AAUSC 2012 Volume – Issues in Language Program Direction

Hybrid Language Teaching and Learning: Exploring Theoretical, Pedagogical and Curricular Issues

Fernando Rubio (Editor)
Joshua J. Thoms (Editor)
Stacey Katz Bourns (Series Editor)

HEINLE
CENGAGE Learning

Australia • Brazil • Japan • Korea • Mexico • Singapore • Spain • United Kingdom • United States

AAUSC 2012 Volume – Issues in Language Program Direction: Hybrid Language Teaching and Learning: Exploring Theoretical, Pedagogical and Curricular Issues
Fernando Rubio, Joshua J. Thoms and Stacey Katz Bourns

Editorial Director: P. J. Boardman

Publisher: Beth Kramer

Editorial Assistant: Gregory Madan

Managing Media Editor: Morgen Gallo

Executive Brand Manager: Ben Rivera

Market Development Manager: Courtney Wolstoncroft

Executive Marketing Communications Manager: Jason LaChapelle

Rights Acquisitions Specialist: Jessica Elias

Manufacturing Planner: Betsy Donaghey

Art and Design Direction, Production Management, and Composition: PreMediaGlobal

Library of Congress Control Number: 2012950573

ISBN-13: 978-1-285-17467-9

ISBN-10: 1-285-17467-4

Heinle
20 Channel Center Street
Boston, MA 02210
USA

Cengage Learning is a leading provider of customized learning solutions with office locations around the globe, including Singapore, the United Kingdom, Australia, Mexico, Brazil and Japan. Locate your local office at: **international.cengage.com/region**

Cengage Learning products are represented in Canada by Nelson Education, Ltd.

For your course and learning solutions, visit **www.cengage.com.**

Purchase any of our products at your local college store or at our preferred online store **www.cengagebrain.com.**

Instructors: Please visit **login.cengage.com** and log in to access instructor-specific resources.

Printed in the United States of America
1 2 3 4 5 6 7 16 15 14 13 12

Chapter 8

Complementary Functions of Face-to-Face and Online Oral Achievement Tests in a Hybrid Learning Program

Susanne Rott

The recent evolution in audio-based social networking technologies has provided many opportunities for language learners to improve their speaking abilities. A multitude of easy-to-use online voice recording tools allows students to share their research, ideas, opinions, and other types of narratives either asynchronously (e.g., podcasts, voiceblogs) or synchronously (e.g., chat tools). Although many of these tools are integrated into the publishers' websites accompanying textbooks and classroom management systems, they can also be found as shareware on the Internet.

Consequently, speaking prompts can extend the content discussed during class time and complement the use of discourse functions of face-to-face (F2F) class interactions. Partner, group, or whole-class assignments can emulate real-life F2F interactions and, to that extent, constitute meaningful exchanges among students (Abrams, 2011). The advantages for language development are many. The additional time needed to interact synchronously during which learners need to access and retrieve vocabulary and grammar on the fly may further the development of oral fluency (e.g., Hewett, 2000; Sun, 2009). Moreover, asynchronous assignments allow students to plan their output, pushing them to use the target language at the sentence and discourse levels. These assignments thus provide a practice opportunity that is often not available during class time for all students. Likewise, the option to re-record a response encourages students to monitor their language use (e.g., Sun, 2009) as well as their knowledge of the content. As a result, students are more likely to notice their knowledge gaps. They can then fill these gaps by looking up information in the course materials, dictionary, or a grammar reference guide. Additionally, if online assignments provide examples as speaking prompts, they have the potential to increase students' awareness of the sociopragmatic aspects of speech acts and the genre-specific uses of lexicogrammatical phrases (Sykes, 2005).

A natural extension of such online homework assignments is online oral tests. Students generally perceive homework and in-class activities as important if the relationship between these activities and what is being tested is apparent to them. This positive link has generally been described as a beneficial washback effect for learning (Shohamy, Reves, & Bejarano, 1986). That is, learners recognize the relevance of their oral homework to their ability to perform well on exams. In addition, technology can simplify the administration and grading of tests. The class can meet in a computer lab and take about 10 minutes to complete the recording

of a three- to five-minute oral test. Thus, students' performance is not ephemeral, as is the case in F2F oral exams when the instructor needs to assess multiple students at the same time (e.g., role-plays) or when the teacher is the interviewer and evaluator at the same time (e.g., one-on-one interview).[1] The simplicity of conducting online oral exams may allow for the administration of multiple oral assessments during a single semester. As a result, online oral exams may result in a "more well-rounded and accurate picture of student abilities than a single oral interview" (Dykstra-Pruim, 1997, p. 16).

