

## **Development of a test of speaking proficiency in multiple languages**

India C. Plough

Residential College in the Arts and Humanities, Michigan State University

The Residential College in the Arts and Humanities (RCAH) at Michigan State University has a foreign language proficiency graduation requirement. The RCAH has found it necessary to revise its language proficiency program and to develop a local test of language proficiency in lieu of using existing, internationally-recognised assessments of speaking proficiency. Situated within Critical Language Testing (Shohamy, 2001a, 2001b), the paper presents motivations for this decision reached after a yearlong program review. Treating the processes of teaching, learning, and assessment as interdependent, the RCAH's new Cultures and Languages Across the Curriculum program and the new performance-based proficiency test are built on the same methodological principles. Grounded in a social interactional theory of second language acquisition and assessment, the RCAH Test employs a paired format and is intended to assess intermediate speaking proficiency in the more commonly taught and the less commonly taught languages. Initial trials have been conducted with native speakers of English, and native and non-native speakers of French, German, and Spanish. Using discourse analytic methods, preliminary analyses highlight the potential influence of sociocultural context and bring into question the importance of syntactic complexity in the conceptualisation of speaking proficiency.

**Key words:** speaking proficiency, cultures and languages across the curriculum

## Introduction

Proficiency in a language other than English is one of the degree requirements of the Residential College in the Arts and Humanities (RCAH) at Michigan State University (MSU). RCAH students have been taking the Oral Proficiency Interview (OPI), the Simulated Oral Proficiency Interview (SOPI), or the shortened form of the SOPI to fulfill this requirement. After a review of its language proficiency program, the RCAH has found it necessary to replace these tests with a locally created assessment of speaking proficiency. The significant body of research on the OPI and the SOPI, both well-respected, internationally recognised tests of foreign language (FL) speaking proficiency, is not reviewed here as a critique of these tests is not the objective. Rather, the main purpose of this paper is threefold. First, the motivations behind the RCAH's decision to revise its language proficiency program and to take on the role of test developer are discussed by way of an overview of pertinent results from a yearlong program review. This discussion is situated within the democratic assessment practices articulated in Critical Language Testing (CLT) (Shohamy, 2001a, 2001b). Second, the RCAH's new proficiency program and test are summarised. The college treats the processes of teaching, learning, and assessment as interdependent and symbiotic (National Research Council, 2001). Therefore, methodological consistency is maintained throughout the new program (a unique Cultures and Languages Across the Curriculum [CLAC] model) and the performance-based speaking test. The underlying methodological principles are highlighted. Finally, preliminary findings from the first test trials of the RCAH Test of Speaking Proficiency are shared. These have been conducted with native speakers (NS) of English, and native and non-native speakers (NNSs) of French, German, and Spanish. Discourse analytic methods are used for initial analyses of test trial transcripts and feedback provided by examiners and students. Findings highlight the influence of sociocultural context and bring into question the relevance of syntactic complexity in the conceptualisation of speaking proficiency.

### National context

Before discussing the program review and in order to appreciate the relative significance of the RCAH's decision, two distinguishing characteristics of foreign language education in the United States must be briefly described: (a) the integral role of the federal government; and (b) the increasing influence of standards-based instruction and assessment. Dating back to the 1950s, the federal government has taken a prominent position in the development of FL assessment (Fulcher, 2003; Herzog, 2007) and continues to play a key role in the design and orchestration of FL instruction and assessment, often providing

critical financial support to areas of FL teaching and learning in K-16 education (Hudson, 2012). At the post-secondary level, programs of the International and Foreign Language Education Office (within the Office of Post-Secondary Education) provide institutional and fellowship grants; the Foreign Language and Area Studies (FLAS) program provides fellowships to institutions of higher education to assist undergraduate and graduate students learning a modern foreign language and related area studies; Title VI of the Higher Education Act supports international business education, undergraduate international studies, research and materials preparation, and overseas teacher training and conferencing. Title VI also provides support for National Foreign Language Resource Centers, which support K-12 foreign language education and teacher preparation.

Despite federal initiatives to improve the level of foreign language proficiency and to increase the number of foreign languages acquired by the people of the United States, results of the 2006 and 2008 General Social Survey conducted by the Joint National Committee for Language-National Council for Language and International Studies (Rivers & Robinson, 2012) indicate that the proportion of speakers of languages other than English in the US has remained relatively unchanged at approximately 25% for almost three decades (Eddy, 1980). Additionally, the overwhelming majority (90%) who self-report speaking a foreign language 'very well' learned that language at home and not in school. Finally, the proportion of college students enrolling in foreign language courses has remained at about 8% since 1977 (Furman, Goldberg, & Lusin, 2008, as cited in Rivers & Robinson, 2012).

The second distinguishing characteristic of FL education in the US is the increasing influence of the Standards for Foreign Language Learning: Preparing for the 21st Century in K-16 education, which is not unrelated to the first as the government has been instrumental in the creation of the Standards. In fact, the Standards are frequently used as measures of success for federally-funded programs. Magnan, Murphy, Sahakyan, and Kim (2012) note that the Standards have had a significant impact on teacher educators for K-12 levels (Byrd, Hlas, Watzke, & Valencia, 2011), have made their way into the curriculum of Less Commonly Taught Languages (LCTL), and have had a 'massive influence...on world language teachers' professional development, methods courses, and classroom curriculum' (Magnan et al., 2012, p. 171). Highlighting the absence of the student perspective in Standards-related research, the authors investigated the alignment between the goals and expectations of 42,000 college students and the Standards. Two findings are relevant to the current discussion. First, when the goals of students were compared to their expectations, the researchers found that students did not expect that FL programs would allow them to

achieve their goals in 11 of the 12 Standards. The authors speculate that a possible repercussion of this is that 'students might be discouraged to start language study believing that their aspirations will not be completely met' (Magnan et al., 2012, p. 185). Second, Magnan et al. (2012) concluded that while students' goals overall seem to align with the Standards, the importance students placed on individual Standards did not match the attention those Standards are given in curricula. Specifically, Communities Standards, placed last (fifth) by educators, but was ranked first by students. Additionally, students ranked Communication Standards second but it would have ranked first if only Interpersonal and Interpretative modes of Communication are considered, and not Presentational mode. One of the questions posed by the authors is: To what extent should curricula reflect the goals of students or should they reflect the 'goals that teachers might perceive as necessary?' (Magnan et al., 2012, p. 185).

