THE VALIDITY OF COMPUTER-MEDIATED COMMUNICATIVE LANGUAGE TESTS

by

Julian Charles Heather

A Dissertation Submitted to the Faculty of the

GRADUATE INTERDISCIPLINARY PROGRAM IN
SECOND LANGUAGE ACQUISITION AND TEACHING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2003

UMI Number: 3089949

UMI®

UMI Microform 3089949

Copyright 2003 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

THE UNIVERSITY OF ARIZONA ®
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by JULIAN CHARLES HEATHER

entitled THE VALIDITY OF COMPUTER-MEDIATED COMMUNICATIVE LANGUAGE TESTS

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of DOCTOR OF PHILOSOPHY

Renate A. Sch | 4/9/03
Dr. Renate Schulz | Date

Mary E Wildner-Bassett | 4/9/03
Dr. Mary Wildner-Bassett | Date

Robert Ariew | 4/9/03
Dr. Robert Ariew | Date

Jerry D'Agostino | 4/9/03
Dr. Jerry D'Agostino | Date

Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Renate A. Sch | 4/22/03
Dissertation Director | Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: T. C. Heather

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the support of the faculty of the University of Arizona's Second Language Acquisition and Teaching Program, and in particular my dissertation committee: Dr. Renate Schulz and Dr. Mary Wildner-Bassett of the Department of German Studies, Dr. Robert Ariew of the Department of French and Italian, and Dr. Jerry D'Agostino of the Department of Educational Psychology. A special thanks goes to Dr. Renate Schulz who, as my dissertation director, provided incredible amounts of long-distance guidance and support as I worked my way through several drafts of this manuscript.

Many others deserve recognition for the role they played in this study. I am extremely grateful to the teacher and students who welcomed me into their classes and agreed to participate in this research. The idea for this dissertation arose as the result of a conversation I had with my friend and colleague Lisa Jurkowitz in the summer of 2001. Her pioneering use of computer-mediated communication for classroom assessment inspired the initial question out of which this study grew. Along with Karen Barto, Christina Pruchna, and Anne-Marie Brooks, Lisa also provided assistance with parts of this study. Financial support was provided by a TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment.

Sharon Deckert has been a wonderful friend whose willingness to listen, give feedback, and provide emotional support has been invaluable. Finally, a huge thank you goes to my wife, Jamie, whose love has maintained me through the long process of writing.

DEDICATION

I dedicate this dissertation to my parents, Jay and Ted, for their love and encouragement of my desire for learning, and to my wife, Jamie, whose entrance into my life has inspired me to be so much more than I thought I ever could be.

TABLE OF CONTENTS

TABLE OF CONTENTS – Continued

TABLE OF CONTENTS – Continued

LIST OF FIGURES

LIST OF TABLES

LIST OF TABLES - Continued

ABSTRACT

A recent innovation in language testing involves the use of computer-mediated communicative language tests i.e., assessment of individuals' second language ability from transcripts of their interactions via computer-mediated communication (CMC). Studies have shown that such interactions in the first language involve a hybrid discourse with features of both written and spoken language, which suggests the possibility of making inferences about oral language ability from performance in a CMC environment. The literature to date offers little guidance on this matter. Research on computer-mediated communication has focused on its use in the second language classroom rather than in a testing context while studies of the linguistic and interactional features of second language learners' CMC discourse have mostly been descriptive with little direct comparison of CMC and face-to-face discourse.

This study, therefore, examines the validity of making inferences from computer-mediated discourse to oral discourse through a comparison of the performance of 24 third-semester French students on two tests: a computer-mediated communicative French test; and its nearest equivalent format in face-to-face testing, the group oral exam. Using a within-subjects design, counterbalanced for testing condition and discussion topic, the present study focuses on five areas which have important implications for validity: (a) the predictability of ratings of pronunciation on the group oral test; (b) the similarity of scores achieved on the CMC and group oral tests; the presence of similar (c) linguistic and (d) interactional features in the discourse of both tests; and (e) students' attitudes to the two tests. Results show that although scores on the two tests showed no statistically

significant difference, students' discourse differed in many respects which would, thus, invalidate any inferences made about oral ability from computer-mediated performance. Moreover, this study raises an important question about the role of computer-mediated communication in promoting second language acquisition since the computer-mediated discourse contained fewer examples of the negotiation of meaning routines that interactionist theories hold to be important to language acquisition.

CHAPTER 1

INTRODUCTION

1.1 GENERAL INTRODUCTION TO THE STUDY

Technology plays an increasingly important role in our lives, and as new technologies are developed, many of them find their way into second language classrooms. Among the latest to be appropriated by language teachers is computer-mediated communication (CMC). In CMC, learners use computers connected to a local or global network to communicate with each other either synchronously—through chat rooms, MOOs, MUDs, etc.—or asynchronously—through e-mail, listservs, or newsgroups. Traditionally, these interactions have occurred through written text. Although CMC using oral interactions is feasible and currently available—though to a limited number of people—written computer-mediated communication is likely to continue to dominate for some time to come because of the great expense in upgrading computer labs sufficiently to support audio- or video-conferencing.

CMC has been implemented in many different ways, involving interaction with both non-native and native-speakers of the target language, and claims about the potential benefits of its use in language classrooms have been quite consistent: CMC is an interactive tool (Johnston, 1999; Kern, 1998; Pinto, 1996) which can aid in the development of communicative competence (Kelm, 1996; Oliva & Pollastrini, 1995; Pellettieri, 2000); CMC provides an authentic audience for learners' interactions (Johnston, 1999); CMC interactions with native speakers give learners access to authentic input (Kelm, 1996; Oliva & Pollastrini, 1995) and help foster cultural understanding

(Bernhardt & Kamil, 1998; Cononelos & Oliva, 1993; Kern, 1996; Lee, 1997; Oliva & Pollastrini, 1995); CMC is learner-centered and self-paced, (Kelm, 1992, 1996; Kern, 1996; Lee, 1997), and it lowers negative affect such as anxiety (Kern, 1998; Sanchez, 1996); CMC allows learners to engage in metalinguistic analysis of transcripts of their interactions (Brammerts, 1996; Kelm, 1996); and finally, CMC represents a context that is in accord with sociocultural theories of learning (Barson & Debski, 1996; Beauvois, 1997; Kern, 1996; Kern & Warschauer, 2000; Peyton, 1999; Shetzer & Warschauer, 2000).

The growing prevalence of CMC in the classroom has also encouraged its adoption for classroom assessment (Jurkowitz, 2002; Kost & Jurkowitz, 2002). Jurkowitz (2002) offers a multi-layered rationale for the use of CMC as an assessment tool: (a) an assessment in which students negotiate meaning and integrate language and sociocultural knowledge in their interactions with a "natural audience (the interlocutor)" is inherently authentic in nature; (b) when classroom and testing tasks are strongly linked, there is a great potential for positive washback; and (c) "if students feel calm, engaged, and empowered by the computer as a tool in classroom activities, they would feel similarly even if the tool were used in an assessment situation" (pp. 25-26). Jurkowitz analyzed the transcripts of the performance of two of the eight students who took her electronic test; she found that the students' language used many different tenses with a relatively high proportion of correct usage (typically over 75%). There was also a high percentage of complex clauses. Responses to a questionnaire, completed by all the students who took the electronic exam, indicated that the majority of the students had liked the exam and had appreciated the opportunity to produce more connected discourse. As a result, many of the students recommended use of the test format in the future.

Kost and Jurkowitz (2002) describe the results of using CMC for assessment in intermediate level French and German classes. Both classes had used CMC for 12 45-minute discussions during the regular semester; the 12 students in the French class used IRC Français (a synchronous chat program) while the 20 students in the German class used POLIS, a bulletin board system developed at the University of Arizona (i.e., an asynchronous system). An examination of the transcripts produced during the tests revealed that the students produced a wide range of tenses with a relatively high degree of accuracy and used both communication and discourse management strategies. The use of complex sentences, however, varied between the two classes. Around half of the sentences produced by the German students were complex while only about a quarter of the sentences produced by the French students were complex.

Both these studies conclude that computer-mediated communicative tests allow students to demonstrate their language ability in an integrated way that can serve both formative and summative roles. In addition, Jurkowitz suggests a potential role for CMC, in conjunction with an oral exam, in testing general proficiency (2002, p. 35). Thus it appears that, for Jurkowitz, the CMC exam is clearly a form of writing assessment to be used separately from oral assessment. There is, however, an alternative option: To use a CMC test in lieu of an oral exam. This suggestion is not as fanciful as it may initially seem. In English, the use of synchronous CMC to interact with others is typically referred to as ‘chatting’, a nomenclature which hints at the oral qualities of the interactions and the language produced. In a similar way, CMC discourse has been referred to in the literature as “talking in writing” (Spitzer, 1986) and “written speech” (Maynor, 1994). Such descriptions are supported by studies of the CMC discourse of native speakers which provide empirical evidence of the hybrid nature of a CMC discourse which combines features of written and spoken language (Collot & Belmore, 1996; Ferrara, Brunner, & Whittemore, 1991; Gaines, 1999; Wilkins, 1991; Yates, 1996). Thus synchronous CMC is said to contain several features which resemble spoken language: disfluencies, hesitancies, and features which show involvement with one’s audience such as the use of direct questions, general emphatics, adverbs of time, references to the speaker’s mental processes, and the use of names and second person pronouns (Ferrara et al., 1991; Wilkins, 1991; Yates, 1996). Features of spoken language found in asynchronous CMC include the use of rhetorical questions, responses to imagined echo-questions, the use of informal lexical items, and the omission of subject pronouns, modals, auxiliaries, and copulas (Gaines, 1999; Maynor, 1994).

The fact that CMC discourse contains features of spoken language does not mean, *per se*, that a computer-mediated test can be used in place of an oral exam for the purpose of testing oral language ability. Any attempt to do so cannot proceed until an important question has first been addressed: Can one make inferences about a student’s speaking ability based on their performance on a computer-mediated (i.e., a written) test? This study will attempt to answer this question by comparing the performance of intermediate-level students of French on a CMC test to that on an oral test. The question is: What would be a suitable benchmark against which to compare the CMC performance? Since the focus here is on classroom rather than proficiency assessment, the answer depends on how CMC is used in regular lessons. The French classes which participated in this study typically used CMC in a similar way to the students in Jurkowitz and Kost's classes—as a discussion forum with groups of three or four students chatting about a topic that the teacher had set for them. Such an arrangement bears many similarities to the group oral exam—itself a relatively recent innovation in language testing—whose use has been successfully documented in Finland (Folland & Robertson, 1976), Israel (Reves, 1980; Shohamy, Reves, & Bejarano, 1986), Zambia (Hilsdon, 1991), Cyprus (Fulcher, 1996), Hong Kong (Morrison & Lee, 1985), and Italy (Lombardo, 1984, cited in Fulcher, 1996).

1.2 THE GROUP ORAL EXAM

In a group oral exam, examinees are assigned to groups of four or five students and are given a task to complete or a topic to discuss within a specified time limit. If students have lower second language (L2) proficiency, they may be allowed time beforehand to discuss the procedure in their native language, but once the formal test begins, all interactions are in the target language. A key feature of group oral exams is that, just as the instructor need not interact with the students in CMC interactions, examiners almost always remain in the background, silently observing the interactions of examinees.

According to Venugopal, the validity of a group oral is derived from replication of a "real life situation in the context of a discussion or chat between 3-5 individuals" (Venugopal, 1992, p. 48). Many other proponents have also argued in favor of the authenticity of the group oral, though Fulcher (1996) cautions that this claim has not yet been supported empirically. Several other advantages are claimed for group oral exams: increased efficiency (Berkoff, 1985); increased reliability of test scores (Folland & Robertson, 1976; Reves, 1991); high face validity (Berkoff, 1985); and a positive washback effect (Hilsdon, 1991). Few of these claimed advantages have yet been supported by empirical evidence. Hilsdon (1991) offers anecdotal evidence of a washback effect in secondary schools in Zambia. Fulcher (1996) used questionnaire and interview data from 47 Greek-speaking students in Cyprus to identify student reactions to three tasks, two involving face-to face interviews, and one involving group discussion. Compared to the other tasks, the group discussion test was perceived as being a more natural situation for conversation, provoked less anxiety among students, gave students more confidence to say what they wanted to, and was more preferable to students. Fulcher also estimated task difficulty using the Rasch partial credit model and found that the group discussion task was the easiest of the three tasks. In recent work, Berry (2000, cited in Swain 2001) has examined the relationship between extraversion and performance on a group oral test, finding that the level of introversion/extroversion demonstrated by other group members can influence the scores of an individual test-taker.

1. 3 LANGUAGE ABILITY AND TEST PERFORMANCE

The notion of performance has been mentioned several times so far but has yet to be either defined or related to underlying ability. Hymes' (1972a) re-interpretation and elaboration of Chomsky's (1965) performance-competence distinction to include usage rules has enormously influenced the models of language used by second language researchers and testers. One influential early model was Canale and Swain's (1980) model of communicative competence, which has been refined and developed in the past decade by Bachman (Bachman, 1990; Bachman & Palmer, 1996).

In the first iteration of his model, Bachman (1990) proposes a model of language ability containing two competencies: language competence and strategic competence. The former expands on and re-organizes Canale and Swain's components of language competence to reflect research from linguistics. Language competence is divided into two broad categories: organizational competence, which is used "in controlling the formal structure of language for producing or recognizing grammatically correct sentences, comprehending their propositional content, and ordering them to form texts" (Bachman, 1990, p. 87); and pragmatic competence, which is "the relationships between . . . signs and referents on the one hand, and the language *users* and the *context* of communication, on the other" (p. 89, italics in original). Strategic competence is a general ability by which individuals use their available resources efficiently; it also serves an explanatory role for differential performance of individuals with the same underlying competence since individuals may differ in their desire and flexibility to use their available resources. In a later elaboration of the model, Bachman and Palmer (1996) propose an interactional framework in which context interacts through strategic competence with personal characteristics such as language ability, personal characteristics, topical ("real-world") knowledge, and affective schemata. An important addition here is affective schemata, which provide a mechanism for describing how users previous emotional experiences of a context such as a test may influence their current response to a similar context.