Most of the test banks accompanying textbooks, as well as the pedagogical literature (e.g., Bachman & Palmer, 1996; Dykstra-Pruim, 1997), provide a variety of example materials for oral achievement tests. These range from narratives and teacher–student and student–student interviews to role-play situations. And yet, these pedagogical resources tell little about which aspects of oral language abilities can be elicited from a specific test format or individual tasks. Obviously, one criterion for a successful oral test is whether the learner has accomplished the task (e.g., ordering a meal, describing a trip). At the same time, however, each test format or task requires a number of competencies that may vary from task to task. For example, although a narrative requires a genre-specific discourse structure and the use of cohesiveness markers (i.e., discourse competence), role-plays and interviews require context-specific turn-taking behaviors and possibly repair strategies (i.e., pragmatic and interactional competencies). Additionally, each task requires specific linguistic tools to accomplish the task (e.g., to convince, explain, or describe). That is, learners need to demonstrate grammatical competence by choosing and producing the grammatical structures that best encode their ideas and the conventionalized phrases that most clearly express their intended meaning (lexical competence).

Research on oral language assessment has mostly focused on the validity and reliability of standardized proficiency tests (e.g., Kenyon & Tschirner, 2000) or learners' language development across proficiency levels (e.g., Martinsen, Baker, Bown, & Johnson, 2011). Conversely, research on task-based language assessment has explored the question of whether examinees can accomplish a given task (e.g., Norris, Brown, Hudson, & Yoshioka, 1998) and how the manipulation of task characteristics affects the complexity and accuracy of student production (e.g., Gilabert, 2007b; Robinson, 2001).

In the past 30 years, many aspects of oral tasks and exams have been researched to explain variability in student performance. Yet no generalizable findings that can be applied to a wide variety of teaching and testing settings have emerged. Robinson (e.g., 2001) and Skehan (e.g., 1998) have attempted to establish a framework that predicts the effect of task complexity on fluency and accuracy. Although Skehan assumes that complexity and accuracy of language use are in competition because attentional resources are limited, Robinson argues that

[1]It is certainly possible to record one-on-one exams. However, it is a very time-consuming process to first conduct the exam and then listen to the recording to evaluate it. Moreover, it is possible to record role-plays. Yet it is challenging to recognize individual voices of students throughout a recoding in order to assign a grade.

increasing cognitive demands may improve accuracy and complexity simultaneously. He suggests that tasks can be set up in a way that completion does not draw on the same cognitive resources. Although empirical findings have been mixed for both positions (e.g., Gilabert, 2007a; Mehnert, 1998; Yuan & Ellis, 2003), studies have been able to show that planning time, for example, can be a mitigating factor. Mehnert found that students who had planning time before engaging in a speaking task performed better than students with no planning time. Moreover, whereas learners who had one minute to prepare focused more on accuracy, learners who had 10 minutes tried to produce more complex speech. Likewise, Yuan and Ellis showed that pre-task planning time promoted higher complexity and lexical variety but not necessarily accuracy. However, when learners had unlimited time for speaking, they performed less fluently yet more accurately because they engaged in reformulation and self-correction. Addressing the correct use of collocations during exam tasks, Wang and Shih (2011) found that students' language for thinking, as they work on a task, affects collocational language use. Test takers who reported that they thought in both the first language (L1) and the second language (L2) produced more accurate collocations than learners who reported that they were thinking either in the L1 or the L2.

Many studies have explored the social dynamics in interview and paired tests and group-based tests. Brooks (2009) summarizes research findings emphasizing what Lazaraton (1996) calls a wild-card effect. Performances among multiple test takers are intricately linked and therefore can vary tremendously. Outlining effects of extraversion, personality, and gender, among other variables, Nakatsuhara (2011) questioned how scores can be assigned for individual performances. Likewise, issues of authority and dominance affect student performance during paired tasks and, in particular, teacher–student interviews. Brooks pointed out that student performance depends on whether the examiners simplify their language or use display questions in order to accommodate students. In fact, based on the discourse analysis of two different tester–student interactions, Brown (2003) suggested that a student's performance strongly depends on the examiner's goal orientation and the way of asking questions, as well as the type of interactional feedback a learner receives.

Other studies have examined how a conversation task triggers negotiation moves. Nakahama, Tyler, and Van Lier (2001), for example, found that information gap activities resulted in more negotiation of meaning than conversational situations in which learners interacted with native speakers. Interestingly, they also observed that in the conversational situation, learners paid more attention to the entire discourse, but in the information gap activity, they attended mostly to individual lexical items.

Few studies have directly compared the use of F2F and different modalities of computer-mediated communication tools. Comparing negotiation patterns of a jigsaw task through video-conferencing, audio-conferencing, or F2F chats, Yanguas (2010) revealed that the group using audio-conferencing engaged in the most negotiation moves. Yanguas suggested that this was because of the lack of visual information. Further investigating the acquisition of speech acts, Sykes (2005) compared the effectiveness of group discussions in written and oral chat as

well as in a F2F format. She discovered that the written and oral chat resulted in better learning than the F2F format. Moreover, the two online modalities demonstrated differences in interaction strategies. Although the written chat led to more complex interactions, the oral chat produced more diverse strategies for managing interactions.