### **Local context**

Within this national context, the faculty, staff, and students of the Residential College in the Arts and Humanities remain committed to creating a signature program that makes the study and use of world languages an essential part of the RCAH experience. The RCAH is a young undergraduate college, accepting its first students in 2007. With an enrollment of approximately 300 students, the major is built on four cornerstones: world history, art and culture, ethics, and engaged learning. As previously mentioned, the RCAH degree requires proficiency in a language other than English. Importantly, foreign languages are not formally taught nor has proficiency been tested in the RCAH; FL instruction and standardised assessment (i.e., OPI or SOPI) are housed in various departments across the University. To meet the language proficiency graduation requirement, RCAH students must

- complete two RCAH Language and Culture courses; and
- complete one 300 level (i.e., third year) foreign language course; or
- pass the OPI or the SOPI; or
- participate in an immersive study abroad/study away, which may be approved as a substitute for courses.

### **Program review**

The college's mission and the democratic assessment practices of Critical Language Testing (CLT) (Shohamy, 2001a, 2001b) guided the review of the language proficiency program. Specifically, as is explained in this section, all stakeholders collaborated in the review, which included an examination of test

uses and test consequences. As is discussed in the next section, Program Revision, findings led to the implementation of three additional CLT practices. First, the RCAH has assumed responsibility for the development and administration of its own test of proficiency; second, the knowledge and perspectives of diverse populations are incorporated into the design of the test and accompanying rubric; and, third, the rights of test-takers are respected by providing students with alternative forms of assessment.

### **Background**

The RCAH's language proficiency program consists of material resources as well as personnel. Established at the inception of the college, the Language and Media Center makes available a multitude of resources (e.g., current print media, language software). Students are also able to take advantage of the language resources (e.g., language clubs, foreign language TV) provided by other units on campus. Within the RCAH, graduate students from other colleges have served as language mentors, leading informal language tables in both the more commonly taught languages and the LCTLs (e.g., Japanese, Thai). A tenure system faculty member serves as director, overseeing the entire program.

As mentioned, FL classroom teaching within the University is not coordinated centrally but is managed through various units across the university. That is, each FL department functions relatively independently and is responsible for its own curriculum and classroom testing. The RCAH's language proficiency program was created to offer additional, optional support that would supplement classroom instruction by providing material resources and personnel as distinct from involvement in FL curriculum, instruction, and (standardised) proficiency assessment.

### **Collaboration among stakeholders**

The goals of the review that are particularly relevant to the current discussion were to obtain

- stakeholder feedback on the RCAH's language proficiency requirement and the language proficiency program;
- stakeholder feedback on the language proficiency test, including the testing process; and
- information on stakeholder use of test scores.

Input from students, alumni and alumnae, and faculty and staff (in RCAH and in FL units across campus) was gathered by means of surveys, semi-structured, one-on-one interviews, and focus group discussions. Analyses of language

resources available to students, language textbooks currently used in FL classrooms, of enrollment trends in language courses, and of student performance on the OPI or the SOPI were completed.

As shown in Table 1, 179 students completed a survey designed to gain insight into students' opinions of the proficiency requirement and the proficiency program in general. This represented 70% (179 of 255) of the student population enrolled Spring 2013. Special effort was made to reach out to as many freshmen and sophomores as possible as these are the students who arguably will be the most affected by any changes and should thus have a voice in shaping the revised program and assessment.

**Table 1.** Student respondents

Class level	n	% of enrolled	% of total respondents
Freshmen	68	100	38
Sophomore	35	63	20
Junior	31	56	17
Senior	45	59	25
TOTALS	179	70	100

All of the faculty interviewed support the language proficiency requirement. As indicated in Table 2, 75% (132 of 176) of the students also support the requirement. This response was used as the point of departure in post-survey interviews. When asked why the requirement was important, responses were equally divided between reasons of social/civic obligation, employment in a globalised economy, personal enrichment, and cognitive development. It is mainly for practical or affective reasons that students are resistant to the requirement. For example, students who have dual majors and/or are active in extra-curricular organisations find it difficult to incorporate language study into their schedules, particularly if they want to graduate in four years. Additionally, students commented that the requirement is 'very intimidating', 'a little difficult', and 'a little unnerving.'

**Table 2.** Language proficiency requirement

Class level	Keep requirement				Totals
	Yes	No	Unsure	No response	
Freshmen	51	10	4	3	68
Sophomore	29	3	3	0	35
Junior	23	6	2	0	31
Senior	29	15	1	0	45
Total	132	34	10	3	179

Interestingly, faculty and students noted that neither coursework nor passing the test would necessarily constitute proficiency, defining proficiency (and thus the intent of the requirement) as the ability to communicate and negotiate effectively and meaningfully. This feedback as well as the affective variables that seem to be impeding student engagement in language learning led directly to the design of the new program and the new proficiency test.

Seemingly contradictory to the belief that knowledge of a foreign language is important and that the RCAH should keep the FL proficiency requirement, students are not pursuing language learning. Fewer than 50% of the students who participated in the survey were enrolled in foreign language courses, and of those, more than 50% were at the beginning to low-intermediate levels (100 or 200 level courses). This is consistent with overall trends in foreign language enrollment of RCAH students from Fall 2007 through Spring 2013. Each fall and spring semesters during that six-year time period, the percentage of RCAH students enrolled in a foreign language course never exceeded 61%, with a median of 48%. Enrollment in 100 and 200 level courses ranged from 50% to 84%, with a median of 70%.

