ross, Steven. (1992) Accommodative questions in oral proficiency interviews. *Language Testing* 9, 173-186.

Ross, Steven. (2007) A comparative task-in-interaction analysis of OPI backsliding. *Journal of Pragmatics*. 39(11), 2017 – 2044.

Ross, Steven and Berwick, Richard. (1992) The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14(2), 159-176.

Skehan, Peter. (2009) Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics* 30(4), 510-532.

Skehan, Peter. (2001) Tasks and language performance assessment. In Bygate, Martin, Skehan, Peter and Swain, Merrill. (Eds.) *Researching pedagogic tasks: second language learning, teaching and testing*. Harlow, England: Longman, 168-185.

Swain, Merrill. (2001) Integrating language and content teaching through collaborative tasks. *The Canadian Modern Language Review*. 58(1), 44 – 63.

Tavakoli, Parvaneh and Foster, Pauline. (2008) Task design and second language performance: The effect of narrative type on leaner output. *Language Learning* 58(2), 439-473.

Van Lier, Leo. (1989) Reeling, writhing, drawling, stretching, and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly*. 23(3), 489 – 508.

Vermeer, Anne. (2000) Coming to grips with lexical richness in spontaneous speech data. *Language Testing* 17(1), 65-83.

Selected Proceedings of the 2010 Second Language Research Forum: Reconsidering SLA Research, Dimensions, and Directions

edited by Gisela Granena, Joel Koeth, Sunyoung Lee-Ellis, Anna Lukyanchenko, Goretti Prieto Botana, and Elizabeth Rhoades

Cascadilla Proceedings Project Somerville, MA 2011

Copyright information

ISBN 978-1-57473-448-5 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Plough, India C., Fabiana MacMillan, and Stephen P. O'Connell. 2011. Changing Tasks … Changing Evidence: A Comparative Study of Two Speaking Proficiency Tests. In *Selected Proceedings of the 2010 Second Language Research Forum*, ed. Gisela Granena et al., 91-104. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2618.