Recent research has attended to various aspects of oral tasks. Yet there has been very little analysis of how different test tasks and formats are complementary in terms of the skills they elicit. One of the challenges facing practitioners who are interested in integrating multiple oral assessments into their language curriculum is determining which oral language competencies are elicited by a specific assessment task and, conversely, which tasks should be used to test the wide spectrum of oral competencies. That is, which oral assessment tasks would complement each other by eliciting the use of different strategies by the test taker? In particular, it is of interest to curriculum designers to better understand which test format lends itself best to assess a certain competence. The current investigation sought to compare students' performances on oral tests to determine whether a particular assessment task provided a good snapshot of their language abilities. Moreover, to align instruction and assessment, oral tests should be an integral part of the language curriculum. Accordingly, in a hybrid language program, the integration of oral assessments would mean the use of both F2F and online oral tests. Therefore, this study sought to shed light on the role of online oral tests compared with F2F tests. The three oral tests compared were a monologue narrative performed online, a role-play among three students who performed F2F, and a teacher–student F2F interview.

Research Questions

1. Do assessment tasks (monologue, interview, and role-play) elicit different aspects of oral competence in students' performance?
2. If the tasks elicit different aspects of oral competence, can the tasks be considered complementary?
3. Is there a relationship between the test task and students' perception of their ability to demonstrate their speaking abilities?

Method

Participants

Forty-five learners of German participated in this investigation. Students were enrolled in the third-semester course of the basic language sequence because they either had earned a passing grade in the previous two semesters or had placed into the course by taking a placement test. The basic language courses are delivered in a blended learning (BL) format. Students meet F2F in class three days each week. They also complete online learning materials outside of class in place of a fourth F2F class session. Online materials engage learners either in input-based

preparatory work (e.g., learning vocabulary and its pronunciation as well as reading texts with the support of audio, cultural information, and an online dictionary) or in communication-focused output tasks. The latter tasks expand on the content covered in class mostly at the sentence level to written and oral language use at the discourse level (e.g., written blogs, oral voice blogs).

Procedure

Students were assigned to three treatment conditions. Two of the conditions took place F2F: one-on-one oral interviews and small-group role-plays. The third condition, monologue, took place in the computer lab, where students recorded their responses with an online recording tool. Students from three sections of third-semester classes participated in the study. Each condition presented an intact class. Random assignment of students from each class to the different treatment conditions was not possible. The three classes took place at different times during the day, and logistically it was not possible to administer the three different types of exams within one class section. In each treatment condition, students received the speaking prompt (see below) and had about two minutes to think about the topic and take some notes on the sheet with the prompt. They were not allowed to write complete sentences or read from their notes while they engaged in the oral exam. Notes were collected after the exam to verify that the participants had complied with the instruction to not write complete sentences. The instructor monitored students and intervened if students read from their notes as they were speaking during the exam. After completing the test, the participants were asked to complete a questionnaire about their learning and motivational background as well as their perception of the types of oral exams. For all three conditions, the data collection took place during a single 50-minute regular class session.

Materials

The oral achievement exam was based on a textbook chapter (Augustyn & Euba, 2008) that covered visiting the German city of Leipzig. Participants were asked to respond to the following prompt:

> During Spring Break you are going to visit the city of Leipzig with your classmates. You are going to fly from Chicago to Berlin and then take the train from Berlin to Leipzig. Plan your stay in Leipzig. You may want to address the following topics: How long do you want to stay in Leipzig; how to get from the train station to the hotel; what do you want to visit in Leipzig and why; where are you planning to eat; and whether you would like to visit the book fair. (translated from German)

Students were not limited to these questions. Rather, the questions served as a prompt reminding students of what they had discussed and learned in class. At the same time, the questions served as a guide for what was expected of them. All students were familiar with all three test tasks in their respective format (interview and role-play in the F2F and monologue in the online format) because they had been used as achievement tests in previous chapters and semesters. Privacy

settings were available only for podcasting in the password-protected classroom management system. Therefore, the monologue task was chosen for the online format. For any of the software that allows recording synchronous interactions, a privacy setting was not an option. Consequently, all students had access to each other's exam recordings.

Test Tasks

Students were informed about the format of the oral exam in which they would participate (interview, role-play, or monologue) one week in advance of the study. Likewise, they received information about the general topic (the city of Leipzig), the length of the oral exam (see below), and the grading criteria. Grading criteria were the following: content/task completion (35%), appropriate use of vocabulary (20%) and grammar (15%), pronunciation (15%), and fluency (15%). Additionally, the researcher visited the three classes, informed the instructor and the students that the exam was used for a research study, and asked the students for their consent of participation. The researcher emphasized that the format of the exam would not deviate from the format of previous exams.

One-on-One Oral Interview. The course instructor met with each participant individually F2F for about three to five minutes. The role of the instructor was to listen to each student's report, ask questions if he did not comprehend what the student was saying, and function as an interlocutor. He encouraged students to continue their explanations by stating something to the effect that the plan sounds interesting and whether there would be any other items on the itinerary. The entire interaction was recorded with a digital voice recorder.

All instructors in the program received training on how to administer oral interviews. Emphasis is placed on the procedure, including how to make students feel comfortable (warm-up), how to encourage students to provide more details if the speech sample is too short to evaluate, how to request clarification if students' production is unclear, and how to finish the interview (wind-down).