Students indicated that they had not taken FL courses because of a lack of confidence in the effectiveness of FL courses or the unavailability of certain courses (i.e., specific language courses were not offered or enrollment was full). With respect to language resources, the majority of the students (82%) are aware of the resources, but only 32% make use of them. All students reported that they need guidance in using the resources and that they would like more language support to be offered within the RCAH, such as one-on-one tutoring and integration of a language component into the college's civic engagement courses.

### **Examination of test consequences**

The OPI and the SOPI were chosen when the RCAH was established because of documented test validity and the fact that the scores are recognised internationally. During the initial discussions of the RCAH proficiency requirement, all faculty and administrators (i.e., RCAH, FL departments, and central administration) agreed that it was a priority for students to graduate from the RCAH with a proficiency credential that would be accepted and understood by organisations and institutions worldwide. Unfortunately, after six years, the overall sentiment toward the current form of testing is negative. Students offer three main reasons: (a) lack of authenticity of either a phone-based (OPI) or a tape-mediated (SOPI) test; (b) the rigidity of the actual content

of the tests; and (c) difficulty in scheduling. In other words, it would seem that the OPI and the SOPI have lost face validity among our student population.

Table 3 lists the total number of pass scores from Fall 2007 to Fall 2012. As can be seen, 57% (103 of 181) of students passed the test on the first attempt. The five-year summary of scores shows that the pass rate was 59% (140 of 237), which suggests that students are unprepared for the test. One can speculate as to the reasons for this poor performance. For example, perhaps students are taking the test too soon and should complete more coursework prior to testing; or, perhaps the loss in face validity is having a significant adverse effect on student performance.

**Table 3.** Total number of pass scores (Fall 2007 through October 2012)

	Pass	No Pass	Total Tests
First Tests	103	78	181
Retest	32	17	49
Second Retest	5	2	7
Total	140	97	237

### Test score use

In Spring 2013 a survey was sent to all RCAH graduates (n=145) to determine if alumni/ae were using the FL they had studied and if so, for what purpose; graduates were also asked if they used their OPI or SOPI score. Results are shown in Table 4. Response rate was 41%. Of the 60 respondents, the majority (62%; 37 of 60) are using the foreign language they studied. The majority of these students (62%; 23 of 32) use the language for professional reasons, either solely or in combination with another purpose or purposes. Fewer than half of the students (46%; 17 of 37) have used their test score, including it on resumes or on graduate school applications. No graduates reported that a test score was required for employment or graduate school admission. Even though the sample size is extremely small, it is not unreasonable to conclude that the high stakes nature of the test is much greater within the local context as part of the RCAH graduation requirement than it is as a credential for employment or future study.

**Table 4.** FL Language use and score use.

	Language Only	Language & Score	Score Only	TOTAL
Professional	3	5	3	11
Personal	7	0	0	7
Academic	3	1	0	4
Professional & Personal	5	4	0	9
Personal & Academic	2	1	0	3
Professional & Academic	0	2	0	2



---

Professional, Personal, & Academic	0	1	0	1
Total	20	14	3	37

---

To address the issues that were discovered as a result of the program review, the RCAH has adopted a comprehensive strategy, one that includes not only assessment but also teaching and learning. There was an apparent misalignment of the purpose for language learning, the form of assessment, stakeholder values, and the RCAH's approach to teaching and learning, which has been characterised as engaged pedagogy. This misalignment may have resulted, in part, in the language program slowly becoming a test-driven one. If students arrived on campus having just completed advanced FL courses in high school, some would take the OPI or the SOPI, pass it and then never take another FL course; if they did not pass the test, some would continue to take it until passing it (perhaps without enrolling in a FL course). This is in direct opposition to the goal, shared by the majority of RCAH stakeholders, for the study and use of world languages to be woven throughout a student's educational career. The goal has been overshadowed by a concentration on meeting yet one more requirement in order to graduate.

Four findings from the program review that are particularly relevant should be highlighted. First and foremost, there is support among all stakeholders for the RCAH's commitment to language proficiency, noting reasons of social obligation, personal enrichment, cognitive development, and employment in a globalised economy. Second, all students would like more language support and they would like that support to be offered within the RCAH. Our student population chose a small residential college because of the individualised attention that that environment affords. Third, students voiced a lack of confidence in the effectiveness of traditional language courses and criticism of the current form of language assessment. Finally, the majority of students do not use or need internationally recognised test scores. These findings were used to make two relatively significant programmatic revisions, one of which was the decision that the RCAH must take entire responsibility for proficiency assessment, a role previously performed by another unit on campus. The other was to create a Cultures and Languages Across the Curriculum program within the RCAH.

## Program revision

### Cultures and Languages Across the Curriculum (CLAC)

It is necessary to keep in mind that in the absence of meaningful engagement in language learning, student performance is never going to be at the required or expected level of proficiency. As noted, students were not pursuing, either formally or informally, ways to improve their proficiency. Research into models of language support offered in non-language units in higher education indicated that the Cultures and Languages Across the Curriculum (CLAC) model has potential as a viable framework to meet the programmatic needs of the RCAH. 'Within this large framework, CLAC can take many forms, depending on specific content and curricular goals within a discipline' (<http://claconsortium.org>). The description of CLAC provided on the consortium website reads, in part:

Cultures and Languages Across the Curriculum builds upon this basic idea: Knowledge exists within and is shaped by culture and, therefore, just as materials in many languages can and should be incorporated into all parts of the curriculum, intercultural perspectives can and should inform the teaching of academic content in many curricular contexts. The program works closely with the longstanding Languages Across the Curriculum, striving to make translingual and transcultural competence a reality for all students, not simply for those who major in a foreign language or participate in immersive study abroad programs. CLAC engages languages...to achieve a better and more multifaceted understanding of content. It focuses less on bringing disciplinary content or culture into the language classroom than on assimilating languages and cultures into instruction and research across a wide range of disciplinary and interdisciplinary context.

The incorporation of a CLAC program into the RCAH curriculum provides students with a guided, relatively structured, supplemental route leading to language proficiency. The program is formed in such a way that students receive additional support while concurrently enrolled in language courses; it also allows them to continue their engagement in language learning activities during those 'gap' semesters when a particular course may not be available. Most important, a CLAC program responds to students' desire, voiced in the program review, to have more language support offered within the RCAH.