One criticism of Bachman and Palmer's model of test performance is that although it provides a detailed model of how several psychological elements interact with each other to affect performance and acknowledges the potential for external factors to similarly influence language performance, it fails to provide an elaborate description of those external factors. McNamara (1996) and Skehan (1998) fill this gap. McNamara proposes a model of language test performance in which the translation of candidates' underlying ability into performance can be affected by both the task they are asked to perform and the interlocutors with whom they interact (i.e., by examiners or other test takers). The scores candidates receive for their performance are determined by the interaction between the performance itself on the one hand and the raters' personal characteristics and their interpretation of the rating criteria on the other hand. Skehan (1998) adapts and expands McNamara's model in two ways. First, he adds a component, *ability for use*, which serves to mediate the realization of underlying competencies in language performance. He also sub-divides the 'task' component of the model into two sub-categories which may affect performance differentially: task conditions and task qualities. The full model is described in greater detail in chapter 2 and can be seen in Figure 2.3.

The McNamara-Skehan model of oral test performance has several implications for this study, which will be discussed in greater detail in chapter 3. For now, they will be listed briefly:

1. The same raters should evaluate the computer-mediated and the group oral exams.
2. A common set of criteria should be used to evaluate performance on the two tests.

3. The relative weighting of criteria should be identical for both tests.
4. Group composition should be maintained across both tests. Students should not interact with different sets of students on the group oral exam and the computer-mediated test.
5. The tasks for each exam format should be as similar as possible.

1. 4 VALIDITY

The central question in this study asks whether individuals' performance on a computer-mediated communicative test allows us to make inferences about their speaking ability. The focus on inferences places this question at the heart of test validity, which Messick (1989) defines as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). In other words, the key issue in test validation is finding evidence that supports (i.e., validates) the inferences we make about candidates on the basis of the scores they are assigned by the assessment instruments we use. Validation, thus, involves both the testing of hypotheses (Cronbach, 1988; Landy, 1986)—since an inference is a form of hypothesis—and the construction of an argument (Messick, 1989; Shephard, 1993).

Messick's (1989) seminal article on test validity provides a detailed examination of the different methodologies and types of evidence that could be used in constructing a validation argument. However, a concern that the complexity of Messick's conceptualization of test validation deters test developers from conducting adequate validation studies has led several researchers to propose simplified frameworks for test validation (Kane, 1992; Shephard, 1993). One particularly promising framework is that of Kane (1992) who suggests that a validation argument is only as strong as the assumptions on which it rests. The aim of a validation study, therefore, should be to collect evidence about the weakest assumptions, that is, those which a study's potential audience might use to question the validity of the inferences and uses made from test scores.

Making inferences about oral language ability from a test that does not involve oral language production rests on a number of assumptions about the importance of pronunciation, the comparability of test scores, the similarity of test takers' attitudes toward the two tests, and the similarity of linguistic and interactional features in spite of the different modalities of production. The following section serves two purposes: (a) it discusses in detail how each of these assumptions might weaken the validity of the inferences made about speaking ability from the results of a computer-mediated test; and (b) as a result of the discussion in (a), it derives the specific research questions which will be answered by this study.

1.5.1 ASSUMPTION 1: THE ROLE OF PRONUNCIATION

Anyone who has listened to the speech of a second language learner with a heavy accent will be aware that pronunciation can influence the effectiveness of L2 (and indeed L1) communication. One implication of this is that any measure of oral language ability will be influenced by pronunciation, either explicitly by its inclusion as a criterion on the rating scale, or implicitly, by its effect on comprehensibility and, thus, on communication. How then, can a written test, which by definition does not provide any information about students' aural comprehensibility, be a valid measure of oral language ability?

The answer lies not in rejecting the importance of pronunciation to effective communicative but in examining whether pronunciation needs to be measured to obtain an accurate measure of speaking ability. If it can be shown that pronunciation correlates highly with either an individual criterion or the sum of scores on multiple criteria, then there may be no need to measure pronunciation since it would not influence the relative rankings of students. It should be noted that, a priori, this is probably the weakest assumption on which we could base inferences about spoken language from computer-mediated discourse because it relies on statistical predictions derived from group data for the interpretation of individual cases, which often produces flawed results. However, even with this caveat, the assumption should still be addressed in a validation study, which results in the first research question:

> Research Question 1: To what extent do measures of students' intelligibility in a group oral exam correlate with measures of other criteria on a computer-mediated communicative test?

1.5.2 ASSUMPTION 2: SIMILARITY OF SCORES

In a validation study of a semi-direct test of Hebrew, Shohamy et al (1989) showed that the test had high concurrent validity, that is, scores on the test correlated highly with scores on a direct oral test of Hebrew (the Hebrew OPI). Similarly, if a group oral exam and a computer-mediated communicative test are tapping a similar construct, one would expect students' scores on the two tests to be highly correlated. Thus, the second research question is:

> Research Question 2: To what extent do students achieve similar scores in the group oral exam and the computer-mediated communicative test?

1.5.3 ASSUMPTION 3: SIMILARITY OF LANGUAGE

Shohamy (1994) argues that comparisons between two productive tests should examine not only the correlation of scores on the two tests but also the language that they tend to solicit: "correlations *per se* cannot provide sufficient evidence that two tests measure the same language" (p. 99). In a follow up to the earlier (1989) comparative study of a semi-direct and a direct test of Hebrew described in section 1.5.2, Shohamy (1994) found that although the scores on the two tests correlated highly, there were significant differences in the language produced by examinees, many of which were attributable to the lack of an interlocutor in the semi-direct test.

Although both tests being compared in the present study involve interaction with interlocutors, they use different modalities (writing versus speech). The studies cited earlier in this chapter have suggested that the discourse of synchronous CMC may be a hybrid discourse, combining features of both written and spoken language. Whether this holds true for learners using CMC in a second language is not so clear. To date, two studies have directly compared the language production of a group of students in both face-to-face and CMC interaction (Kern, 1995; Warschauer, 1996).

Warschauer's (1996) experimental study involved 16 students from an ESL composition class at the University of Hawaii. Students discussed two questions, one in a face-to-face discussion and one electronically. Warschauer found that in the CMC discussions, students produced language that was more formal and more complex (in terms of type/token ratios and coordination index) than that produced in the face-to-face discussion. Participation rates were also more equal in the electronic discussion than in the face-to-face discussion. However, the study looked at a very narrow range of linguistic variables and only checked for statistically significant results for the variables of complexity. Kern's (1995) study used data from an intermediate level French class. Kern examined a greater range of linguistic variables than Warschauer and replicated Warschauer's findings that CMC results in more equal participation and the production of more complex language (defined by Kern in terms of morphosyntactic features and range of functions). However, Kern was also unable to check his results for statistical significance, which makes generalization impossible.

The studies by Kern and Warschauer are useful starting points for a comparison of the language produced by second language learners in electronic and face-to-face environments. However, both studies are flawed, and neither deals with language production under testing conditions when a different set of affective factors may be influencing student performance. Thus, the third research question is:

> Research Question 3: In what ways is the language produced by students on the computer-mediated test similar to or different from that produced on the group oral exam?

1.5.4 ASSUMPTION 4: SIMILARITY OF INTERACTIONS

In spite of the tendency by some researchers to refer to the language produced during CMC as 'written conversation' or, as Beauvois (1992) calls it, 'conversation in slow motion', relatively few studies have examined the discourse and interactional features of CMC in a second language. Pinto (1996) concluded that conversations between 15 ESL students during four 90-minute MOO sessions were not very fluent and lacked the give and take of regular face-to-face interaction; he failed, however, to consider that the large group size may have caused difficulty in following threads. In Pellettieri (2000), patterns of interaction between intermediate-level Spanish students in a chat room were found to be similar to those that occur during oral conversation in language classrooms. The negotiation of meaning that occurred between students made the language input more comprehensible and, when combined with corrective feedback, caused learners to attend to form and to modify output. Other studies by Smith (2001) and Blake (2000) have also found learners engaged in significant amounts of negotiation of meaning on several different language learning tasks, but Smith's results suggest that these negotiations included closing elements not predicted by the most popular interactionist models. Sotillo (2000) found significant differences in learner's production in asynchronous versus synchronous CMC, with the latter producing syntactically simpler sentences with discourse features characteristic of spoken language. Davis and Thiede's analysis (2000) found that over time, non-native speakers in on-line discussions modified their writing style to match their interlocutor's status. Gonzalez-Bueno's (1998) study of the use of e-mails as electronic dialogue journals in low-level Spanish classes found enhanced student participation and discourse that had a conversational tone.

In summary, the literature suggests that negotiation of meaning takes place in CMC environments, but the nature of that negotiation is unclear. Given this uncertainty and the problem outlined in the previous section—that the context of testing may influence performance in ways that are not found in regular language classrooms—we cannot predict the ways in which students' interactions may be similar or different during the two tests. However, any use of CMC performance to infer oral abilities will involve the assumption that the two test modes produce similar interactions. Thus the fourth research question is:

Research Question 4: What are the differences in students' interactions on the two tests?

1.5.5 ASSUMPTION 5: STUDENT RESPONSES TO THE TEST

The final assumption is that learners perceive the two tests in similar ways. The importance of this assumption to the validation argument being made in this study may not be immediately obvious. It is not a question of whether learners perceived the computer-mediated test to lack face validity because of its use to make inferences about oral ability based on essentially written interaction; such a use had not been articulated to the learners in this study, so it is unlikely to have influenced their attitudes towards what was for them basically a classroom test which incorporated the type of computer-mediated activity they had practiced all semester. Instead, this assumption implies that students saw both tests as equally challenging, as providing them with equal opportunities to perform to the best of their abilities, and as invoking similar affective responses. What this assumption is examining is whether any differential performance on the two tests can be attributed to affective factors. The more the students' perceptions of the computer-mediated and face-to-face tests were similar, the more likely it is that they approached the tests with a similar attitude and tried to perform equally well on both tests. Thus, the fifth research question is:

Research Question 5: What are students' perceptions of the two modes of testing?

1.6 SUMMARY OF RESEARCH QUESTIONS

The present study explores the validity of making inferences about spoken language ability from performance on a written computer-mediated communicative test. In developing this validation argument, the study addresses five research questions (RQ).

RQ1: To what extent do measures of students' intelligibility in a group oral exam correlate with measures of other criteria on a computer-mediated communicative test?

RQ2: To what extent do students achieve similar scores in the group oral exam and the computer-mediated communicative test?

RQ3: In what ways is the language produced by students on the computer-mediated test similar to or different from that produced on the group oral exam?

RQ4: What are the differences in students' interactions on the two tests?

RQ5: What are students' perceptions of the two modes of testing?

1.7 ORGANIZATION OF THE DISSERTATION

Chapter 2 provides a detailed description of the theoretical and empirical literature that inform this study. Given the centrality of the concept of communicative competence and its relationship to test performance and to language teaching and assessment, the chapter begins with a detailed examination of models of communicative competence and of test performance. The following section investigates the key concept of test validity and discusses principles for the design of validation studies. Since an important issue in this study is the relative similarity of CMC discourse to spoken and/or written discourse, the relationship between written and spoken language and empirical studies of first language CMC discourse are explored next. Finally, the chapter discusses CMC in the second language classroom, examining the rationales presented for its use and presenting a critical summary of classroom-based research on CMC.

Chapter 3 restates the research questions and provides a detailed description of the data collection, coding, and analysis methods for this study. Chapter 4 presents the results obtained for each of the research questions and includes a preliminary discussion of those results. Chapter 5 recaps the most important findings reported in chapter 4, discusses and interprets those results in detail, and suggests some general implications these findings may have for second language assessment. Finally, Chapter 5 elaborates on the limitations of the present study and suggests potential areas that this study opens for future research into computer-mediated communicative testing in the foreign/second language classroom.

CHAPTER 2

REVIEW OF LITERATURE

2.1 INTRODUCTION

Chapter 1 established that the goal of this study was to answer the following overarching question: Can one make inferences about students' speaking ability based on their performance on a computer-mediated (i.e., written) test? The purpose of this chapter is to present the theoretical and empirical work which informs this study. Section 2.2 begins with a critical examination, from the perspective of Hymes' (1972a) notion of *ability for use*, of influential models of communicative competence proposed by Canale and Swain (1980) and more recently by Bachman (Bachman, 1990; Bachman & Palmer, 1996). This section also discusses models which describe contextual and internal factors that mediate and influence the translation of communicative competence into performance on language tests (McNamara, 1996; Skehan, 1998) and derives implications from them for the design of this study.

Section 2.3 examines the notion of validity, tracing its development from a tripartite concept established through correlational studies to the current view as a unitary construct which encapsulates several facets. With regard to the latter, Messick's (1989) seminal article on validity is discussed in some detail, both to exemplify the current approach and to determine why many current practitioners are unable to translate Messick's theoretical basis for validation studies into practice. As a result, this study will use Kane's (1992) alternative model for test validation studies, which will be presented at the end of the section.

Since one of the research questions of this study requires a comparison of the oral language produced in a face-to-face test and the written language produced in an electronic test, section 2.4 presents an overview of literature pertaining to differences between writing and speaking. Writing and speaking have moved from representing opposite poles of a simple dichotomy (O'Donnell, 1974; Olson, 1977) to being points on a continuum (Chafe, 1982; Tannen, 1982) to the current view, developed from research by Biber (1988), which sees them as being multi-dimensional constructs which vary and overlap along several continua. The similarities and differences of spoken and written language will be examined in terms of the influence of mode and context. Section 2.5 will extend this discussion to the context of CMC discourse, presenting several L1 studies which suggest that CMC combines features of both spoken and written language.