Role-play. Three students were randomly assigned to each group to discuss their trip to Leipzig F2F. In the two-minute preparation and note-taking phase, students worked as a group. During the exam, the instructor did not intervene in the dialogue among the students. He observed students' interactions and took notes to assign a grade later. Only in two instances when a student did not participate sufficiently to award a grade, the instructor asked follow-up questions. The role-plays took about three to five minutes and were recorded with a digital voice recorder. When the group of students entered the classroom, the instructor made them feel comfortable by asking them, in German, how they were doing and talking about the weather (warm-up).

Monologue. Students were sitting at individual computer stations recording their monologues with an online voice recorder integrated into the classroom management system. Students were told that the monologue should be about three minutes long and that the maximum recording time was five minutes. The recordings were completed with a microphone that was attached to a headset. The use of the

headset prevented students from overhearing what other students were recording. Before the recording of the exam monologue, students were required to complete a 10-second test recording to ensure that their microphone and recording devices were working.

Questionnaire. The purpose of the questionnaire was to determine the participants' attitudinal reactions to the test format and test tasks. The questionnaire was adapted from Kenyon and Malabonga (2001). The questionnaire sought to shed light on whether students had a preference for one of the formats and whether they perceived one test format as a better measure of their ability. Students were asked to respond to the following three statements on a four-point scale: (a) I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking on the test; (b) I feel the test was difficult; and (c) I feel someone listening to my test responses would get an accurate picture of my current ability to speak in real-life situations outside the classroom.

Analysis

The recordings of all three treatment conditions were transcribed by research assistants. Transcriptions included false starts, repetitions, hesitations, and self-corrections. The recordings were analyzed for the following aspects:

Communication

C-units. Following Abrams (2003), C-units were used as the unit of analysis comparing the amount of content that learners produced in each condition. C-units are "isolated phrases not [necessarily] accompanied by a verb, but they have a communicative value" (Crookes, 1990, p. 184). For example, in the following interactional exchange, the question and the response are counted as a c-unit: "Q: Where is my hat? A: On the table" (Crookes, 1990, p. 184). C-units were used because they capture phrases not accompanied by a verb, which is a natural phenomenon in oral interaction. The commonly used utterance as a unit of analysis for spoken discourse was not adopted for the study because the analysis relies on intonation contour and pauses. Participants in the study were intermediate learners who cannot be assumed to have full control over intonation contours. The transcriptions were coded by two research assistants and the researcher. Interrater reliability reached 94.3 percent. Discrepancies were resolved by the principal investigator. Discrepancies mostly occurred with single-word responses in the role-play and the interview condition. Single-word responses referring to time, such as *afternoon* (*Nachmittag*) or *Monday* (*Montag*), counted as c-units even if the preposition was missing. In turn, single-word contributions, such as *opera* (*Oper*), were not counted as a c-unit because they did not specify whether the speaker suggested to visit the building, watch a performance, or walk by the opera building.

Number of Words Produced. The number of words produced by each student provided an additional measure to determine the size of the language sample obtained. The count included false starts, reformulations, self-corrections, and repetitions.

Interactional Competence.

Editing Features. Editing features, such as self-corrections, false starts, and reformulations, generally demonstrate that learners monitor their output. The use of editing features was considered an indicator of the students' concern about being understood and getting their meaning across clearly. Therefore, only linguistic edits (vocabulary and grammar) were counted. For example, self-corrections happened when learners corrected verb endings; false starts happened mostly when learners changed verbs (*Dann stehen wir drei . . . dann bleiben wir drei Tage in Leipzig; Then we stand three . . . then we stay for three days in Leipzig*); reformulations mostly happened when participants attempted to improve their pronunciation. For the analysis, no strict distinction between the three editing features needed to be made. Each attempt to use any of these three features counted as one point even if the editing did not result in a more comprehensible utterance. The attempt to edit one's output is obviously higher on the interactional competence continuum (no attempt to a successful attempt) than no attempt at all.

Negotiation of Meaning. Any clarification request or comprehension check initiated by a learner was awarded one point. Negotiation moves were initiated when miscomprehension occurred because of lexical, grammatical, or pronunciation issues. This comparison was conducted only between the role-play and the interview conditions. The online monologue condition did not require the use of this aspect of interactional competence.

Questions

Content questions asked in order to solicit further information or an opinion from interlocutors were awarded one point. Again, this comparison was conducted only between the role-play and the interview conditions. The online monologue condition did not require the use of this aspect of interactional competence.

Discourse Competence

Word-to-Clause Ratio. Discourse competence describes learners' ability to speak at the sentence and multi-sentence level as compared with the multi-word level. A higher level of discourse competence indicates that learners support and provide details for their statements. This aspect of discourse competence was calculated by the ratio of words per total number of clauses (main and subclauses combined).

Discourse Markers. Participants' use of discourse markers was used as a measure of discourse competence. Discourse markers included but were not limited to adding (e.g., and, besides, in addition), structuring (e.g., first of all, finally, then), and logical consequence (e.g., as a result, consequently, then) markers. The use of discourse markers demonstrated that participants were able to produce cohesive discourse at the multi-sentence level. Participants earned one point for each attempt to use any discourse marker. Instances of partial or incorrect use of discourse markers were also awarded one point because it indicated the emerging development of discourse competence.[2]

[2]No explicit difference was made between correct use and emergence of discourse markers.