Integrated Language Options (ILOs), organised around a single issue or theme, constitute the core elements of the program. Following the RCAH's engaged pedagogical approach, the topic of an ILO is student-initiated, stemming from a question that originates in an RCAH course. Students work collaboratively, with other students and/or community partners, to investigate an issue of their

choosing. ILOs consist of a series of related, sequential tasks that lead to a final performance-based project. ILOs are woven throughout the curriculum and, thus, throughout a student's career in the RCAH.

The CLAC program is being implemented simultaneously with the development of the RCAH Test of Speaking Proficiency. Program requirements now include multiple forms of assessment that are more closely aligned with the goals and values of the RCAH. That is, in addition to successfully completing course work in traditional FL classes, students may: a) participate in a specified number of ILOs, b) create a language portfolio that consists of self-selected projects that were completed in their ILOs, and/or c) take the newly developed RCAH Test of Speaking Proficiency.

### **RCAH Test of Speaking Proficiency**

The RCAH Test of Speaking Proficiency is based on information obtained from stakeholders during the program review, current research in speaking assessment, and the engaged pedagogical practices of the RCAH. It is also grounded in a social interactional theory of learning and assessment, which maintains '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 RCAH test is designed to measure language ranging from the upper beginner to upper intermediate level, very broadly speaking. In terms of the Common European Framework of Reference (CEFR), this is language within the A2+ (Basic User) to B2 (Independent User) levels. Language use at these levels ranges from the ability to 'manage simple, routine interactions and exchange ideas and information on familiar topics [to the ability] to discuss hypothetical situations and explain opinions' (Council of Europe, 2001, p. 20). The RCAH test is not designed to measure any specific curricular content.

The design and content of the test are aligned with the teaching and learning philosophy of the RCAH, are based on the definition of the speaking construct provided by stakeholders, and on the future intended uses of the foreign language. Consistent with an engaged pedagogical approach, which encourages student-centered, collaborative work, the test employs a paired format and consists of three different tasks. As shown in Table 5, after a brief introduction (Task 1), which serves as a warm-up, students co-present (Task 2) on a topic that they have selected and prepared in advance. The topic is one that they have investigated in an ILO and with which they are, therefore, quite familiar. The preparation for the presentation takes place well before the students sit for the

test, and there are no restrictions on the length of time students can prepare. This is followed by an examiner-led discussion of the topic among the students and the examiners. The purpose of the discussion is not to test students' content knowledge of the topic they have chosen; rather, the discussion serves as an extemporaneous speech event that provides students with an opportunity to show the range and level of their proficiency. Similar to many of the problem-based activities found in the engaged pedagogy of the RCAH, the test ends with a social scenario (Task 3), an authentic decision-making task in which only the students participate. This is followed by an examiner-led discussion of the decision among the students and the examiners. The test lasts approximately 25-35 minutes.

**Table 5.** RCAH Test of Speaking Proficiency

Task	Description	Condition	Time
1	Introductions		3-4 minutes
2	Joint presentation and discussion	Student-selected prepared topic Examiner led	6-7 minutes 5-6 minutes
3	Social decision-making and discussion	Set topic Examiner led	6-8 minutes 5-7 minutes
Total			25-35 minutes

Based on the growing body of task-based research (Brown & Hill, 2008; Bygate, Skehan, & Swain, 2001; Foster & Skehan, 1996; Foster & Tavakoli, 2009; Fulcher, 2003; Keck, Iberri-Shea, Tracy-Ventura, & Wa-Mbaleka, 2006; Skehan, 2009), the RCAH speaking test incorporates task characteristics that studies have indicated are instrumental in supporting interlocutors by eliciting interactional features from which meaningful inferences of proficiency can be made. Specifically, tasks that require learners to reach an agreement on a decision while allowing for multiple outcomes are likely to generate more interaction than tasks that do not include a consensus-building element. Additionally, open tasks (i.e., tasks in which relatively indeterminate or unrestricted information is provided to students) encourage individual expression, an important feature of learner-centered assessment (Chalhoub-Deville, 2001). Furthermore, requiring students to state the pros and cons of a decision seems to promote 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.

The test is also informed by the ever-growing body of empirical and theoretical research supporting the use of the paired format for assessments of speaking proficiency. The early work of Van Lier (1989), Young and Milanovic (1992), and Lazaraton (1996) argued that the inherent power differential between

interviewer and candidate distorts and limits the linguistic production of students in the one interviewer: one candidate format. This set the stage for studies focusing on the effects of examiner discourse on rater perceptions of candidate performance and thus the ratings given to candidates. For example, a number of studies investigating the OPI (Berwick & Ross, 1996; Ross, 1992; Ross & Berwick, 1992) concluded that a majority of final ratings could be predicted based on the degree of examiner accommodation during the test. In this same vein, Brown (2003) discovered that differences in the ways examiners provide feedback, structure talk, and formulate questions could influence candidate performance, which in turn affects the interviewer's perceptions of candidate proficiency.

The use of the paired format is now widely accepted and there is an active research agenda in the testing community to describe and define the interaction that is constructed as well as to investigate the possible effects of one's testing partner on the co-construction of discourse. Galaczi (2008) has contributed to our understanding of the construct in the identification of four general patterns (collaborative, parallel, asymmetric, blended) of interaction realised during a two-way collaborative task. Focusing on the interactional features attended to by raters, Ducasse and Brown (2009) found that rater commentary on student interaction could be grouped into three general categories: non-verbal interpersonal communication, interactive listening, and interactional management. Investigating the possible effects of the proficiency of one's partner on variables such as amount of talk, topic initiation, and topic continuation, Nakatsuhara (2006), Brooks (2009), and Davis (2009) could not conclude that the proficiency level of one's partner affected one's rating. Building on this work, May (2009, 2011) has investigated 'the separability of individual candidate's contributions to an asymmetric interaction' and suggests that interactional competence might be more accurately reflected in rating scales with the incorporation of 'shared scores for interactional effectiveness' (May, 2009, p. 419).