The use of CMC in the second language classroom will be examined in sections 2.6 and 2.7. Section 2.6 examines the many justifications for classroom use of CMC that are found in the literature. These will be examined in three groups: those drawing on interactionist approaches to SLA, those drawing on sociocultural approaches, and those dealing with affective concerns. Section 2.7 presents research on classroom-based CMC that has investigated students' development of cultural understanding, the transfer of skills from written CMC interaction to oral proficiency, the effect of individual traits on CMC participation, the textual and interactional features of CMC discourse, and the creation of the collaborative learning communities favored in sociocultural approaches to SLA.

2.2 COMMUNICATIVE COMPETENCE AND ORAL PERFORMANCE

Although developed to describe first language ability, Hymes' (1972a) model of communicative competence has been a highly influential—some would say the most influential (McNamara, 1996)—model for communicative language tests. Hymes re-interpreted Chomsky's (1965) performance-competence distinction, which, he felt, missed an important social dimension. Thus, he argued that "there are rules of language use without which the rules of grammar would be useless" (p. 278). Moreover, competence is not simply knowledge of grammatical and usage rules: It also includes a systemic component, *ability for use*, which models underlying capacity to translate knowledge (of grammatical and usage rules) into actual use (which he refers to as performance). *Ability for use* is a very broad concept, including a number of non-cognitive factors such as motivation and, following Goffman (1967), "capacities in interaction such as courage, gameness, gallantry, composure, presence of mind, dignity, stage confidence" (p. 283). As we shall see, it is precisely these affective and/or volitional factors with which second language models have had the greatest difficulty, with McNamara (1996), for example, referring to *ability for use* as a 'Pandora's Box' which few researchers have opened successfully.

One of the most influential of the early models based on Hymes' work was Canale and Swain (1980) who distinguished between communicative competence and performance. Although Canale and Swain's use of communicative competence may suggest that they are discussing the same concept as Hymes, in fact, their definition is fundamentally different since they have reduced it solely to the knowledge component of

Hymes model, with no place for *ability for use* (McNamara, 1996). Thus, for Canale and Swain, communicative competence has three components: grammatical competence (knowledge of lexical items, rules of morphology, syntax, sentence-grammar semantics, and phonology); sociolinguistic competence (sociocultural rules of use and rules of discourse); and strategic competence (compensatory verbal and non-verbal communication strategies). Their model implicitly includes elements of *ability for use* in strategic competence, as McNamara (1996) notes: "surely 'coping' is an aspect of performance, involving general reasoning or problem-solving capacities, as well as imaginativeness, and also possible personality factors—preparedness to take risks, versatility, and adaptability" (p. 62). However, Canale and Swain clearly state their skepticism about explicitly modeling *ability for use*: "We doubt that there is any theory of human action that can adequately explicate 'ability for use' and support principles of syllabus design intended to reflect this notion" (p. 7).

Bachman's (1990) definition of communicative language ability as "both knowledge, or competence, and the capacity for implementing, or executing that competence in contextualized communicative language use" (p. 84) does not show the same skepticism. His model differs significantly from Canale and Swain's. A new component, psychophysiological mechanisms, is added, and language competence is separated from strategic competence, with the latter being presented not as a 'coping' mechanism employed to compensate for linguistic deficiencies but "as a general ability, which enables an individual to make the most effective use of available abilities in carrying out a task, whether the task be related to communicative language use or to non-verbal tasks" (p. 106). This separation of language and strategic competence is an important development because it allows Bachman to describe a mechanism by which people with the same language competence may differ in their language performance. Strategic competence has three components: an assessment component, in which the individual identifies needed and available resources to realize a linguistic goal; a planning component, where relevant items are retrieved and a plan is formulated; and an execution component, where the individual draws on relevant psychophysiological mechanisms to implement the plan. Differences in performance may be due to individuals' "willingness to exploit what they know and their flexibility in doing so" (p. 105). Clearly, strategic competence has elements of Hymes' *ability for use*, but, though the previous quote seems to allow for the effect of volitional elements, the components of strategic competence are mostly limited to cognitive factors (McNamara, 1996, p. 71). Thus, it is more restricted than Hymes' *ability for use*.

In addition to its redefinition of strategic competence, Bachman's model also considerably expands on and re-organizes the components of language competence to reflect research from linguistics (see Figure 2.1). He divides language competence into two broad categories: organizational competence and pragmatic competence. Organizational competence refers to "those abilities involved in controlling the formal structure of language for producing or recognizing grammatically correct sentences, comprehending their propositional content, and ordering them to form texts" (p. 87); it is sub-divided into grammatical competence and textual competence. Pragmatic competence describes "the relationships between . . . signs and referents on the one hand,

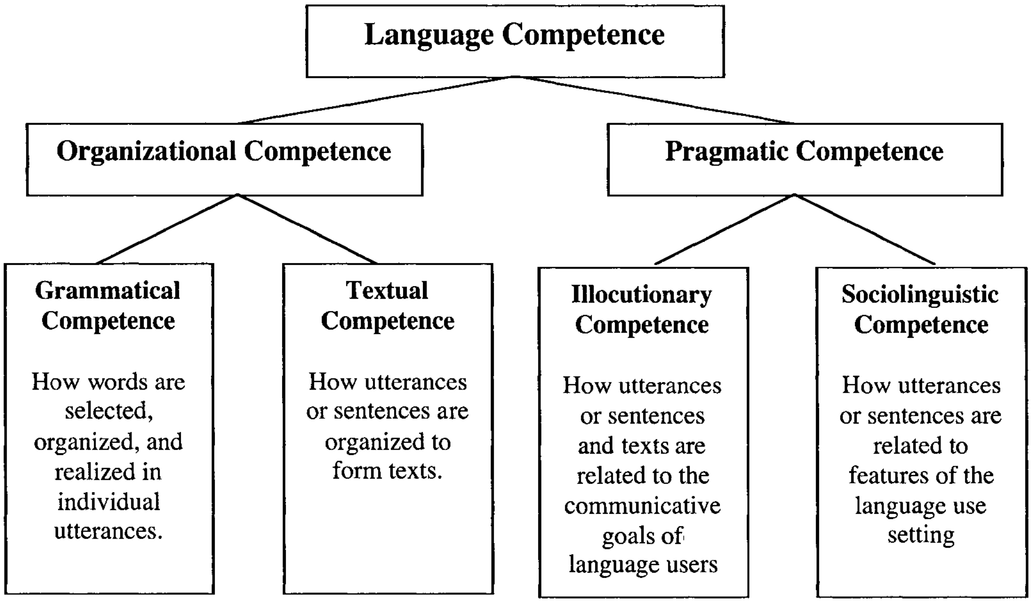
Language Competence
Organizational Competence
Pragmatic Competence
Grammatical Competence
How words are selected, organized, and realized in individual utterances.
Textual Competence
How utterances or sentences are organized to form texts.
Illocutionary Competence
How utterances or sentences and texts are related to the communicative goals of language users
Sociolinguistic Competence
How utterances or sentences are related to features of the language use setting

Figure 2.1: Components of Language Knowledge. From *Fundamental Considerations in Language Testing* (p. 87), by Lyle F. Bachman, 1990, Oxford: Oxford University Press. © Lyle F. Bachman 1990. Adapted with permission of Oxford University Press.

and the language *users* and the *context* of communication, on the other" (p. 89, italics in original); it is divided into illocutionary competence (defined in terms of language functions) and sociolinguistic competence. The greater specificity of this model in regard to language competence is undoubtedly a boon to test design and the research agenda.

Bachman and Palmer (1996) build on the earlier work of Bachman (1990). Their model of language ability is, bar some nominal changes, essentially unchanged from Bachman (1990), but the way it works has changed (see Figure 2.2). Bachman and Palmer propose an interactional framework in which context interacts with personal characteristics such as language ability, topical ("real-world") knowledge, and affective schemata. The key new element here is affective schemata, which, according to Bachman and Palmer:

> provide the basis on which users assess, consciously or unconsciously, the characteristics of the language use task and its setting in terms of past emotional experiences in similar contexts. The affective schemata, in combination with the characteristics of the particular task, determine, to a large extent, the language user's affective response to the task, and can either facilitate or limit the flexibility with which he responds in a given context. (p. 65)

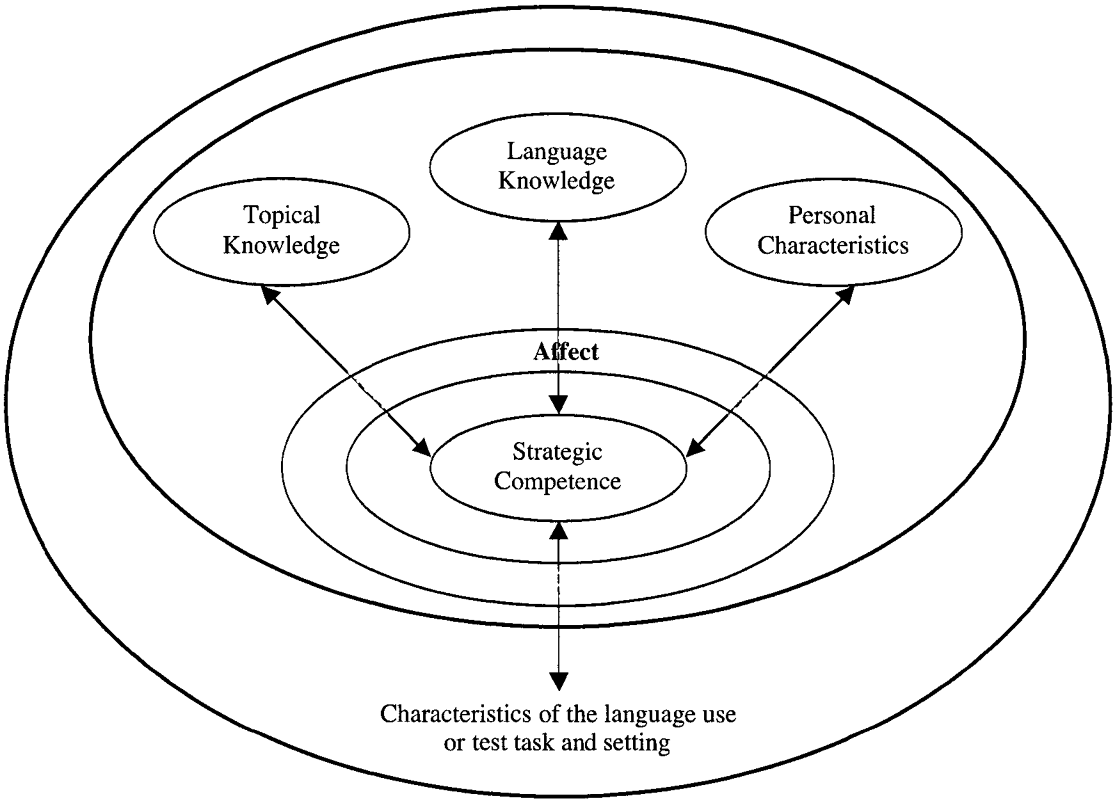

Figure 2.2: Bachman and Palmer's Components of Language Test Performance. From *Language Testing in Practice* (p. 63) by Lyle Bachman & Adrian Palmer, Oxford: Oxford University Press. © Lyle Bachman & Adrian Palmer 1996. Reproduced with permission of Oxford University Press.

McNamara notes that this is the first time that "an attempt has been made to deal explicitly. . . with the aspect of *ability for use* which relates to affective or volitional factors" (1996, p. 74). However, he also comments that Bachman and Palmer have problems in exploring these factors, especially when it comes to deciding when it would be appropriate to include affective responses as part of test content.

One criticism of all three models discussed so far is their static view of communication (McNamara, 1996; McNamara, 1997). Although not the first to worry about the lack of interaction in models of communicative competence (see, for example, Courchene & de Bagheera, 1985). McNamara (1996) follows Kenyon (1992) in creating a model which addresses this issue. His model accepts the view of underlying language ability proposed in Bachman (1990) and Bachman and Palmer (1996) but places it within a framework which integrates the potential role of several other factors on oral test performance. Among these are the task, which provides the need for performance, and the raters who, equipped with rating scales and criteria, judge the performance. The model also allows for the influence on performance of other interlocutors such as examiners or other interactants (the consequences of this aspect of the model will be discussed in more detail later). The inclusion of these factors in McNamara's model means that a test score is no longer viewed solely as the result of the candidate's underlying competence (Skehan, 1998) but reflects the interaction of multiple influences.

While Skehan (1998) sees this model as an advance, he argues that the model could be improved and expanded in two ways. The first change is to specifically incorporate Hymes' *ability for use* as the mechanism by which "a second language performer adjusts to performance conditions by trying to allocate attention in appropriate ways" (p. 168) such as choosing between fluency, accuracy, and complexity. In deciding among these competing goals, Skehan argues, *ability for use* plays a mediating role between underlying competence and performance. The second change Skehan suggests arises as a result of the centrality that he assigns to tasks within the testing context. He argues that performance and the resulting judgments about proficiency are strongly influenced by the qualities, types, and characteristics of the tasks that test takers are asked to complete and also by the conditions in which those tasks as implemented. Thus, although the McNamara-Kenyon model includes task as one of the factors influencing performance, Skehan proposes a more detailed model which makes explicit the aspects of a task that will influence performance: task conditions and task qualities. This model is shown in Figure 2.3.