Lexical Competence

Collocations. Participants' data were analyzed for the correct use of conventionalized multi-word chunks. The use of collocations is considered an important factor in one's ability to express ideas concisely as well as an indicator of the development of advanced language abilities (e.g., Schmitt & Carter, 2004; Wray, 2001). The use of native-like (Pawley & Syder, 1983) expressions demonstrated that students did not need to use L1 transfer as a crutch to express their ideas. Noun–verb combinations, such as *stay in a hotel* (*im Hotel übernachten* compared with **im Hotel stehen*, which translates as *stand in a hotel*), or expressions, such as *I go to Leipzig* (*ich fahre nach Leipzig* compared with **ich gehe nach Leipig*," which translates as *I walk to Leipzig*) were counted as multi-word chunks. The transcripts were coded by two research assistants and the principal investigator. Interrater reliability was 92 percent.

Lexical Richness. Lexical richness is an indicator of learners' breadth of the lexicon. It was measured by means of Giraud's Index (Vermeer, 2000), according to which the total number of word types is divided by the square root of the total number of word tokens. The square root is included to eliminate the effect of differences in text length.

Grammatical Competence

Syntactic Complexity. Syntactic complexity was calculated in order to determine which task pushed learners to produce more complex language as a measure of grammatical competence. Complex language use is generally marked by using subordinate clause structures. Therefore, complexity was measured by determining each learner's subordinate-to-clause ratio. Following previous studies (e.g., Foster & Skehan, 1996; Michel, Kuiken, & Vedder, 2007; VanDaele et al., 2006), complexity was computed by dividing the total number of subordinate clauses by the total number of clauses. Previous research has shown that students produced syntactically simpler structures in dialogues than in monologues (Michel et al., 2007).

Results

The first research question sought to determine whether different types of oral assessment tasks elicit similar or different oral competencies. Table 8-1 presents descriptive statistics from the three treatment conditions tasks—monologue, interview, and role-play—for all aspects of the production data that were compared. For all comparisons, one-way ANOVAs (analyses of variance) were conducted. Post hoc contrasts between the three different treatment conditions were submitted to Tukey's HSD (honestly significant difference) test because the three sample sizes were equal and homogeneity of variance was met.

Communication

The first objective of the study was to determine whether the three treatment tasks led to the same amount of information exchange and whether the tasks produced the same amount of language from each participant. Although the

Table 8-1. Descriptive Statistics (Means and Standard Deviations) of All Measures for All Three Conditions

Measure	Interview (n = 15)	Role-Play (n = 15)	Monologue (n = 15)	Results
Communication				
content units	32.00 (13.40)	24.40 (11.61)	29.60 (17.04)	NS
Number of words	178.00 (43.80)	138.13 (41.68)	203.60 (58.41)	Monologue > role-play Interview > role-play
Interactional competence				
Editing features	0	0	0.44 (0.73)	
Negotiation of meaning	0.20 (0.41)	0.40 (0.63)	0	NS
Questions	0.87 (0.91)	3.4 (2.41)	0	Role-play > interview
Discourse competence				
Clause-to-word ratio	0.09 (0.05)	0.05 (0.03)	0.11 (0.04)	Monologue > role-play Interview > role-play
Discourse markers	4.87 (3.24)	4.40 (2.89)	4.93 (3.01)	NS
Lexical competence				
collocations	3.60 (3.52)	1.8 (2.68)	3.73 (2.46)	Monologue > role-play Interview > role-play
Lexical richness Giraud's index	9.02 (1.50)	7.67 (1.87)	10.87 (1.47)	Monologue > role-play Monologue > interview
Grammatical competence				
Syntactic complexity Subclause-to-clause ratio	0.08 (0.12)	0.1 (0.05)	0.21 (0.12)	Monologue > role-play Monologue > interview

NS, not significant.

quantity of content exchanged in each condition was indeed the same ($F[2, 42] = 1.12$; not significant [NS]), the size of students' speaking samples varied significantly ($F[2, 42] = 6.93$; $p < .05$). The post hoc comparison revealed that the overall production of words was not significantly different when participants engaged in the monologue or the interview tasks. Not surprisingly, the role-play condition resulted in a significantly smaller language sample of each student's speech than the interview (mean difference = 65.47; 95% confidence interval [CI] = 22.41, 108.52; $p < .05$) and the monologue (mean difference = 39.87; 95% CI = −75.63, −4.1, $p < .05$) conditions.

Interactional Competence

Student production data were further analyzed for their use of interactional language characteristics, such as using editing features and posing questions, and the negotiation of meaning. None of these features was used much in any condition. Few participants self-corrected and reformulated their speech when they produced the monologue online ($\bar{x} = 0.44$), and none of the participants in the two other tasks used these features. By contrast, some students who participated in the role-play or the interview initiated clarification requests and confirmation checks to negotiate meaning ($F[2, 42] = 1.05$, NS). Naturally, this was not possible in the monologue task in which students sat individually at computer stations. Also, these students were unable to ask questions. However, participants in the role-play condition used the interactional feature of asking questions significantly more than students in the interview condition ($F[2, 42] = 21.07; p < .05$).