### *Rubric*

Using an intuitive (experiential) method, validated, operational scales and proficiency descriptors (e.g., ACTFL [American Council on the Teaching of Foreign Languages, 2012], CEFR, ILR [Interagency Language Roundtable, 2014], TOEFL iBT [ETS, 2014]) have been referenced to create a preliminary scoring rubric (see Fulcher, 2003). Categories and features are: Interaction (communicative functions; contribution to topic development; facilitative discourse behavior), grammar (accuracy and complexity), vocabulary (accuracy and range), and pronunciation. During test trials, examiners provide feedback

on these features and also assist with revising and creating level descriptors; in this way, empirical development of the scale has begun.

### **Test adaptation**

The RCAH Test of Speaking Proficiency was first developed, trialed and revised in English with RCAH students. This was done in order to get direct feedback on the test from the test taking population and to observe the effectiveness of test tasks. Even though the intended test population remains the same, for purposes of test comparability, it is essential to determine the extent to which the linguistic and cultural differences that exist among the different languages that the test is intended to assess might affect and/or be affected by the test format, task designs, the rating descriptors, and/or test examiners who are native speakers of those languages. The International Test Commission (ITC) Guidelines for Test Adaptation (Hambleton, Merenda, & Spielberger, 2005) were consulted in developing RCAH tests in French, German, and Spanish. ITC provides a series of 22 guidelines in four categories (Context, Test Development and Adaptation, Administration, and Documentation/Score Interpretations), which ongoing test validation must address. Of relevance to initial test development is ITC Guideline D.1:

the expertise and experience of translators are perhaps the most crucial aspects of the entire process of adapting tests as they can significantly affect the reliability and validity of the test (Bracken & Barona, 1991 as cited in Hambleton et al., 2005)...Because a single translator cannot be expected to have all of the required qualities and brings a single perspective to the task of translation, in general, it seems clear that a team of specialists is needed to accomplish an accurate adaptation. (Hambleton et al., p. 24)

Table 6 provides demographic and academic information of each language team. In total, there were 13 members on 3 different translation teams. First languages (L1) listed are self-reported. In addition to translating the test and the instructions provided by examiners during the test, team members discuss the sociocultural appropriateness of test format and design, specific content of Task 3 (the social scenario), as well as rubric development. All but four of the translators are trained as RCAH test examiners. This allows for revisions to be informed by first-hand experience with test administration.

**Table 6.** Test translation teams

	Number of members	L1	Country of Origin	Area of study
French	5 (F)	French	France	French language & literature
		Arabic & French	Tunisia	Engineering
		French		French literature
		Romanian & English	Romania	French history; French language & literature
German	4 (3 F; 1 M)	English (2)	US	
		German	Germany	Education
Spanish	4 (1 F; 3 M)	English (3)	US	German language & literature (2); education
		Spanish (3)	Colombia, Honduras, Venezuela	Education, cultural anthropology, second language studies
		English	US	Second language studies

### Connecting pedagogy and performance-based assessment

The connections between the RCAH's CLAC program and the different forms of proficiency assessment adopted by the RCAH warrant highlighting. The program and the assessments are built on the assumption that the processes of teaching, learning, and assessment are interdependent and symbiotic. As noted by the National Research Council (2001), teaching and assessment must be informed by a model of learning; assessment must be aligned with learning objectives and this alignment is better accomplished if teaching and assessment are actually conducted in the same way. Both the ILOs and the RCAH Test of Speaking Proficiency are based on similar methodological principles (Long, 2011): both create opportunities for collaboration; both respect diverse ways of knowing; both provide authentic input; and both cultivate learning by doing (see Chalhoub-Deville, 2001).

More specifically, in both an ILO and an RCAH test, students collaborate to accomplish a project or task; in fact, collaboration is required for task completion. Students choose the ILO issue they will investigate just as they choose their testing partner and the topic that they will co-present in the test; that is, learner autonomy is respected. Additionally, Task 2 (co-presentation) of the test mirrors RCAH classrooms where student-led discussions/presentations are the rule rather than the exception. ILOs are based on real-life issues (e.g., poverty, transitional justice) that students are studying in an RCAH content course; similarly, Task 3 of the test (the social scenario) is placed within the context of the RCAH and represents an authentic situation in which students could, or even have, found themselves (e.g., deciding which visiting artist to invite to the RCAH). By design, the interactional and experiential nature of the

ILOs and the RCAH test create opportunities for students to learn by doing. Throughout the test, students are engaging in purposeful uses of language to complete a series of tasks that model future communicative events in which they will participate. Students are collaborating in order to accomplish these tasks that are embedded in a larger project (that is, the testing event). Finally, increasing learner awareness of aspects of their test performance is crucial. Through the richness of the feedback provided, learning by testing is achieved.

## Test trials

### Students

Following Foster and Tavakoli's (2009) recommendation to establish a native-speaker baseline in order to disentangle second language (L2) processing constraints that may be influencing performance from other influences (e.g., task characteristics, the interaction of task and interlocutors), the RCAH test is being trialed with native-speakers of each language in which the test will be administered.

English trials were conducted with RCAH students. Test trials in French, German, and Spanish are ongoing with undergraduate students of the target languages who have completed at least one 300-level (third year) FL course. RCAH students are not permitted to take FL trial versions of the test as this would create complications once the test becomes operational and students might opt to take it as partial fulfillment of the proficiency requirement. Descriptive information of trial participants is provided in Table 7. A total of 30 students have participated in 15 tests to-date. Students range in age from 18 to 22; 23 females and 7 males have participated. First languages listed are self-reported. All tests were video and audio-recorded.