The Kenyon-McNamara model and Skehan's extension of it represent a major advance in that they provide a model of the interaction between second language testing environments and underlying competencies that can be explored systematically through research and, where necessary, revised. However, in terms of assigning meaning to test scores, the inclusion of interlocutors in the model raises some important questions about test score interpretation. McNamara (1997) provides a provocative discussion of the impact of adding an interactive component to models of language ability. He notes that 'interaction' can be thought of in two ways: psychologically, as "mental activity within an individual"; and socially/behaviorally, as co-constructed "joint behavior between individuals" (p. 447). Current models, he argues, have stressed the former over the latter

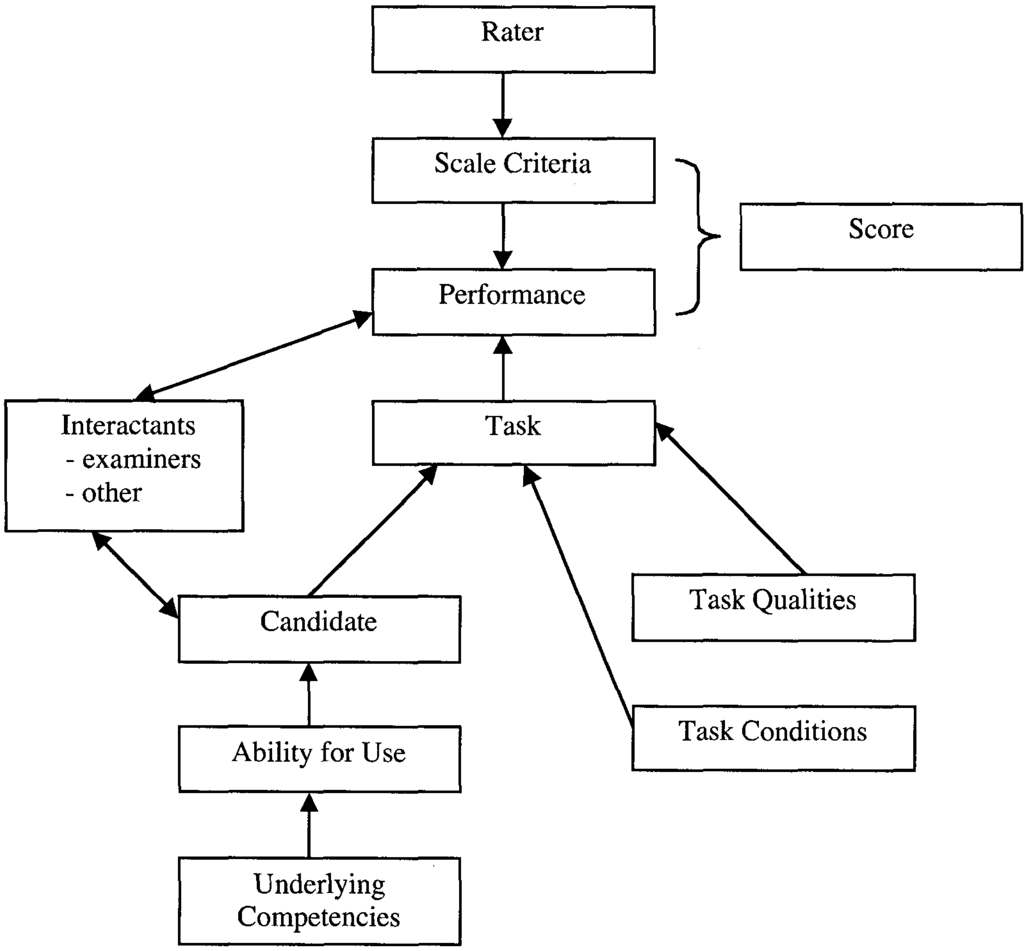

Figure 2.3: Skehan's Model of Oral Test Performance. From *A Cognitive Approach to Language Learning* (p.172), by Peter Skehan, Oxford: Oxford University Press. © Oxford University Press 1998. Reproduced by permission of Oxford University Press.

with several unfortunate consequences: placing "the brunt of the responsibility for the performance" on the candidate (p. 453); judging the candidate according to different standards than would be used for native speakers; and failing to view judgment of performance as "an inherently social act" (p. 453) involving "an interplay of socially derived understandings of the nature and purpose of the activity on the part of test developers, interlocutors, and raters" (p. 458).

McNamara (1997) calls for models of language ability to add the dimension of social interaction, but he is keenly aware of the implications of doing so. Some implications are obviously beneficial to the task of test design. For example, closer examination of naturally occurring discourse and target language situations may help us identify (a) ways in which the standards proposed by language testers differ from those that apply in reality and (b) the extent to which the characteristics of simulated interactions match those of the target language situations. However, other implications have the potential to complicate score interpretation enormously:

> If the performance is co-constructed, how can we build the interlocutor into our assessment of an individual's communicative abilities, when the type of potential interlocutor is so variable? . . . . Further, if communication is a joint responsibility, then who are we to blame if communication goes awry?" (McNamara, 1997, pp. 458-459)

Given these complications, the difficulties experienced by Bachman and Palmer, and the skepticism of Canale and Swain, one may wonder why we should include *ability for use* in models of second language ability, let alone add a dynamic, social dimension. What, if anything, do we gain by doing so? The answer is that understanding the contribution of *ability for use* and aspects of the testing context to second language learner's performance permit more valid interpretations of the scores that are assigned to that performance. The next section of the chapter discusses the concept of validity in testing in greater detail.

2.3 VALIDITY IN LANGUAGE TESTING

In her discussion of validity, Chapelle (1999) points out the importance of this concept to the language testing field:

> The definition of validity affects all language test users because accepted practices of test validation are critical to decisions about what constitutes a good test for a particular situation. In other words, assumptions about validity and the process of validation underlie assertions about the value of a particular type of test. (p. 254)

Much of the literature on validation in language testing draws upon parallel work in the field of educational psychology, where validity is "pre-eminent among the various psychometric concepts" (Angoff, 1988, p. 19). As the latter field has developed new perspectives on validity and validation practices, language testing researchers have appropriated and adapted them to their own context. Thus, this section presents an overview of the conceptual development of validity and its implications for test validation practices in language testing.

Early definitions of validity saw it as a property of tests. A typical perspective is that of Lado (1961) who wrote: "Does the test measure what it is supposed to measure? If it does it is a valid test" (p. 30). This view of validity as inherent in a test, however, came to be viewed as "naive" as researchers such as Cronbach and Meehl (1955) realized that they were validating "a principle for making inferences" (p. 297). Thus, in the first edition of *Educational Measurement,* Cureton (1951) explicitly linked test validity to the purpose(s) of a test. In the second edition of *Educational Measurement,* Cronbach (1971) distinguished between a narrow and broad sense of validity. In the narrow sense, researchers and test developers were interested in “the process of examining the accuracy of a specific prediction or inference made from a test score” while in the broad sense, “validation examines the soundness of all the interpretations of a test—descriptive and explanatory interpretations as well as situation-bound predictions” (p. 443). Thus for Cronbach, validation had much in common with the process of evaluating any scientific theory. Cronbach’s views were echoed in the language assessment literature by Palmer and Groot (1981) who also point out the potential for confusion in the literature where the word ‘test’ is often used to refer to a combination of the test itself and the inferences made from that test so that the notion of “the ‘validity of a test’ can then have meaning—as long as the distinction between the two uses is kept clearly in mind” (p. 1).

The last 50 years have also witnessed a shift in the categorization of types of validity. While the majority of studies in the 1940s could be categorized as predictive in nature (Angoff, 1988), an awareness of “the chaotic state of test construction procedures” (Anastasi, 1986, p. 2)—which included the validation of tests—led the American Psychological Association to publish *The Technical Recommendations for Psychological Tests and Diagnostic Procedures* in 1954 (Anastasi, 1986) which classified validity into four types: content, predictive, concurrent, and construct. In a later article (Cronbach & Meehl, 1955, p. 282), two of the authors of the *Technical Recommendations* define these validity types as follows: content validity is the extent to which test items represent an adequate sample of the universe that an investigator wishes to examine; construct validity is involved whenever a test must use an indirect measure of an attribute, and it identifies the extent to which that measure operationalizes the attribute or quality of interest; predictive validity and concurrent validity both examine the correlation between scores on the test and scores on an independent criterion, but they differ in that predictive validity is concerned with a criterion to be measured in the future while concurrent validity measures a criterion measured at the same time. For Cronbach and Meehl, predictive and concurrent validity can be considered as "criterion-oriented validation procedures" (p. 281), a view which was integrated into later revisions of the *Technical Recommendations*, leaving a tripartite division of validity which was the orthodoxy for more than 20 years among educational psychologists and language assessment researchers (for a representative sample of the latter, see Lado, 1961; Oller, 1979; Palmer & Groot 1981).

During this time, validation studies were mostly correlational. While this is to be expected for studies of criterion-referenced validity which, by definition, is underpinned by a correlational relationship between a predictor and a criterion variable, it was also true for investigations of construct validity. For example, included in the types of evidence that Cronbach and Meehl (1955) identify as relevant to construct validity are "interitem correlations, intertest correlations, test-criterion correlations, [and] studies of stability over time" (p. 300), that is, reliability, which is itself typically measured through correlations. In a similar fashion, Campbell and Fiske's (1959) model of construct validation uses a matrix in which multiple trait-method units are correlated (a trait-unit is a measurement of a single trait using a single method). Underlying their model are two assumptions: (a) construct validity depends both upon convergent validation—high correlation between independent attempts to measure the same trait—and discriminant validation—low correlation between measures of traits that theory holds should be different; and (b) multiple measures of multiple traits allow researchers to identify the "relative contributions of trait and method variance" (p. 81). While correlational studies were the predominant mode for collecting validation evidence for many years, their domination has been challenged more recently as measurement specialists have argued that validation is a form of hypothesis testing (Kane, 1982; Grotjahn, 1986; Landy, 1986; Cronbach, 1988) and, thus, can be investigated using any of the methods and evidence types traditionally utilized in investigating scientific hypotheses (Landy, 1986). One of the most complete analyses of possible sources of evidence is found in Messick (1989), which will be discussed in greater detail later.

In the 1980s, conceptualizations of the taxonomy of validity in the language assessment and educational psychology fields began to diverge. Language assessment researchers were concerned with the addition of concepts such as "affect, [which is] particularly the extent to which our test causes undue anxiety" (Madsen, 1983, p. 179) and response validity, that is, "the extent to which examinees responded in the manner expected by test developers" (Henning, 1987, p. 96). A growing interest in performance assessment with its related concern of authenticity led to a debate in the literature about face validity, a concept which briefly gained favor, but which Stevenson criticizes for its potential to allow "the perilous jump from face validity to construct validity: we think it's valid, therefore it is. The critical mediating role of—above all—criterion-related validity and validation is passed over" (Stevenson, 1985, p. 46).

In educational psychology, however, researchers criticized the implementation of the tripartite division of validity in empirical studies and, by extension, the usefulness of that division itself. In separate papers, Landy (1986) and Anastasi (1986) argued that the tripartite division became "a small and fixed set of validity models" (Landy 1986, p. 1184) which served in practice to limit the types of inquiries conducted—and, thus, the range of evidence collected—in validation studies. Both authors advocate Messick's view of validity as a unitary construct with multiple facets, a framework for which he had argued forcefully in a number of articles throughout the 1980's (see, for example, Messick, 1980; Messick, 1988), culminating in his seminal article in the third edition of *Educational Measurement* (Messick, 1989). The influence of Messick's 1989 article in the field of measurement is universally acknowledged—Shephard, for example, hails it as a "landmark treatise" (1997, p. 5)—and Messick's perspective on validity has been adopted by a number of language testing researchers (Chapelle & Douglas, 1993; Cumming, 1996; Kunnan, 1998). For instance, both Cumming (1996) and Kunnan (1998) use Messick's progressive framework as a taxonomy for classifying the studies in their respective volumes. The most influential adoption of Messick's framework is found in Bachman (1990) whose chapter on validation replicates Messick's conceptualization of validity and discusses how validation studies of language may utilize the sorts of evidence proposed by Messick. Given the importance of Messick's article to both the educational psychology and language testing fields, it is worth examining in greater detail.

Messick's article reiterates two validation concepts that had come to be widely-accepted by assessment researchers: that validity is a property not of tests themselves but of the uses and inferences made from those tests; and that since inferences are hypotheses, validation of those inferences involves hypothesis testing. Thus, Messick defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 13). Evidence is both the data that is collected about the test and "the rationale or arguments that cement those facts into a justification of test-score inferences" (pp. 15-16). He identifies six possible sources of evidence, which he views as supplementary rather than alternative:

> We can look at the content of the test in relation to the content of the domain of reference. We can probe the ways in which individuals respond to the tasks or items. We can examine relationships among responses to the tasks, items, or parts of the test, that is the internal structure of test responses. We can survey relationships of the test scores with other measures and background variables, that is the test's external structure. We can investigate differences in these test processes and structures over time, across groups and settings, and in response to experimental interventions—such as instructional or therapeutic treatment and manipulation of content, task requirement, or motivational conditions. Finally, we can trace the social consequences of interpreting and using the test scores in particular ways, scrutinizing not only the intended outcomes but also unintended side effects. (p. 16)

Messick rejects the traditional approach in which some of these forms of evidence have been seen as validity types in their own right. Instead, he defines validity as "a unified though faceted concept" (p. 14) and proposes a validity framework which integrates examination of "value implications and social consequences" and in which content- and criterion-related evidence play an "important though subsidiary role" (p. 20) in construct validation.

Messick's validity framework is a fourfold classification (see Figure 2.4) represented by a matrix constructed by the intersection of two facets: one facet representing the source of a test's justification, which may have either an evidential or a consequential basis; and a second facet representing the function or outcome of testing, that is, test interpretation and test use. The matrix is progressive because construct validity appears in each of the cells. Thus the evidential basis of test interpretation (top left cell) is construct validity itself. The consequential basis of test interpretation (bottom left cell) refers to the value connotations inherent in constructs: "the evaluative overtones of the construct labels themselves; the value connotations of the broader theories or nomological networks in which constructs are embedded; and the value implications of still broader ideologies about the nature of humankind, society, and science that color our manner of perceiving and proceeding" (p. 59). The evidential basis of test use (top right cell) combines general evidence for construct validity with specific evidence concerning the relevance of score interpretation in a specific context and for a specific purpose. Finally, the consequential basis of test use (bottom right cell) examines the intended and unintended individual, institutional, societal and systemic effects of test interpretation in order to determine "whether the proposed testing should serve as means to the intended end" (p. 84).