Discourse Competence

Participants' level of discourse competence was measured by the clause-to-word ratio and the use of discourse markers. The production data showed that the assessment task affected whether participants spoke more at the sentence and multi-sentence level than at the word or multi-word level ($F[2, 42] = 10.51; p < .05$). The post hoc comparison revealed that participants produced significantly more discourse-level language when they engaged in the monologue and the interview tasks as compared with the role-play (mean difference = .06; 95% CI = .03, .09; $p < .05$ and mean difference = .04; 95% CI = −.07, −.01; $p < .05$, respectively). Discourse markers were used with equal frequency in all three conditions ($F[2, 42] = .09$, NS).

Lexical Competence

Two aspects of lexical competence were compared: the correct use of conventionalized collocations as well as lexical richness. The treatment condition did have an impact on the number of correctly used collocations ($F[2, 42] = 8.6; p < .05$). Students who participated in the role-play used significantly fewer collocations than students in the interview (mean difference = 5.67; 95% CI = 2.85, 8.48; $p < .05$) and students in the monologue (mean difference = 3.87; 95% CI = 1.05, 6.68; $p < .05$) conditions. Likewise, the treatment condition had an impact on the students' ability to demonstrate the depth of their mental lexicon ($F[2, 42] = 14.66; p < .05$). In the monologue condition, participants used more varied vocabulary than in the interview (mean difference = 1.85; 95% CI = 1.73, 4.67; $p < .05$) and in the role-play (mean difference = 3.20; 95% CI = .38, 3.33; $p < .05$) conditions. The interview and role-play tasks resulted in the same level of lexical richness.

Grammatical Competence

Grammatical competence was determined by assessing participants' ability to use syntactic subordination. Similar to the lexical richness, the assessment task affected syntactic complexity ($F[2, 24] = 14.47, p < .05$). The monologue task

Table 8-2. Descriptive Statistics for the Questionnaire for All Three Conditions*

	Condition		
	Interview	**Role-Play**	**Monologue**
I feel I had the opportunity to adequately demonstrate both my strengths and my weaknesses in speaking on the *test*.	3.15 (.55)	3.33 (.48)	2.33 (.72)
I feel the *test* was difficult.	2.15 (.55)	2.39 (.78)	2.30 (.77)
I feel someone listening to my *test* responses would get an accurate picture of my current ability to speak in real-life situations outside the classroom.	3.08 (.90)	2.72 (.26)	2.13 (.92)

*$n = 45$
4 = strongly agree, 3 = agree, 2 = disagree, 1 = strongly disagree.

led participants to use more syntactically complex sentences than the interview (mean difference = .20; 95% CI = .11, .29; $p < .05$) and role-play (mean difference = .13; 95% CI = .04, .22; $p < .05$) conditions.

Student Perception

Table 8-2 reports the means and standard deviations of the questionnaire that assessed the participants' perception of the test. Participants in each of the three treatment conditions ranked the three assessment tasks as equally difficult. They thought that the test was not very difficult. However, when participants were asked whether they thought that they had the opportunity to demonstrate their strengths and weaknesses (question 1), students who participated in the online monologue task mostly disagreed (xM = 2.33), but learners in the role-play (xM = 3.33) and interview (xM = 3.15) conditions thought that they had been able to demonstrate their abilities adequately. Likewise, participants who had completed the monologue task online (xM = 2.13) thought that they had not been able to demonstrate their ability to speak German in real-life situations outside the classroom. In contrast, when students engaged in the role-play (xM = 2.72) or the interview task (xM = 3.08), they thought that they had had the opportunity to demonstrate speaking abilities similar to those needed in real-life situations.

Discussion

This study is of particular relevance to language program directors who are involved in pedagogical decisions and the design of BL courses. The reduction of F2F time in BL courses often results in a prevalence of online speaking tasks. Thus, an understanding of what types of online and F2F tasks provide opportunities to

practice specific competencies, and thereby have a direct impact on students' linguistic development, is essential. Because assessment tasks should be a natural extension of classroom tasks and homework assignments, the main goal of this investigation was to determine whether different oral assessment tasks trigger and consequently allow to measure different language competencies. Understanding which language competencies can be elicited with a single test task is vital for program directors in order to make informed pedagogical decisions regarding the oral assessment component for a language program. The assessment tasks used in this investigation were three commonly used exam tasks: role-play, a teacher–student interview, and a monologue. The study showed that each task elicited a combination of different competencies.

The first aim of the analysis was to establish that participants in all three conditions accomplished the task of planning a trip to Leipzig. Although students' responses as to which sites they wanted to visit and which means of transportation they were planning to take varied, the overall number of ideas produced was similar in each task condition. This quantitative measure served to establish a content baseline for further qualitative analyses of oral competencies. A second quantitative baseline measure assessed the amount of language produced in each condition. The rationale for this analysis was to ensure that the production data collected from each student provided a large enough sample to determine students' current language abilities. Because the role-play condition required the co-construction of content by the three participants, student samples were significantly shorter than in the interview and monologue conditions. That is, these students had fewer opportunities to demonstrate their competencies.