**Table 7.** Test Trial Participants

Test	Participants	Age	L1	Country of Origin
English	10 (7 F; 3 M)	18-22	English	US
French	8 (F)	20-21	2 French	France
		18-21	6 English	US
German	8 (4 F; 4 M)	18-23	6 English	US
Spanish	4 (F)	21,22	2 Spanish (Mexican)	US
		19,22	1 Chinese; 1 French	US

### Coding protocol

All tests are currently being transcribed and then translated into English; one other person checks each transcription and translation. Initial coding of original



(i.e., L1) transcriptions is underway for syntactic complexity, communicative functions, and active listening strategies, defined as follows:

- Syntactic Complexity. Analysis of Speech Units, which are 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' (Foster, Tonkyn, & Wigglesworth, 2000, p. 365);
- Range of Communicative Functions. Using Plough, MacMillan, and O'Connell (2011) as a starting point, a functional taxonomy is being developed from the data; and
- Active listening strategies, indicated through backchannel cues, confirmation checks, clarification requests, supportive sentence completions (Ducasse & Brown, 2009; May, 2011).

Video recordings are being reviewed for non-verbal facilitative discourse behavior. Definitions of all features are agreed upon by all researchers (i.e., two for each language, in addition to the primary investigator). Each researcher is coding for all features. Transcripts and notes from video reviews are then exchanged so that one other researcher checks each coding.

Additionally, videotapes of the tests are being analysed for facilitative discourse behavior as evidenced through non-verbal behaviors, including head nods, body posture/position and eye contact (Ducasse & Brown, 2009).

### **Analysis**

As mentioned, data collection and coding are still underway. Therefore, for the purposes of this paper, the discussion focuses on critical feedback offered by participants (examiners and students); observations from initial reviews of video recordings; and preliminary findings of linguistic complexity.

Feedback sessions are conducted with students after each test trial. Among the questions asked are 'Do you think the test is fair?' and 'Do you think the test allowed you to show your level of proficiency?' to which all students thus far have responded in the affirmative. To the question 'what did you like best?' all students said that they liked testing with a friend; most said they liked that they got to choose the topic of the presentation (Task 2); finally, several students testing in different FLs said that they liked 'how the test wasn't really structured like the usual speaking test.' When asked to elaborate, students described tests in which teachers ask a series of relatively unrelated, de-contextualised questions (e.g., 'What's your favorite movie?' 'What are your career plans?'). Based on this initial feedback from students, it appears that the test has face validity.

Of key interest to the researchers were responses to the question 'What would you change?' A number of responses focused on the content of Task 3. Two students taking the FL test in German said that the topic of Task 3, (social scenario), should be 'more personal.' As examples, they said they would like to be given choices of different vacation destinations and they would need to agree on a single vacation spot. Because these tend to be topics that are primarily for beginning level students, their use will be limited. A number of RCAH students stated that they would like to discuss more controversial topics (e.g., similar to ethical issues addressed in many RCAH courses); however, the remaining students said that they would feel too uncomfortable discussing certain issues in a test. Based on fairness and sensitivity standards, certain topics will continue to be excluded from test content. Students are, however, free to select the topic that they co-present in Task 2. Of particular interest was a comment on the structure of the test. The native speakers of Spanish asked if there were ways to build in opportunities for the students to ask the examiners questions, a feature that may be incorporated into the format of the test and into the pre-test instructions given to students.

Interesting variations in body position during the collaboration portion of Task 3 have been observed in preliminary reviews of the videos. What was particularly noticeable is if students did not face each other, one student of the pair did not face the other, or as in the case of the NS Spanish pair, the students simultaneously stood up and changed the orientation of their chairs to be facing each other, where they stayed for the remainder of the test. The importance of nonverbal behavior on examiners' perceptions of candidate proficiency has become increasingly evident in recent studies focusing on the interactional features attended to by raters. This takes on added significance in the current context of the development of a valid assessment of speaking proficiency in multiple languages. Results of Jenkins and Parra (2003) 'indicated that those candidates who employed active listening behavior considered appropriate by *North American* evaluators, such as frequent eye contact, smiling, forward leaning, head nodding, and back channeling were able to compensate for weaker linguistic proficiency' (Plough & Bogart, 2008, p. 199) [emphasis added]. Ducasse and Brown (2009, p. 437) report that 'raters found body language, or non-verbal behavior, to be a contributing feature to the success or lack thereof of interpersonal interaction'; one rater commented that 'the girl...uses her hands when she talks. It gives a nice color and is more in tune with the *Latin American* speech and culture' (p. 434) [emphasis added]. As concisely articulated by May (2009, p. 417),

...other factors related to personality and culture might determine, or at least contribute to, the pattern(s) of interaction that emerge...That both body

language and demonstrating assertiveness through communication were salient to raters in determining the effectiveness of a candidate's interaction may be of concern, in that these characteristics could be seen as aspects of culture, and L1 usage.

Given that nonverbal behaviors seem to be integral features of interactional competence attended to by raters, these must be included to a greater extent into programs of language study. To contribute to our understanding of nonverbal behavior and its implications, the literature within the fields of intercultural communication and behavioral psychology is currently being explored. Preliminary reviews indicate that there may be certain 'universal' patterns of behavior (e.g., status-related body posture/movements) that only vary from culture to culture in terms of the magnitude of the expression (Aguinis, Simonsen, & Pierce, 1998; Kowner & Wiseman, 2003; Mast & Hall, 2004; Mignault & Chaudhuri, 2003).

Turning now to Analysis of Speech Units in Task 3 of the test, generalisations, of course, cannot be made based on the limited number of students examined thus far. Nonetheless, when combined with verbal feedback from participants, two related observations of the data become more meaningful. First, as shown in Table 8, the median AS Units produced by the native speakers of English is greater than the top range of AS Units produced by the native speakers of French. Second, the lower and the upper ends of the range of AS Units for non-native speakers of French are higher than those for the native speakers of French.

**Table 8.** Analysis of Speech Units, Task 3

	English	French
Native Speaker	2.3 – 5.14 (median 4.2; n=10)	1.59 – 3.93 (n=2)
Non-Native Speaker	NA	2.2 – 4.3 (median 2.8; n=6)

Comments made by one of the NS French examiners during discussions of the rubric descriptors may be informative in analyzing these initial data: 'Using more complex grammatical structures isn't natural. It's forced. The students might use them if I set an example, but I wouldn't normally use those structures in everyday conversations.'