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance/utility
CONSEQUENTIAL BASIS	Construct validity + Value implications	Construct validity + Relevance/utility + Social consequences

Figure 2.4: Messick's Validity Framework (adapted from Messick 1989, pp. 20-21)

Messick examines each of these facets systematically, presenting a rationale for inclusion of each facet in the framework as well as multiple methodologies for collecting evidence about the facet. While educational measurement specialists have, in general, accepted the thrust of Messick's definitions of validity and validation, his article has sparked discussion regarding several aspects of his validation model: What constitutes appropriate evidence for validation of alternative assessments?; is investigation of social consequences an integral part of test validation?; and should his proposed test validation process be simplified to improve the adequacy of validation arguments? The following paragraphs examine each of these issues in turn.

The validity of alternative assessments has been challenged by research which uses traditional criteria such as reliability, efficiency, and the year-to-year comparability of assessments. While accepting in principle the value of these criteria, proponents of alternative assessment such as Linn, Baker and Dunbar (1991) argue that the traditional criteria for validity "should not be the only, or even the primary, criteria in judging the quality and usefulness of an assessment" (p. 16). Instead, they argue for a broader view of validity which includes eight criteria: a greater emphasis on the consequences of testing; an examination of fairness issues; a study of generalizability across tasks and raters; analysis of the cognitive complexity of assessment tasks; the quality of test content; the breadth of content coverage; the meaningfulness of tests to students; and the cost and efficiency of test administration.

A slightly different perspective is offered by Fredericksen and Collins (1989) who are more concerned with the consequences of assessment. They propose that tests should be designed to enhance systemic validity, which they define as an "extension of the notion of construct validity to take into account the effects of instructional changes brought about by the introduction of the test into an educational system" (p. 27). They argue that tests with the qualities of directness, scope (covering all the knowledge, skills, and strategies required), reliability, and transparency to students will be systemically valid because "instruction that improves the test score will also have improved performance on the extended task and the expression of the cognitive skills within the task context" (p. 29).

Messick (1994) takes issue with both of these approaches, which he characterizes as being "consistent with but less extensive than . . . general validity standards" (p. 13). He argues that the authors of these articles present a definition of validity that may cause the omission of important types of validity evidence. For example, he describes systemic validity as a special type of the criterion of social consequences, but criticizes Fredericksen and Collin's implied assumption that other aspects of the educational system beyond classroom instruction are working well, an assumption which should, he argues, be investigated and established empirically. Similarly, Messick sees the concern for issues of authenticity and directness not in terms of the specialized criteria presented by these two articles, but in terms of "the nature of the evidence accrued to counter the two major threats to construct validity, namely, construct under-representation (which jeopardizes authenticity) and construct-irrelevant variance (which jeopardizes directness)" (p. 14). This evidence, he argues, can only come from the general validity criteria outlined in his earlier work (Messick 1989); none of these criteria should be ignored.

The second aspect of Messick's framework that has provoked much discussion is consequential validity. The most extreme of Messick's critics is Popham (1997) for whom the inclusion of social consequences in the concept of validity "will not only muddy the validity waters for most educators, it may actually lead to less attention to the intended and unintended consequences of test use" (p. 13). Instead, Popham argues that investigations of the social consequences of testing—which he supports—should be conducted separately from test validation, which should only concern itself with examining the inferences derived from tests rather than the uses of those scores. However, such a divorce between inference and use is mistaken since it opens the door to both the misuse of tests and to the realization of Shephard's (1997) critique of Messick's segmented presentation of validity, namely, that test developers will focus mostly on construct validation leaving investigation of the social consequences to "moral philosophers and the politically correct" (p. 6). In a similar vein to Popham, concern about the confusion surrounding test bias led Cole and Moss (1989) to distinguish between *validity*, which answers the question "*can* a test be used for a given purpose?", and *extra-validity*, whose concern is "*should* a test be used for that purpose?" In a later article (Moss, 1992), however, one of the authors notes that "since then, we have expanded our definition of validity to include the consequential component, in part because we were concerned that excluding consideration of consequences from the definition of validity risks diminishing its importance" (p. 235).

Even though Popham's calls for a simplified concept of validity are misplaced, there is certainly a stronger argument to be made for a less complex approach to the validation process. In his discussion of the causes of the gap between validity theory and practice, Shephard (1993) suggests that the often inadequate validity evidence found in studies may result from a lack of understanding of "the integrative nature of construct validation", from the perception that validation is a process whose complexity and on-going nature mean that it can never be fully realized, or from a lack of suitable examples of complete validation studies (p. 407). As a result, Shephard calls for a simpler validation model which would allow researchers to identify the key validity questions to be answered by their study: "[A model] that clarifies which validity questions must be answered to defend a test use and which are academic requirements that go beyond the immediate, urgent questions" (p. 407). To satisfy this requirement, Shephard argues in favor of the approach developed by Cronbach (1988, 1989) and extended by Kane (1992).

Cronbach (1988) sees validation as a clarification, through persuasive argument, of a test's meaning for a particular audience. Central to the development of a validity argument is the talent of being devil's advocate by identifying and addressing those aspects of your argument that will be seen as weaknesses from the perspective of your audience. Kane (1992) adopts a similar approach. He views validation as the development of interpretive arguments about tests. Such arguments may be evaluated in terms of their coherence, their clarity, and the plausibility of the assumptions which underlie them. However, like Cronbach, Kane believes that "validity evidence is most effective when it addresses the weakest part of the interpretive argument" (p. 528). Typically, he argues, weak assumptions constitute the most serious problem for interpretive arguments:

> An assumption can be questioned because of existing evidence indicating that it may not be true, because of plausible alternative interpretations that deny the assumption, because of specific objections raised by critics, or simply because of a lack of supporting evidence. (p. 530)

Thus, Kane argues that validation should concern itself not with providing evidence for highly plausible assumptions but with investigation of the most questionable assumptions, those to which a validation argument is most susceptible to attack and refutation.

The literature reviewed in this section suggest that, even a study such as this one, which uses the simplified validation framework outlined by Kane, must create a potentially complex and detailed argument which considers both the theoretical rationale and the empirical evidence supporting the assumptions on which test interpretations rest. Considering a single type of evidence, such as correlations between test scores, will not provide sufficient evidence for the validity of using scores on computer-mediated tests as indicators of oral language ability. Instead evidence regarding all the major assumptions must be collected, examined, and combined with theoretical rationales into a meaningful argument which results in the justification or the rejection of test-score inferences.

2.4 ORAL AND WRITTEN LANGUAGE

Kane (1992) identifies several types of inferences that may appear in an interpretive argument, one of which is extrapolation from the observed behavior to another type of behavior. This study explores the extent to which performance in the context of written computer-mediated communication can be extrapolated to performance in the context of oral interaction. However, before making such an extrapolation, the relationship of written to oral language and the extent to which linguists view written and oral language as discrete constructs must be explored. This section, thus, reviews the increasingly complex ways in which linguists have examined oral versus written language.

The view of some early researchers (for example, O'Donnell, 1974; Olson, 1977) that the relationship between written and spoken language was a dichotomous one was challenged by researchers such as Tannen (1982) and Chafe (1982) who argued that rather than a dichotomy, there was, in fact, an oral/literate continuum. Thus Chafe (1982), in his discussion of the differences between two forms of discourse—the informal spoken language of dinnertime conversations and the formal written language of academic papers—notes that his "seemingly categorical statements about spoken and written language apply in fact to extremes on a continuum" and that the differences he describes are for samples which are "maximally differentiated" (p. 49).

In a similar way, Tannen (1982) argues that the strategies of involvement and content, which had been associated with oral and written language respectively, "are not limited to orality vs. literacy, and certainly not to spoken vs. written language, but rather can be seen to interplay in spoken and written discourse in various settings" (p. 4). She provides two types of evidence for this assertion. The first comes from a series of cross-cultural studies which examined Greek and American story-telling practices and the use of formulaic language. A common thread to these studies was the extent to which the two groups responded in culturally conventionalized ways. Americans tended to eschew formulaic language and focus on content (i.e., simply telling the events of the story) while the Greeks tended to favor formulaic language and to focus on interpersonal involvement by drawing upon their own experiences and interpreting the story.

The second type of evidence comes from her analysis of conversational story-telling at a Thanksgiving dinner. Tannen notes that the point (i.e., the speaker's evaluation) of the story can be communicated either internally—through the use of paralinguistic or non-verbal cues—or externally—through explicit statement—and that literacy tends to focus on external evaluation while non-literate communication depends more on features of internal evaluation. In the stories told over the Thanksgiving dinner table, she found a marked difference in tendencies towards internal and external evaluation, with New Yorkers of Jewish background relying on internal evaluation and other participants preferring external evaluation. Thus, she concludes "individuals and groups can make use of strategies that build on interpersonal involvement and make maximal use of paralinguistic and prosodic channels that are lost in writing; or strategies that focus on content and make maximal use of lexicalization, as these serve their context-bound needs and as these have been conventionalized in their speech habits" (pp. 13-14).

Tannen's notion of strategy choice being contextually-driven typifies one of the two ways of characterizing oral versus written language discussed in Wold (1992). For Wold, oral and written language can be examined in terms of their modality, which typically leads to an examination of the differences between them, or in terms of their communicative situation, which permits an understanding of the "broad variation with respect to communication and performances that exists for both oral and written language" (pp. 175-76).

Wold summarizes several ways that the modality of writing and speech leads logically to differences between the types of language associated with each. Where oral language is auditory, temporally structured, evanescent, and modulated by prosody, written language is visual, spatially organized, enduring, and modified by graphic characteristics (e.g., punctuation). Oral language is produced under time constraints and requires simultaneous comprehension by the audience; written language has no time limits for either production or comprehension. In oral language production, the speaker chooses the sequence of information intake; in written language, the ability to jump to the end of a section means that it is the reader who has control over information intake. A speaker's audience includes all those who are within earshot of the speaker; the writer's audience is limited only to those who actively focus on the paper the text is written on. The writer's message can be obscured through insufficient light, the speaker's through excessive background noise.

Wold also presents several differences between oral and written language that have been claimed to arise from the different communicative situations in which each is used. Thus, oral language is said to be typical of informal dialogues where a combination of a shared situational context, shared knowledge, and opportunities for feedback from the listener means that the speaker can be less explicit. In contrast, writing is typically seen as formal, monologic, and involving an unknown or invisible reader who shares no common here-and-now with the author, who, thus, must focus on making meaning as linguistically explicit as possible. However, Wold critiques the fundamental approach which allows such clear cut distinctions between oral and written language. She argues that these characterizations depend on prototypical communicative situations which ignore theoretically important variations between situations which may cause a great variety in language products: "In some of these situations oral language becomes explicit, independent of context, and monologic; while written language may become context dependent and embedded in ongoing dialogues" (p. 186).

Given the importance of communicative situation ascribed by Wold, how can we classify such situations in a way that will allow us to systematically examine the variety of their language products? Several approaches have been suggested, two of which will be discussed here: Jakobsen's speech event model (Jakobsen, 1960); and Hymes ethnography of communication (as described in Duranti, 1997).

Duranti (1997) identifies Jakobsen's (1960) speech event model as one of the pioneering models which embedded speech in social units. Jakobsen's model uses the speech event as the unit of analysis and identifies six constitutive factors of a speech event: an *addresser* who sends a *message* to his/her audience, the *addressee*. The message operates in a *context*, through a *code* shared by the addresser and the addressee who are connected either physically or psychologically by a *contact* (p. 73). Each of these factors has an associated function of language. The *emotive* function—that is, conveyance of information about the speaker's attitude towards the topic on which s/he is conversing—is associated with the *addresser* and can be expressed at the phonemic, grammatical, and lexical level. A focus on the *connative* function reflects an orientation towards the *addressee* since it is an "act of communication which transforms or attempts to transform reality or people, which aims to affect the course of events or the behaviors of individuals" (Yaguello, 1998, p. 12). References to the state of the world show the *referential* function which is associated with the *context* of a situation. A focus on the *message* for its own sake represents the poetic function. An orientation towards establishing, prolonging or discontinuing communication represents the *phatic* function with its emphasis on *contact* between the participants. Finally, speakers may attempt to focus on the *code*, that is to use language in its *metalingual* function where language is both the topic and the medium of communication.

Duranti (1997) notes that the central focus of Jakobsen's model is essentially the linguistic code and how different forms of participation could be linked to grammatical features of the language; however, in Dell Hymes' (1972b) ethnography of communication, the central unit of analysis was the *communicative event*, "a social unit which includes or is based on speech" (p. 289). The key components in Hymes' model are grouped under the acronym SPEAKING: Situation (i.e., the setting and scene), Participants (speaker/addressor, hearer/addressee), Ends (the purposes of the communicative event in terms of goals or outcomes), Act sequences (the form and content of the message), Key (i.e., the tone in which talking is done), Instrumentalities (i.e., channels and forms of speech), Norms (norms of interaction and interpretation), and Genre.