In fact, students in the role-play condition were not able to demonstrate their competencies to talk at the multi-sentence discourse level, their depth of lexical knowledge, or their syntactic abilities as well as participants in the other two conditions. The role-play task led students to talk in multi-word strings, with a smaller variety of words, and with little subordinate sentence structure. The need to interact with multiple task participants and under time constraints may have depleted the resources necessary for using more complex language. Nevertheless, the role-play task required students to demonstrate interaction skills that participants in the other two conditions did not need to show. Role-play participants had to comprehend and immediately respond to their peers' statements and questions without much time to plan their responses. That is, they needed to adhere to the talk patterns and dynamics of a group conversation. Consequently, it is vital that evaluation criteria for role-plays reflect this evidence. Moreover, future research needs to analyze more closely turn-taking and talk patterns. Some students' performances seemed very short and demonstrated parallel structures of turn taking. These students' participation was marked by contributing sentences or multi-word chunks without a clear reference to what was said in the previous turn, an issue outlined by Dimitrova-Galaczi (in Gan, 2010). In turn, one may require students in the role-play tasks to perform twice as long in order to provide them with more opportunities to demonstrate their language abilities. However, intermediate learners may not have more content knowledge to talk about to fill an additional three to five minutes.

Besides asking questions, the role-play condition nevertheless did not result in much use of interactional competencies. Although group interaction depended on the students' successfully comprehending each other, they did not seem to monitor their comprehensibility by correcting themselves or reformulating their ideas. Likewise, not many of the role-play participants initiated the clarification of other participants' statements. Of course, one reason could have been that participants did not feel the need to request clarification. Even though their output had lexical and grammatical errors, they may have comprehended each other's utterances—or at least they comprehended each other's utterances well enough to continue the dialogue. After all, the context and goal of the role-play dialogue was clear to all of the participants because they had had pre-task planning time before the exam. That is, planning time before engaging in a role-play may limit the need to understand precisely the interlocutor's message. It may be interesting to explore in future research whether eliminating planning time will lead to more clarification and negotiation moves because students will not know the outcome of the interaction from the outset of the task. Future research should further examine which types of errors specifically trigger learners to self-correct or to initiate negotiation of meaning. Likewise, a more in-depth analysis of students' interaction data could illustrate whether their interaction was indeed logical and coherent.

Participants in the monologue and interview conditions were equally successful in showing their level of discourse competence by using discourse markers and producing multi-sentence stretches of speech. However, neither of these two tasks elicited many instances during which students used language demonstrating their interactional competence. Although students in the monologue task showed a tendency to focus on lexical and grammatical form by correcting themselves, students in the interview demonstrated a tendency to initiate clarification requests. Considering that in the role-play situation, learners did not produce many clarification moves, it seems safe to say that the unplanned interaction with the instructor required students to engage in clarification moves in order to avoid communication breakdown.

In many ways, the online monologue and the interview tasks were similar: in both tasks, students provided responses to the question prompts on the instruction sheets. But whereas students in the online environment were completely left to their own devices, students in the interview situation received support through nonverbal cues, such as encouraging facial expressions as well as guiding questions from the instructor. Still, learners who completed the monologue task online produced lexically richer and syntactically more complex language than learners who engaged in the interview with the instructor. One could argue that the combination of the monologue task and the online environment pushed learners to produce oral competencies at a linguistically (lexically and syntactically) more sophisticated level. Future research should make direct comparisons between a monologue performed online and a presentation performed in front of an instructor to further tease out the effect of the online environment.

Interestingly, even though participants in the online monologue condition were able to demonstrate the highest level of linguistic complexity (lexical

richness and grammatical competence) compared with the two other test tasks, the participants themselves did not think that they had been able to most adequately demonstrate their language abilities.[3] This may have been an artifact of the monologue task itself in which they did not have the teacher's or their peers' encouraging nonverbal cues or may have been attributable to the online test format. Even though these students had experience with online speaking activities, the online environment may have made them feel self-conscious because they may have noticed their linguistic limitations more than in a F2F speaking situation. Future investigations should tease apart the online factor and the monologue task factor to determine how they affect the level of comfort students experience in the assessment situation. Moreover, it would be interesting to see whether students considered the task to emulate a real-life speaking opportunity. To gain further insights into students' perception about online speaking prompts, future questionnaires could ask a question to the effect of whether they think that what they recorded is similar to what they would have said in front of a group of people.

Although the limitations of the current investigation prevent drawing definitive conclusions, the study suggests that a better understanding of which aspects of linguistic and interactional competencies a specific assessment task elicits is essential to test task design and the evaluation of student performance. Analyses of student production data revealed that the role-play task was distinctive in that the participants had to co-construct the content. Although the role-play task did not give learners the opportunity to display their linguistic competencies to the fullest extent, they were able to display and interpret nonverbal aspects of interactional competence. By contrast, participants who performed the monologue task in the online environment were able to demonstrate complementary competencies that demonstrated more sophisticated and complex language use. It seems too early to discard teacher–student interviews as a valuable oral assessment task. Nonetheless, the current findings revealed that learners neither demonstrated a higher level of interactive features compared with the role-play task nor elicited the same level of linguistic complexity as the monologue performed online. Moreover, considering that one-on-one interviews are time consuming and cumbersome to administer, monologue tasks recorded online may be an effective alternative. That is, a hybrid solution to oral testing seems to be a natural extension of the curricular design to the testing situation.