A number of questions immediately arise as a result of the examiner's comment. First, might it be the case that the AS Unit totals (or syntactic complexity) for both the French NS and French NNS students are a result of the interaction with this particular examiner? Research on the effects of examiner

discourse indicates that this may certainly be the case. Indeed, the examiner's AS Units for Task 3 range from 2.4 to 3.3, which is well below the range for the English NS students. However, if we assume that this style is not entirely an individual idiosyncrasy, one must ask to what extent this definition and measure of complexity is tied to a particular socio-cultural norm of discourse.

Additional insight may be gained by looking at comments made by the Spanish examiners. In the following excerpt, the examiner refers to the language of the student as exhibiting 'a very western organization,' which, based on the examples he provides, may be equated with syntactic complexity. In the excerpt, Juan is the second examiner and Estrella is one of the students, who, to repeat, stated that Spanish is her first language. NS Spanish Examiner:

We wouldn't do that. We just say what we have to say. There are...some subtleties in the way in which they [i.e., students] said some of their sentences but it is hard to point them out since they are not grammatically incorrect but simply not the way a native speaker would typically say them. For instance, Juan asks Estrella if she had the choice of traveling somewhere other than France where would she go and she answers "I don't know. I think somewhere where they have a really nice beach" or "No sé, pienso que alguna parte donde esté la playa bonita." [We] would probably say something like "algún lugar que tenga playas bonitas." Or "some place with nice beaches".

Similar to the comments made by the French examiner, if we assume that the Spanish examiners' interpretation of the discourse is not simply personal preference, how might the students' apparently non-native speech be explained? Perhaps the sociocultural context – both the immediate context of voluntary participation in a testing situation as well as the students' position in the larger context of immersion in a dominant L2 educational setting – are influencing the L1 discourse of the students.

Somewhat related, one of the students who is a native speaker of English commented during the feedback session that he felt the RCAH test involved real communication and that '[a]ll the kids in my high school knew that all we had to do was throw in a couple sentences with the subjunctive and we'd get a high score on the AP [Advanced Placement] Spanish Test'. In other words, it seems that there is a preconception of a specific formula for passing the test, regardless of actual speaking proficiency. Whether or not this is an accurate perception, is, in some regards, secondary. That is, it is revealing that getting a high score is thought to be the result of producing complex grammatical structures, and not the result of, for example, engaging in active listening strategies.

A word of caution must be repeated given the preliminary nature of the current investigation. However, taken together, initial observations – examiner commentary questioning the ‘naturalness’ of grammatical complexity or the appropriateness of ‘western organization’ and the somewhat disparate AS Unit totals between the native speakers of English and the native speakers of French – may suggest that a re-examination of the place for syntactic complexity in descriptors of FL speaking proficiency is warranted; specifically, the assumption that increased complexity is directly related to higher proficiency comes into question. Grammatical accuracy, complexity, and/or range are standard descriptors on scales for well-respected, validated tests of speaking proficiency. Research is broadening our understanding of interactional competence with the discovery of more and more features attended to by examiners when evaluating the discourse co-constructed in paired speaking tests. It may be that what examiners are not attending to is of equal importance. Recall that the RCAH test examiners are trained in the administration of the test and not in the use of a specific rating scale. Of note is the fact that none are trained examiners in any international test of language proficiency, while all have experience as candidates taking those tests to show English proficiency for admission to US universities and for applications for graduate assistantships. Rather than a disadvantage, it can be argued that their status as novice examiners is actually advantageous – they come to the process with minimal preconceptions of what descriptors of speaking proficiency should be. While perhaps an extreme interpretation, the current investigation may indicate that syntactic complexity is a product of Anglo-American educational systems and that a diminished role of this feature as an indicator of communicative effectiveness for particular cultures and languages should be considered.

### **Conclusion**

That the complex constellation of interactional and linguistic features that are important to effective communication in one language, culture, and context are not necessarily equally important, if at all, in another language, culture, and context is now a well-accepted fact; but, it is one that takes on added significance in the grassroots development of comparable tests of speaking proficiency in multiple languages that are going to be administered to a single population. Research within the social interactionist perspective, employing both quantitative and qualitative methods, has begun to chart these constellations.

Based on a comprehensive and critical examination of its language proficiency program, the Residential College in the Arts and Humanities at Michigan State

University decided to embark upon the rigorous task of program revision. This has involved the incorporation of a Cultures and Languages Across the Curriculum program into the college as well as the development of an accompanying RCAH Test of Speaking Proficiency. This process has been guided by the democratic assessment practices provided by Shohamy (2001a, 2001b). Specifically, following CLT, in collaboration with all stakeholders, the uses and consequences of existing tests were examined; the RCAH has assumed responsibility for all aspects of the proficiency test; new forms of assessment have been developed, most notably with key input from the test taking population; and, finally, the test development process continues to include the knowledge of different groups into assessment material as exemplified by possibly repositioning the place traditionally held by syntactic complexity in construct definition and related scoring rubric.

The growing body of research on paired speaking assessments has provided a solid foundation, both theoretical and methodological, on which this project could build. Producing more questions than answers, preliminary analysis of test trial transcripts and feedback from students and examiners have, nonetheless, suggested limited usefulness of quantitative measures of syntactic complexity in capturing the intricacies of what it means to be a proficient speaker in a language other than English. And, while the issues of the powerful impact of testing context and educational setting on performance are certainly not new, it is hoped that their appearance here might serve to remind us to keep them at the fore.

### Acknowledgements

I would like to thank the anonymous reviewers for their useful feedback. I am grateful to the Editor of this PLTA issue for her probing questions and insightful suggestions. All shortcomings remain my own.