Hymes inclusion of *genre* in his model is interesting because genre has frequently been employed as a means of categorizing spoken and written language. Swales (1990) suggests that different disciplines have applied the concept of genre—defined broadly as "a distinctive category of discourse of any type, spoken or written, with or without literary aspirations" (p. 33)—in different ways: In folklore studies, genre is seen as a method of classification, as a permanent form, or as an indicator of sociocultural value within a community; the focus in literary studies is on genre analysis as a "clarificatory" tool which allows exploration of how writers break conventions; and linguistics employs an ethnographic approach which combines empirical observation with elicitation of the community's category-labels. The implication that genre is defined by the members of a particular community is stated in a number of other definitions (Carter and McCarthy

1997; Georgakopoulou and Goutsos 1997; Wales 2001) and forms part of Swales' (1990) own definition:

> A genre comprises a class of communicative events, the members of which share some set of communicative purposes. These purposes are recognized by expert members of the parent discourse community, and thereby constitute the rationale for the genre. This rationale shapes the schematic structure of the discourse and influences and constrains choice of content and style . . . . [E]xemplars of a genre exhibit various patterns of similarity in terms of structure, style, content, and intended audience. If all high probability expectations are realized, the exemplar will be viewed as prototypical by the parent discourse community. (Swales 1990, p. 58)

Several aspects of this definition have been echoed in later definitions. Thus, Georgakopoulou and Goutsos (1997) view genre as a concept which overarches both style and register and which allows communicative events to be classified by participants in terms of a shared set of "formal, functional, and contextual properties" (p. 33). Similarly, Wales (2001) sees genre as categories which members of the discourse community view as sharing communicative purpose and audience; he also distinguishes between literary genres, which share a set of stylistic and structural properties, and speech genres, which are a kind of contextually-driven social practice. Finally, Bauman (2001) takes a more linguistic approach to the notion of genre, seeing it as "a constellation of systematically related, co-occurrent formal features and structures that serves as a conventionalized, orienting framework for the production and reception of discourse" (p. 79) that occurs in recurrent communicative contexts.

The co-occurrence of linguistic features mentioned by Bauman has been the focus of much research by Biber (Biber, 1988; Biber, 1996; Biber, 1999; Biber & Conrad, 2001). Biber's work represents a breakthrough in the study of written versus oral language because of his use of a methodology which combines investigation of the co-occurrence of features, or "association patterns" (Biber, 1996, p, 173), in large-scale corpora of naturally-occurring language with the advanced statistical procedure of factor analysis. The results are analyses which look at variation in terms of multiple parameters and which support a functional analysis with empirical evidence rather than a set of a priori relationships. For example, in *Variation Across Speech and Writing* (1988), Biber presents the analysis of 481 texts which contain 960,000 words and represent 17 written and 6 spoken genres. The factors identified by his analysis lead him to propose six dimensions along which the genres vary: (1) 'informational versus involved production'—i.e., "high informational density and exact informational content versus affective, interactional, and generalized content" (p. 107)—is the most important dimension, in terms of the number of features which load onto it; (2) 'narrative versus non-narrative concerns', that is, "active, event-oriented discourse and static, descriptive or expository types of discourse" (p. 109); (3) 'explicit versus situation-dependent reference'; (4) 'overt expression of persuasion'; (5) 'abstract versus non-abstract information'; and (6) 'on-line informational elaboration', a dimension which "distinguishes discourse that is informational but produced under real-time conditions" (p. 117).

Having identified the dimensions along which the texts vary, Biber undertakes a macro-analysis of the genres relative to one another by computing factor scores for each genre. That is, for each text, Biber sums the frequency of each feature on a factor. Examining the results for each genre, Biber notes that because speakers and writers sometimes ignore the situational forces of each mode to produce discourse that is atypical of the mode they are using, there is no clear cut distinction between written and spoken language: "in the present study, no absolute difference is observed; with respect to each dimension, written and spoken texts overlap" (p. 160). For example, on the first dimension, 'informational versus involved production,' personal letters were grouped with interviews and spontaneous speeches while broadcasts were grouped with professional letters and general fiction. Furthermore, Biber's results do not support a unidimensional view of written and spoken language. In other words, while it is possible to define oral/written discourse in terms of situational characteristics that can each be defined along a single dimension, no single dimension adequately represents the features of language produced, leading Biber to argue that "consideration of all six dimensions is required for an adequate description of the relations among spoken and written texts" (p. 169). In fact, he argues that rather than representing differences between oral and written texts, the dimensions he identifies represent the "fundamental parameters of linguistic variation among English texts" (p. 200). In later research, Biber and his colleagues have applied this methodology to explore linguistic variation both within a single language—such as Biber's exploration of the variability across registers in the form and use of English complement clauses (Biber 1999)—and cross-linguistically through analysis of register variation in English, Somali, and Korean (Biber & Conrad, 2001).

This discussion of written and oral language needs to briefly address one final issue: Which of the two modes is primary? Halliday (1989) comments that writing is usually the more highly valued mode in literate cultures, a view which originated in Ancient Greece. However, many 20th century linguists with anthropological backgrounds who work with non-literate cultures have seen spoken language as the primary form. This is exemplified by Chafe (1994) for whom conversation is the basic use of language from which all other uses are derived. Halliday, however, sees this distinction as artificial: "The two [writing and speaking] are both language; and language is much more important than either. It is a mistake to become too much obsessed with the medium" (p. 92). Indeed, Halliday argues that rather than writing being seen as more complex, each mode is complex in its own way: Where written language is lexically dense—containing a higher proportion of content than function words—spoken language is grammatically intricate. Such a stance is congruent with Biber's (1988) claim that writing and speaking are different systems, each of which are worthy of analysis.

To summarize, views of written and spoken language as opposite poles of a dichotomy were called into question by research in the early 1980s which posited an oral/written continuum. Written and oral language could be defined not only in terms of their modality, which tended to emphasize differences, but also in terms of communicative situation. Jakobsen and Hymes provided two models for classifying a communicative situation, the former concerned more with functional/linguistic aspects of the situation, the latter with social aspects. A key concept in the discussion of spoken and written language has been genre, which Biber has investigated using language corpora. His study suggests that rather than a single written-oral language continuum, there are multiple dimensions which need to be considered when comparing oral and written discourse.

With regards to this study, the literature on oral versus written language suggests that although computer-mediated communication utilizes a different mode from oral interaction, that does not, a priori, imply that the discourse of CMC differs from that of face-to-face discussion. Studies have shown that written and spoken genres exist which are closer to each other than they are to other genres with whom they share a modality. The extent to which this is the case for computer-mediated communication vis-à-vis face-to-face communication is an empirical matter. The next section examines data from first language studies which shed light on this issue.

2.5 WHAT IS CMC DISCOURSE?

What this discussion has ignored so far is the question of where CMC discourse fits. There are several reasons for believing that the language produced in computer-mediated communication may pose a challenge to current categorizations of written and oral language. Discussing other forms of technology such as tape recorders, Wold (1992) noted that the use of technology blurred many of the logical distinctions between written and spoken language that arose as a result of the mode of production. Similarly, in their discussion of genre, Georgakopoulou and Goutsos (1997) note that the use of CMC "allows for texts that do not fall neatly into any particular category" (p. 37). The cause of these problems lies partly in the interactivity inherent in much CMC discourse. This interactivity goes beyond the rejection of writing as a monologic activity by scholars whose adherence to theories of social constructivism allowed them to claim that writing was a dialogic act (see, inter alia, Clark, 1990; Wold, 1992). Instead, synchronous, and to some extent asynchronous, computer-mediated communication involves a real audience with whom one interacts, which explains why CMC discourse has come to be called "interactive written discourse" (Holec, 1985; Ferrara, Brunner et al., 1991), "talking in writing" (Spitzer, 1986), and "written speech" (Maynor, 1994). The following paragraphs review several studies which have examined the linguistic features of CMC discourse.

In an early study of synchronous CMC, Ferrara, Brunner and Whittemore (1991) examine interactive written discourse, the language produced during "simultaneous terminal-to-terminal typed dialogues" (p. 9) in which interchanges scroll across the screen in real time and are not available for later access. The subjects for their study were 23 computer professionals who were asked to solicit information from an individual playing the role of travel advisor. Each subject interacted with the travel advisor for approximately one hour and forty-five minutes, during which time they were videotaped and encouraged to think aloud. The authors' analysis of the messages showed several interesting features. First, they found many omissions of subject pronouns, articles, and finite forms of the copula. Together with the shortening of words through abbreviations, symbols, or informal spellings, these omissions lead the authors to suggest that the e-messages may represent a reduced register. They also found evidence of a mixture of features from spoken and written language. For example, the use of yes/no questions and first/second person pronouns is typical of the interactiveness found in spoken language, while the presence of adverbs of time, direct questions, and general emphatics (e.g., *just* and *real*) represent features of involvement that are found in face-to-face conversation, as do the use of informal discourse particles such as *okay*, *sure*, and *sorry*. Features of writing include the use of more formal language that is elaborated and expanded and includes relative clauses, adverbial clauses and subordination. Ferrara, Brunner and Whittemore conclude that interactive written discourse represents a "hybrid register that resembles both speech and writing, yet is neither" (p. 10), but they caution that the high degree of variability between subjects may mean that it is still an emergent register.

In another study of synchronous CMC, Wilkins (1991) examines the messages produced in a chat room over a three-month period by 33 'novice' computer users who had never previously participated in computer conferencing. She found that since turn could be maintained by the writer of the message, there was no need to negotiate turns. While new topics were not directed to individuals, responses to previously posted messages were indicated by addressing the person who posted the original message by name. This reference to names did not, in itself, serve to maintain topic; instead, topic maintenance was achieved through lexical repetition, synonyms, and shared cultural knowledge. Wilkins claims that the use of names and lexical repetition were also indicative of interactive language use in that they validated a previous speaker's contribution to the conversation. Other features of the messages which demonstrated interactive language use are those which showed participants' high level of involvement:

references to a speaker's mental process (ego involvement); the use of names and second person pronouns; and devices which show the speaker's interest in the subject (e.g., exaggeration, exclamation, and expressive vocabulary). Finally, the messages showed evidence of the type of disfluencies found in face-to-face conversation: hesitancies, false starts, and statements of afterthoughts. Thus, Wilkins concludes that "in computer conversations, which provide a means for a traditionally oral activity to take place in the written form, we may observe a shift in the boundaries between spoken and written discourse" (p. 75).

Collot and Belmore (1996) use Biber's (1988) multi-dimensional multi-feature framework to examine a 200,000 word corpus of electronic language collected from a bulletin board. The absolute frequencies for Biber's 59 linguistic features were determined and used to calculate factor dimension scores for each of the six dimensions in Biber's framework. By plotting these dimension scores onto Biber's graphic representation of the dimensions for his corpus, the authors show that the electronic language corpus displays features commonly found in some forms of written language and other features commonly found in some forms of spoken language. Their conclusion about electronic language is that the "genres which it most closely resembles are public interviews and letters, personal as well as professional" (p. 21).

Yates (1996) compares three corpora representing written (Lancaster-Oslo/Bergen corpus), spoken (London-Lund corpus), and computer-mediated language. The computer-mediated corpora were obtained from a database of computer conferences and included over 2,000,000 words produced during asynchronous CMC. Yates analyzed all three corpora for three lexical measures (type/token ratio, unweighted lexical density, and weighted lexical density), use of personal pronouns, and use of modals. He found that CMC discourse bore striking similarities to written discourse on all of the lexical measures but differed significantly in terms of pronoun use (which was closer to spoken discourse) and modal use (which differed from both writing and speaking).

While the above studies have focused on synchronous CMC, the discourse of asynchronous CMC has also been the subject of investigation. Using anecdotal evidence, Maynor (1994) suggests that e-mail has its own style characterized by the omission of subject pronouns, modals, auxiliaries, and copulas, the use of informal words such as *yep* and *nope*, a lack of capitalization, and the use of simplified spellings, abbreviations, and icons. Her findings are supported to some extent by Gaines (1999), who uses a variety of tools to examine the textual features of a small corpus of 136 e-mail messages collected from two settings in the United Kingdom, including 62 e-mails from an insurance company and 54 e-mails from an academic setting. His results show a marked difference between the two settings. The commercial e-mails, with the exception of surface adaptations to forms of greeting and closing, resemble the language of formal business settings with the fully-formed, correctly punctuated, grammatical sentences of paper-based documents such as business letters or faxes. Gaines attributes these qualities to the legal status afforded to such messages by financial services legislation in the United Kingdom. In contrast, the academic e-mails display a much wider range of registers, including many features associated with conversation such as those which Maynor described. For example, the academic messages contain such interactive features as rhetorical questions and responses to imagined echo questions, and they show selection of lexical items with oral qualities (e.g., the use of phrases like *no sweat man* and of *just* as a modifier). Messages also included features which echoed other forms of oral language such as answer machine talk and the opening sequences of telephone talk. Thus, Gaines cautiously concludes that in spite of the small sample from which his analysis is drawn, the creativity and conversational tone seen in some messages may be evidence of a genre that differs from previously identified genres.

Gaines' study is interesting because its finding of distinct differences between e-mails produced in a business setting versus messages produced in academia suggests that our discussions of CMC should refer to *discourses* in the plural, with the nature of each discourse being a function of the context in which it is constructed. Thus, in a business setting where e-mails may have the status of legal documents, CMC discourse is more likely to incorporate features found in other legally important documents such as letters or faxes. In a similar way, what the studies described here show is that when CMC is perceived by its users as being a 'conversation by other means', its discourse incorporates many features associated with spoken language to produce a discourse that is hybrid in nature.

2.6 COMPUTER-MEDIATED COMMUNICATION IN THE L2 CLASSROOM

CMC has been implemented in language classrooms in many different ways, but many of the theoretical arguments for its use can be identified with two traditions which are not necessarily mutually exclusive. One tradition emphasizes CMC's compatibility with interactionist models while the second tradition places it within social theories of learning. A third, separate set of arguments are more practical in nature, dealing with CMC and affective concerns. The following sections will discuss all three perspectives on using CMC and will review relevant research on the use of CMC in the classroom.

2.6.1 INTERACTION AND CMC

One of the most often cited rationales for CMC is its ability to provide communicative interaction since newsgroups and listservs offer students more people to communicate with and communicative opportunities for collaborative learning (Lee, 1997). These rationales draw implicitly or explicitly on interactionist theories. Hatch (1978a, 1978b) was among the first to suggest a link between interaction and language acquisition. Drawing on data from first and second language learners, both child and adult, she argued that language learning—i.e., the development of syntactic structures—evolved from learning how to communicate through interaction.