Limitations

The current study has a number of limitations that should be addressed in future investigations. Future studies should assess each student's performance on multiple test tasks and formats to gain more generalizable insights from a larger data set. Likewise, obtaining student performances across a variety of

[3]It may be possible that students may have interpreted the questions differently. They may have thought that what they do in front of a computer is not related to what happens in real life. That is, they may be assessing the comparability of the task rather than comparing the language produced.

different topics will help to determine whether the speech and interaction patterns identified in this study hold across different topics. Moreover, instead of using intact classes for each treatment, students should be assigned randomly to a condition. Additionally, it would be interesting to make a direct comparison of online and F2F testing formats. As collaborative technology tools become available for testing purposes, they may be a valuable alternative to time-consuming and difficult-to-administer F2F exams and may allow for multiple oral exams that test complementary competencies. Finally, students' self-evaluation of performance may be more valuable if the evaluation took place after several different exams. Other factors could have intervened to affect a student's perceptions besides the test task or mode of delivery, such as testing conditions (e.g., temperature, noise in the testing site, nervousness, composition of the small group, topic).

References

Abrams, Z. (2003). The effect of synchronous and asynchronous CMC on oral performance in German. *Modern Language Journal*, *87*, 157–167.

Abrams, Z. (2011). Interpersonal communication in intracultural CMC. In N. Arnold, & L. Ducate (Eds.), *Present and future promises of CALL: From theory and research to new directions in language teaching* (pp. 61–92). Texas: CALICO.

Augustyn, P., & Euba, N. (2008). *Stationen: Ein Kursbuch für die Mittelstufe*. Boston: Thompson/Heinle.

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.

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

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

Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied Linguistics, 11,* 183–199.

Dimitrova-Galaczi, E. (2004) *Peer-peer interaction in a paired speaking test: The case of the First Certificate in English*. Unpublished PhD dissertation, Teachers College, Columbia University.

Dykstra-Pruim, P. (1997). Integrating a series of oral assessments: Quick, low-stress options for the idealist. *Unterrichtspraxis, 30*, 16–29.

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

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing*, *27*, 585–602.

Gilabert, R. (2007a). The simultaneous manipulation of task complexity along planning and +/- Here-and-Now: Effects on L2 oral production. In M. P. García-Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 44–68). Clevedon, UK: Multilingual Matters.

Gilabert, R. (2007b). Effects of manipulating task complexity on self-repairs during L2 oral production. *IRAL, 45*, 215–240.

Hewett, B. (2000). Characteristics of interactive oral and computer-mediated peer group talk and its influence on revision. *Computers and Composition, 17*, 265–288.

Kenyon, D., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology, 5*, 60–83.

Kenyon, D., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *Modern Language Journal, 84*, 85–101.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing, 13,* 151–72.

Martinsen, R., Baker, W., Bown, J., & Johnson, C. (2011). The benefits of living in foreign language housing: The effect of language use and second-language type on oral proficiency gains. *Modern Language Journal, 95*, 274–290.

Mehnert, U. (1998). The effect of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20,* 83–108.

Michel, M. C., Kuiken, F., & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *Language Acquisition, 45*, 241–259.

Nakahama, Y., Tyler, A., & Van Lier, L. (2001). Negotiation of meaning in conversational and information gap activities: A comparative discourse analysis. *TESOL Quarterly, 35*, 377–405.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*, 483–508.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). Designing second language performance assessments. *Technical report 18, Second Language Teaching and Curriculum Center, University of Hawaii at Manoa*. Honolulu: University of Hawaii Press.

Pawly, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Native like selection and native like fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191–226). London: Longman.

Robinson, P. (2001). Task complexity, task difficulty and task production: Exploring interactions in a componential framework. *Applied Linguistics, 21*, 27–57.

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 1-22). Amsterdam: John Benjamins.

Shohamy, E., Reves, T., & Bejarano, T. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal, 40*, 212–20.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Sun, Y. (2009). Voice blog: An exploratory study of language learning. *Language Learning & Technology, 13*, 88–103.

Sykes, J. M. (2005). Synchronous CMC and pragmatic development: Effects of oral and written chat. *CALICO Journal, 22*, 399–431.

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

Wang, H., & Shih, S. (2011). The role of language for thinking and task selection in EFL learners' oral collocational production. *Foreign Language Annals*, *44*, 399–416.

Wray, A. (2001). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Yanguas, Í. (2010). Oral computer-mediated interaction between L2 learners: It's about time! *Language Learning & Technology, 14*, 72–93.

Yuan, F., & Ellis R. (2003). The effects of pretask planning and on- line planning on fluency, complexity, and accuracy in L2 monologic oral production. *Applied Linguistics, 24,* 1–27.