### References

- Aguinis, H., Simonsen, M.M., & Pierce, C.A. (1998). Effects of nonverbal behavior on perceptions and power bases. *The Journal of Social Psychology*, 138(4), 455–469.
- American Council on the Teaching of Foreign Languages. (2012). *ACTFL proficiency guidelines*. Alexandria, VA: ACTFL, Inc.
- Berwick, R., & Ross, S. (1996). Cross-cultural pragmatics in oral proficiency interview strategies. In M. Milanovic, & N. Savielles (Eds.), *Performance*

- testing, cognition and assessment: Selected papers from the 15<sup>th</sup> LTRC* (pp. 34–54). Cambridge: Cambridge University Press.
- Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating and using psychoeducational tests in cross-cultural assessment. *School Psychology International*, 12, 119–132.
- Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, 26(3), 341–366.
- Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing*, 20(1), 1–25.
- Brown, A., & Hill, K. (Eds.). (2008). *Tasks and criteria in performance assessment*. Germany: Peter Lang.
- Bygate, M., Skehan, P., & Swain, M. (Eds.). (2001). *Researching pedagogic tasks: Second language learning, teaching and testing*. Harlow, England: Longman.
- Byrd, D., Hlas, A., Watzke, J., & Valencia, M. (2011). An examination of culture knowledge: a study of L2 teachers' and teacher educators' beliefs and practices. *Foreign Language Annals*, 44, 4–39.
- Chalhoub–Deville, M. (2001). Task-based assessment: a link to second language instruction. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 210–228). Harlow, England: Longman.
- Council of Europe. (2001). *Common European Framework of reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Davis, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing*, 26(3), 367–396.
- Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. *Language Testing*, 26(3), 423–443.
- Eddy, P. (1980). Foreign languages in the USA: A national survey of American attitudes and experience. *Modern Language Journal*, 64, 58–63.
- Educational Testing Service. (2014). *iBT/Next Generation TOEFL Test Independent Speaking Rubrics*. Princeton: NJ. Retrieved from [http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking\\_Rubrics.pdf](http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf).
- Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18, 299–323.
- Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, 59(4), 866–896.

- Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, 21(3), 354–375.
- Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Education.
- Furman, N., Goldberg, D., & Lusin, N. (2008). *Enrollments in languages other than English in United States institutions of higher education, Fall 2006*. New York: The Modern Language Association.
- Galaczi, E. (2008). Peer–peer interaction in a speaking test: The case of the First Certificate in English Examination. *Language Assessment Quarterly*, 5(2), 89–119.
- Hambleton, R. K., Merenda, P. F., & Spielberger, C. D. (Eds.). (2005). *Adapting educational and psychological tests for cross–cultural assessment*. New Jersey: Lawrence Erlbaum.
- Herzog, M. (2007). An overview of the history of the ILR language proficiency skill level descriptions and scale. Retrieved from <http://www.govtilr.org/Skills/>.
- Hudson, T. (2012). Standards–based testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 479–494). London: Taylor and Francis.
- Interagency Language Roundtable. (2014). *Skill level descriptors – Speaking*. Washington, DC. Retrieved from <http://www.govtilr.org/Skills/ILRscale2.htm>.
- Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal*, 87, 90–107.
- Keck, C. M., Iberri–Shea, G., Tracy–Ventura, N., & Wa–Mbaleka, S. (2006). Investigating the empirical link between task–based interaction and acquisition: A meta–analysis. In J. M. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Amsterdam: John Benjamins.
- Kowner, R., & Wiseman, R. (2003). Culture and status–related behavior: Japanese and American perceptions of interaction in asymmetric dyads. *Cross–Cultural Research*, 37(2), 178–210.
- Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*, 13, 151–172.
- Long, M. H. (2011). Methodological principles for language teaching. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 373–394). Hoboken, NJ: Wiley–Blackwell.



- Magnan, S. S., Murphy, D., Sahakyan, N., & Kim, S. (2012). Student goals, expectations, and the standards for foreign language learning. *Foreign Language Annals*, 45, 170–192.
- Mast, M. S., & Hall, J. A. (2004). Who is the boss and who is not? Accuracy of judging status. *Journal of Nonverbal Behavior*, 28(3), 145–165.
- May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing*, 26(3), 397–421.
- May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, 8, 127–145.
- Mignault, A., & Chaudhuri, A. (2003). The many faces of a neutral face: head tilt and perception of dominance and emotion. *Journal of Nonverbal Behavior*, 27 (2), 111–132.
- Nakatsuhara, F. (2006). The impact of proficiency-level on conversational styles in paired speaking tests, *University of Cambridge ESOL: Research Notes*, 25, 15–20.
- National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- National Standards in Foreign Language Education Project (2006). *Standards for foreign language learning in the 21<sup>st</sup> century*. Yonkers, NY.
- Plough, I.C., & Bogart, P.S.H. (2008). Perceptions of examiner behavior modulate power relations in oral performance testing. *Language Assessment Quarterly*, 5(3), 1–23.
- Plough, I. C., MacMillan, F., & O'Connell, S.P. (2011). Changing tasks... changing evidence: A comparative study of two speaking proficiency tests. In G. Granena (Ed.), *Selected Proceedings of the 2010 Second Language Research Forum* (pp. 91–104). Somerville, MA: Cascadilla Proceedings Project.
- Rivers, W. P., & Robinson, J. P. (2012). The unchanging American capacity in languages other than English: Speaking and learning languages other than English, 2000–2008. *The Modern Language Journal*, 96, 369–379.
- Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, 22(1), 27–57.
- Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL*, 43(1), 1–32.
- Ross, S. (1992). Accommodative questions in oral proficiency interviews. *Language Testing*, 9, 173–186.

- Ross, S., & Berwick, R. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14(2), 159–176.
- Shohamy, E. (2001a). Democratic assessment as an alternative. *Language Testing*, 18(4), 373–391.
- Shohamy, E. (2001b). *The power of tests: A critical perspective*. London: Longman.
- Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: second language learning, teaching and testing* (pp. 168–185). Harlow, England: Longman.
- Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, 30(4), 510–532.
- Swain, M. (2001). Integrating language and content teaching through collaborative tasks. *The Canadian Modern Language Review*, 58(1), 44–63.
- Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*, 58(2), 439–473.
- Van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, 23(3), 489–508.
- Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition*, 14, 403–424.