The importance of interaction was further established in a series of pioneering studies by Long (1981, 1983, 1985) in which he examined the modification of input by native speakers in their interactions with non-native speakers. In an early study (Long, 1981), he showed that interactional features were modified to meet the communicative needs of the conversation and hypothesized that modified interaction might be a necessary and sufficient condition for second language acquisition. In Long (1983), he expanded on this notion, suggesting an indirect causal chain which would explain how such acquisition might occur: "If it could be shown that the linguistic/conversational adjustments promote comprehension of input, and also that comprehension of input promotes acquisition, then it could safely be deduced that the adjustments promote acquisition" (p. 189). Subsequent research (Long, 1985) established the first part of this causal chain, that modified input did lead to greater comprehension.

One of the more recent interactionist models is found in Gass (1997) who describes a model of second language acquisition that is clearly influenced by the ideas of Long. In Gass' model, language acquisition occurs as a dynamic system in which input becomes output via a 5-stage process involving apperceived input, comprehended input, intake, integration, and output. In the first stage (*apperceived input*), learners recognize a gap in their linguistic system that needs to be filled. Gass suggests several factors which may influence learners' apperception of input—time pressure, frequency, affect, prior knowledge, saliency of form, and attention— and which may operate in interaction. Gass characterizes apperception as a priming device which allows the learner to notice language for later analysis and which is followed by comprehension (i.e., *comprehended input*). *Comprehended input* is different from comprehensible input in that the focus is on the hearer rather than the speaker and the extent to which the former understands the language to which he or she is exposed. Gass suggests an important role for (a) negotiated interaction of form or meaning and (b) linguistic modification in making input comprehensible to learners. Negotiation occurs when there is a perceived "asymmetry between the message transmission and reception and when both participants are willing to attempt a resolution of the difficulty" (p. 108). The resulting clarification and elaboration result in more input whose interlanguage features have already been marked as deserving greater attention from learners.

In the next stage in the model, the comprehended input becomes *intake*, which Gass describes as the psycholinguistic process by which linguistic material is compared to prior knowledge. However, she cautions that not all comprehended input becomes intake because it may simply serve the immediate conversational needs without being subject to the sorts of analysis implied by intake. Where learners do analyze the input as intake, the result is the fourth stage in the model, *integration*, in which the learner either develops their L2 grammar by using the new knowledge to reject or confirm hypotheses about the target language, stores the new linguistic material for later use because the information from the intake is insufficient to confirm or reject a hypothesis, or does not use the new knowledge at all.

In the fifth stage, learners manifest the result of the process of acquisition through *output*. This should not be interpreted as suggesting that Gass sees output in her model as the culmination of language acquisition; in fact she argues the opposite, that output is an important part of the learning process. In doing so, Gass acknowledges Swain's (1985) notion of "comprehensible output" which was developed to explain the limited L2 development of students in a French immersion program. Swain claims that the language development of these students, who had over seven years of input to the target language, is evidence against the theory that input alone will facilitate language acquisition. Instead, she argues that these students were hindered by a lack of opportunities to use language productively, or as she explains it, to be "pushed towards the delivery of a message that is not only conveyed, but . . . is conveyed precisely, coherently, and appropriately" (pp. 248-49). Gass suggests that output contributes to language acquisition by providing learners with opportunities to test their hypotheses and receive feedback which may help them to notice mismatches in their interlanguage or deficiencies in their output.

Given the influence of interactionist approaches on classroom practice, it is unsurprising that several scholars have used an interactionist perspective to comment on the potential benefits of interaction in a CMC environment. Turbee (1999) notes that MOOs "[bring] teachers, learners, and native speakers together in intense interaction" (p. 361). Kelm (1996) suggests that the use of peer-to-peer communication promotes communicative competence, as does Pellettieri (2000), for whom interaction in chat rooms is similar enough to face-to-face interaction that it can enhance language development. For Oliva and Pollastrini (1995), CMC promotes acquisition by "providing the circumstances necessary for a high degree of communicative urgency" (p. 552). For Johnston (1999), it is "the availability of an authentic audience [that] affects the rate and extent of language learning" (p. 57). While noting the complexity of the notion of audience, he defines an authentic audience as "an audience that is concerned exclusively with the meaning of the speaker's message" (p. 60). Moreover, since the messages are usually on topics that interest the students, and which may have been selected by them (Kelm 1996), computer-mediated exchanges are likely to allow an authentic audience in Johnston's terms since interlocutors are focusing on meaning over form. In fact, Johnson argues that CMC allows learners to exchange their traditional role of 'eavesdropper' in language classes—where they are merely listening in on others' exchanges—for that of 'authentic audience', fully involved in the interaction.

Other arguments focus on the authenticity of the language to which learners are exposed. Drawing on Krashen and Terrel's (1983) distinction between language acquisition and language learning, Oliva and Pollastrini (1995) argue that "interaction with native speakers . . . exposed students to natural models of language usage" (p. 552), an argument also found in Kelm (1996). For Brammerts (1996), native speakers can provide learners with a model of language usage and can assist them with expression of ideas and corrective feedback. Kelm and Brammerts both discuss the possibility of separating discussion of grammar from the act of communication. Since interaction through CMC is written, teachers have access to a record of all interactions, allowing metalinguistic analysis. For Kelm, these written records are beneficial in several ways: they provide instructors with multiple examples of L1 transfer which can serve as 'teaching moments'; they allow individualized grammar instruction; and they permit students to peer-edit each others' work.

Proponents of CMC have also highlighted its potential for increasing cultural understanding. The rich source of cultural information available on the internet can promote "the method of learning experientially rather than through the memorization of facts" (Lee, 1997, p. 411). Students can participate in society and culture newsgroups which "enhance the resources normally available to foreign language teachers because they provide a forum in which students can 'converse' with native speakers about current issues pertaining to the culture" (Cononelos & Oliva, 1993, p. 529). Thus, students may come to "glimpse other ways of seeing the world" (Kern, 1996, p. 106). Oliva and Pollastrini (1995) take this idea further, equating the use of networked resources to a "virtual immersion in Italian language and culture . . . [which] helps students to improve their language skills in a manner similar to full immersion or study abroad" (p. 551). Not all researchers have seen use of the target language as being necessary to promote cultural understanding. Bernhardt and Kamil (1998) have argued that opportunities for the successful integration of culture may not be realized since "the culture form can only be targeted at a level that can be explicated within the limited set of linguistic structures available" (p. 40). In other words, lack of linguistic skills by students at lower proficiency levels can inhibit discussion of culturally important issues. Thus, Bernhardt and Kamil argue for the use of L1 'knowledge sources' supplemented with on-line discussion in learners' L1.

2.6.2 CMC AND SOCIAL CONSTRUCTIVISM

A second group of arguments for CMC in language learning draws on social constructivist models of learning (e.g., Kern & Warschauer, 2000), which are based heavily on the ideas of Vygotsky. According to Bonk and Cunningham (1998), "a primary tenet of Vygotskian psychology is that individual mental functioning is inherently situated in social interaction, cultural, institutional, and historical contexts" (p. 35). Synthesizing the work of several educational theorists, Bonk and Cunningham provide a clear explanation of how key sociocultural principles are related to computer-supported collaborated learning. Learning is assumed to be influenced, or mediated, by the tools and institutional settings learners are exposed to. The use of technology changes both the tools and the institutional setting, and thus may have profound effects on how learning occurs within the zone of marginal proximal development (ZPD), which is the distance between what a learner can achieve independently and what they can achieve under the guidance of a teacher or in collaboration with more able peers. In order to promote learning (i.e., the movement to higher levels of ability), the teacher or the more capable peers provide assistance, or scaffolding, that helps less able students achieve a task that they would not have been able to solve independently. Learners' development as a result of this scaffolding occurs twice: once in the successful social completion of a task with others, and a second time when the skills/ abilities become internalized and the learners are able to complete the task independently of others. Social interaction can thus be seen as a sort of cognitive apprenticeship in which traditional roles of all participants, including teachers, change. Rather than directing learning, teachers assist in the learning process by providing "rich interactive conversations" about learning both with and among students. Such conversations help to develop intersubjectivity, that is "a temporary shared collective understanding or common framework among learning participants" (p. 41) which helps make scaffolding more effective. The implication of this theory is that any study of learning has to link the individual to his or her social setting. Learning cannot be examined outside of the context in which takes place.

In recent years, second language acquisition scholars have linked social theories to the use of CMC in language classrooms (Barson & Debski, 1996; Kern, 1996). For Barson and Debski (1996), the introduction of technology to language learning represents "a partnership" which is "redefining language learning and altering the fundamental notion about how best to create suitable environments for language acquisition in academic settings" (p. 51). They describe the computer as a "facilitator" which allows students to use language constructively and creatively. In the same volume, Kern (1996) contrasts the traditional role of computers as a consultative tool to their new roles in CMC as a medium, an "additional channel through which to communicate" (p. 108). He draws on Vygotsky and Bakhtin to reconceptualize teachers' roles in CMC as integral participants who scaffold student learning with their own knowledge and experience.

Beauvois (1997) also discusses the use of CMC for scaffolding but does so in terms of its potential for creating a linguistic community, "an interactive situation that . . . occasions the creating of much needed social structures that are so crucial to language learners as they progress along the continuum of interlanguage" (p. 166). According to Beauvois, CMC engenders this feeling of community in two ways: students get to know each others' names (which, she argues, rarely happens in most college level language classes); and having all comments and thoughts freely available on the screen creates a transparency of thinking. Thus, the discussion of ideas is continuously collaborative with multiple authors who build upon each other (Peyton, 1999).

One example of the creation of new communities comes from advocates of using MOOs in classrooms. Sanchez (1996) defines MOOs as "text-based virtual reality systems" which are user extensible yet permanent enough to allow users to develop "a feeling of actual existence (telepresence) in cyberspace" since users not only interact with each other but also with the environment (p. 149). Kern (1998) describes interactions between MOO users as developing in a multilinear and associative fashion where the structure of participation is determined collaboratively by users. He suggests that the continual construction and re-construction of identities that can occur in MOO environments “will give voice to learners’ other language selves, creating a dialogue that may lead to greater self-understanding and perhaps self-transformation” (p. 81), especially if learners are encouraged to reflect on their written interactions. Kern believes these interactions are “governed by a different set of conventions and constraints” (p. 59), as do others such as Schetzer and Warschauer (2000), who highlight the need for students to develop ‘electronic literacy’. Since CMC has particular stylistic and sociolinguistic features and involves new ways of interacting and collaborating, students should be taught those speech acts and conversational strategies necessary for them to join and to interact in on-line discourse communities (Shetzer & Warschauer, 2000). Thus, the incorporation of CMC into language classrooms has social as well as pedagogical utility (Kern & Warschauer, 2000).

Finally, Warschauer, Turbee, and Roberts (1996) present a convincing argument that CMC has the potential to facilitate student empowerment. Their argument presents a fresh perspective on several of the claims that have already been discussed in this chapter. Thus, they point out that the opportunity to control discourse and take the initiative in discussions with their peers provides students with greater autonomy. Additional benefits claimed by the authors are that more egalitarian participation of shy or minority students is seen as a democratizing trend, students benefit from the development of “skills of inquiry, interpretation and application” (p. 8) that arise when writing is seen as an aid to thinking, and CMC allows collaborative development of ideas. Warschauer, Turbee, and Roberts conclude that if students receive adequate computer training and are clear about the teacher's expectations for their roles, this combination of autonomy, equality, and learner skills can be empowering.

2.6.3 CMC AND AFFECTIVE CONCERNS

The final group of arguments in favor of the use of CMC in language classrooms deals with affective concerns. Several scholars make claims about CMC's ability to respond to students' affective needs, but there is less agreement on this topic. Sanchez (1996) cites Pantelidis' (1995) list of reasons to use text-based virtual reality systems such as MOOs. Several of his reasons address the affective domain: providing a "social atmosphere"; allowing passive students to become active; allowing experimentation with different personalities; and allowing role playing. To this list, Sanchez adds that reticent students may lurk (i.e., watch a conversation for some time before joining in), teachers may use 'whispering' to correct errors without embarrassing students, and students may interact through their 'telepresence' (a character that they define through a written description) rather than their real self.

Many scholars would agree with Lee (1997) that CMC allows lessons to be self-paced and learner-centered. Kelm (1996) emphasizes that slower students may work at their own pace and thus may be less intimidated and more likely to participate. For Kern (1998), one benefit of using synchronous conferencing is that students can "voice their thoughts at will without interrupting other participants' thoughts or expression" (p. 59). A corollary of this is that students do not need to worry about other people interrupting them, which may lead to greater participation. However, Colomb and Simutis' (1996) discussion of synchronous CMC in L1 writing classes offers an interesting counterpoint to the ideas of Kelm and Kern. Colomb and Simutis note that sending a message in a networked classroom is not the same as 'gaining the floor'. Although quiet students may contribute their messages to a discussion, they are not full participants in the interaction unless their message draws a response which influences the direction of the discussion. In addition, contributions during synchronous discussions need to be timely or they will be ignored. Pinto (1996) acknowledges a related problem for L2 learners. While arguing that the novelty and anonymity of MOOs may increase motivation, he also notes that students who are unfamiliar with MOO environments may find them disorienting and may be confused by the speed with which messages flash up onto the screen. The argument that students can work at their own pace may, therefore, be more relevant to asynchronous than to synchronous CMC.

2.7 RESEARCH ON CMC IN L2 CLASSROOMS

Much of the early research on CMC in language classrooms was anecdotal or descriptive, relying on the use of interviews, surveys, and teachers' impressions. For example, the students in Cononelos and Oliva's (1993) 400-level Italian class reported increased confidence in using Italian after participating in cultural newsgroups. The majority of the students (41 of 71) in Oliva and Pollastrini's (1995) study of the use of cultural newsgroups in advanced level Italian classes reported improved writing skills with a significant number (24 of 71) also reporting improvements in reading skills. However, the students were concerned about the inconvenience and time commitment involved in accessing computer labs outside of class. Van Handle and Korl (1998) also reported this problem for their e-mail exchange between intermediate level German classes at two U.S. universities; more positively, they claim, based on anecdotal evidence, that CMC produced a higher level of participation, a higher quality of discussion, and a focus on communication over accuracy. Johnson's (1996) description of a keypal arrangement between two low-level Spanish classes in U.S. high schools leads him to claim that communication was facilitated by similarities in cultural background and proficiency level.

A number of researchers have conducted more rigorous studies, which can be grouped into five broad areas: CMC's role in building cultural understanding; an examination of how individual differences influence performance in a computer-mediated environment; the potential effect of CMC interaction on overall proficiency; sociocultural studies of classroom-based CMC; and analyses of CMC discourse.

2.7.1 CMC AND CULTURAL UNDERSTANDING

Although the literature contains many studies of cultural acquisition which report positive results from the use of CMC, much of that research has been anecdotal. For example, Tella's (1992) ethnographic study of the introduction of technology into secondary school foreign language curricula in Finland found that students who participated in e-mail exchanges with classes in the UK and the USA had increasing awareness of cultural differences. Cononelos and Oliva's study (1993) showed a similar finding for an Italian class in which students had to participate on newsgroups relating to social and cultural issues. Soh and Soon (1991) reported that the teenage Singaporean and French Canadian students in their study learned about their own as well as each others' cultures through on-line discussion of locally and internationally significant issues.

Students in Lee's (1997) study researched cultural topics of their own choice on the world wide web and discussed those topics with peers, their instructor, and native speakers using e-mail. In a follow-up survey, the students reported that the internet was a useful tool for developing cultural knowledge and that they had an improved attitude towards learning culture.

These studies have all focused on interactions between native speakers and learners at the intermediate level or above. Johnson (1996), however, reports on a beginning level, high school Spanish class in rural California which had to pair itself with a similar class in San Francisco because no classes in Spanish-speaking countries were available. Johnson concludes that this arrangement suited his students better than interacting with native speakers would have since the relative cultural similarity and the similarity of proficiency levels facilitated communication. He also believed that the difference in backgrounds—urban versus rural—meant that there were cultural differences between the two groups of students which were explored in the e-mail exchanges. This last point is interesting because it suggests, albeit in anecdotal form, that students may have been developing an awareness and tolerance of multiple viewpoints that would serve them well in learning about Hispanic cultures.

Like Johnson, Bernhardt and Kamil (1998) focus on the problems for beginning level college students of interacting with native speakers. Even though such students are able to understand and conceptualize complex notions of culture, their limited linguistic resources preclude extended discussion in the target language. As a result, Bernhardt and Kamil decided to supplement the regular L2 course materials with an English language cultural history text which students discussed outside of class on an English language newsgroup. Bernhardt and Kamil claimed that students showed significant engagement with the cultural text and that the language class became "a systematic intellectual endeavor".

With the exception of Tella's ethnographic study, all the above studies used surveys and questionnaires to collect data. While surveys may allow researchers to identify students' perceptions of culture learning, they offer little insight into either the processes by which learners arrive at cultural understanding or the nature of that understanding. One potential source of greater insight into these areas can come from qualitative analyses of the interaction as documented in the transcripts provided by many of the software programs used in CMC. Thus, in their analysis of the integration of technology into an EFL curriculum in Bulgaria, Meskill and Ranglova (2000) examined the discourse in e-mail exchanges between the Bulgarian students and graduate students in the U.S. Their qualitative analysis allowed them to conclude that these exchanges had brought in multiple perspectives on the short stories that the EFL students were asked to write responses to. Moreover, since the contributions of the US students were often tentative rather than absolute, Meskill and Ranglova claim that the e-mail exchange empowered the Bulgarian students since they were involved in the co-construction of meaning with the American students. Unfortunately, their article does not provide the sort of detailed analysis necessary to substantiate this claim. However, it does highlight the important potential of discourse analysis in showing learners' negotiation and construction of meaning.

In a study which combined quantitative and qualitative analysis, Meagher and Castanos (1996) describe research in which the attitudes towards American culture of Mexican high school students of English were measured by assigning a semantic differential score to adjectives that the students chose to describe Americans. Differences between pre- and post-test scores show that the 26 students in their study had developed a more negative attitude towards Americans as a result of their year-long e-mail exchanges with a class in San Diego. Meagher and Castanos' analysis of CMC transcripts also leads them to claim that students' discussion of the Rodney King affair had made them more critical of racist policies in the U.S. The authors attribute this negativity to the students being in a stage of culture shock; however, a more likely explanation is found in the authors' suggestion that students awareness of "both differences and similarities between the culture of L1 and L2 . . . . suggests a model of intercultural dialogue according to which simple, general attitudes towards the foreign culture are replaced by complex, diversified opinions about different aspects as knowledge of the culture increases" (p. 200). In other words, the students' changes in attitudes reflect the replacement of a simple, generally positive, view of American culture with a more complex view, in which they are able to see both good and bad aspects of American culture with greater clarity.

2.7.2 INDIVIDUAL TRAITS AND CMC

Only one study to date (Meunier, 1998) has examined the influence on individual traits on the use of CMC. Meunier's study investigated the extent to which individual traits such as personality, motivation, attitude, and gender influence their use of computer-mediated communication. To identify the motivation types of the 64 third-year

French and German students in her study, Meunier used a 43-question survey which combined Likert-scale responses with opportunities for students to comment on their choices. Students' personality types were determined with the Myers-Briggs Type Indicator personality test, which measures four personality traits: Introversion or Extroversion; Thinking or Feeling; Sensing or Intuition; Judgment or Perception.

The picture that emerges from Meunier's results is complex. 'Introverts' and 'Extroverts' were equally stimulated by the use of CMC, but 'Introverts' were more easily overwhelmed by the flow of messages. 'Intuitive' students were more at ease than 'Sensing' students; the latter tended to worry less about accuracy and to prefer the use of pseudonyms. More 'Thinkers' preferred CMC in comparison to 'Feelers', who missed the paralinguistic cues of face-to-face communication and were more sensitive to flaming. 'Judgers' liked the opportunity to work at their own pace; if they were also 'Extrovert' they were less likely to be overwhelmed by the flow the messages than if they were 'Introvert'. Compared to females, males were much more overwhelmed by the flow of messages and tended to write longer messages. Instructional context was found to play a role. Motivation tended to be low in those classes which used a lot of computer peripherals, or where the teacher was overly monitoring, lacked confidence in the technology, provided boring topics of conversation, used CMC infrequently, or failed to integrate CMC activities into the curriculum.

While this study is a valuable first step in examining the effect of individual factors on the use of CMC, it raises more questions that it answers. Meunier suggests that there may be an interaction between personality type and instructional style, but it is also likely that interactions exist between several of the variables that she investigates. All these interactions are worthy of further investigation.

2.7.3 CMC AND ORAL PROFICIENCY

One area which has received some attention is the role of computer-mediated communication in oral proficiency development. Beauvois (1997, 1998) used a quasi-experimental design with a control group to examine the effect of CMC interaction on speaking skills. She found that students who discussed class texts using CMC had statistically significant higher speaking scores than students who discussed the texts face-to-face. Unfortunately, these potentially powerful findings are undermined by measurement issues. Students' speaking scores were determined using four sub-scales, one of which was content or "accuracy of response" (Beauvois, 1998, p. 107). Since the test asked students about the texts they read, it seems reasonable to assume that 'content' refers to the accuracy of their understanding, in other words, comprehension. However, Beauvois (1997) also writes that transcripts of CMC discussions allows teachers to check for comprehension of texts in ways that are not possible during face-to-face discussion and to "comment on what is accurate or inaccurate in the students' perceptions" (p. 169). Thus, the higher speaking scores in the experimental group are not interpretable since they can be attributed either to increased comprehension developed through better teacher feedback or to transfer of skills from writing to speaking.

A more persuasive argument for CMC's role in developing oral proficiency is found in Payne and Whitney (2002). Payne and Whitney augment Levelt's model of language production with Working Memory theory to investigate whether CMC can aid in the development of oral proficiency and, if it does, whether this development may be related to the reduced demands on working memory that the slower pace and textual nature of CMC would seem to allow. Participants were drawn from third semester Spanish classes, all of whom received identical instructional content. While class activities for the two control groups were only conducted in face-to-face interactions, approximately half of the activities for the experimental group was conducted using synchronous CMC. Allowing for differences in pre-test scores, the authors found that the experimental groups' oral proficiency scores were higher at a statistically significant level than those of the control groups. However, they caution that the results should be interpreted as an argument for integrating CMC and oral interaction in the development of oral proficiency rather than relying solely on CMC.

The possibility that CMC may contribute to the development of not only students' writing skills but also their oral proficiency is one that is intriguing and needs to be verified in further studies. Even if CMC is found to play an important role in L2 oral development, a number of questions need to be answered concerning (a) the types of CMC interactions that are most likely to increase overall proficiency, (b) the most effective combination of face-to-face and electronic interaction for furthering language acquisition, and (c) the relative effectiveness of interaction with native-speakers as compared to interaction with fellow non-native speakers.

2.7.4 SOCIOCULTURAL PERSPECTIVES IN CMC RESEARCH

A number of studies have been heavily underpinned by sociocultural models. Kern (1996) describes an e-mail exchange in which a French class at Berkeley and Lycée students in France discussed a book of family histories written by the latter group. He cites evidence of sociolinguistic lessons in the context of real conversation (e.g., use of *tu* and *vous*), linguistic scaffolding (i.e., the use of forms not yet covered in class but present in the exchanges), and the presence of intertextual elements which aid comprehension and provide "a rich source of linguistic input" (p. 116). Warschauer's ethnographic study (2000) found that a combination of institutional factors and teacher beliefs influenced how technology was introduced into four classes. The success of activities using the technology depended on students' perceptions of the relevance and purpose of those activities.

Writing from the perspective of socio-cultural theory, Darhower (2002) suggests that the chat room discourse produced by two fourth-semester Spanish classes shows that students created a "dynamic learner-centered discourse community" (p. 273) in which they negotiated discussion topics of mutual interest (especially when the instructor was not present) and maintained social cohesiveness through a combination of greeting and leave-taking, humor, playful insults, experimentation with identities, role playing, and the strategic use of English to further the conversation. This, he claims allows students to both enjoy their learning experience and develop sociolinguistic competence.

Abrahms (2001) also examines the type of community that learners create in a CMC environment. She investigates students' participant roles during two different types of activities: synchronous CMC discussions and a more traditional written interactive task, the group journal. She found many roles in common to both writing environments; however, learner roles in the CMC environment showed more diversity with learners adopting several roles—attacker, challenger, supporter and joker—not found in the group journals. Abrahms suggests that these roles arose because of learners' perceptions that comments on the computer screen were relatively impermanent and that the immediacy of CMC interactions allowed immediate repair.

2.7.5 TEXTUAL AND INTERACTIONAL FEATURES OF CMC DISCOURSE

One of the first studies to look at patterns of interaction in CMC was Chun (1994). In this early descriptive study, Chun examines the use of synchronous CMC over two semesters of German classes. In the first semester, fourteen students used the Interchange program for five discussions lasting 15-20 minutes. In the second semester eight of the same students plus one new student had nine computer sessions lasting twenty to forty minutes. The students differed enormously both in the number of the turns they produced (individual's averages ranged from 2.8 to 17.8 turns per session) and in their styles of participation (single, grammatically simple sentences versus multi-sentential, grammatically complex entries). More important, however, was the high degree of interactive competence evidenced by the large number of entries showing students answering questions, asking each other questions, and making statements that expanded on an existing topic or started a new one. Moreover, Chun found that students were interacting more with each other than with the teacher.

Kern (1995) found a similar pattern for CMC interaction in a study which compared the features of whole class computer-mediated and face-to-face discussions in two second-semester French classes which met in a computer lab once every two weeks. During the computer session, the 40 students in the study used Interchange to discuss a topic for 45 minutes. In the following class, they discussed the same topic face-to-face. Kern analyzed data from a single CMC and face-to-face session during the 10th week of the semester and found differences in the language produced under the two conditions. When students used CMC, they produced more language, both absolutely and in terms of number of turns and number of sentences. Their CMC discourse also tended to be more complex in terms of morphosyntactic features and range of functions used. However, two caveats apply to Kern's findings (both of which he acknowledges): First, differences in students' language were not subjected to parametric statistical tests; and second, the research design was flawed since the oral discussion always followed the computer-mediated discussion. As Kern suggests, some students may have felt "talked out" after the Interchange session, which would account for their lower levels of language production in the face-to-face discussion.

Like Kern, Warschauer (1996) also compared students production in face-to-face and CMC discussion in a study that was methodologically more robust since it was counterbalanced for both topic of discussion and order of use of CMC. Although Warschauer was able to use parametric tests for his measures of language complexity—type/token ratio and a coordination index (the ratio of independent clauses to total clauses)—the majority of his analyses are, like those in Kern's study, descriptive rather than inferential. Warchauer found that, compared to the face-to-face discussion, the CMC discussion produced more equal levels of participation and more complex language. A qualitative analysis showed that the CMC discussion involved longer turns and more formal language.

Sullivan and Pratt (1996) examine whether the use of computers can influence writing apprehension, attitudes towards writing, or the writing growth of intermediate students in two ESL writing classes—one which met in a regular classroom, and one which met in a computer lab once or twice a week and conducted discussions electronically. Their results showed no difference between the two groups for any of these variables. However, a discourse analysis of the interactions replicated the findings of Warschauer (1996) and Kern (1995): The CMC group had much higher levels of student participation and lower teacher involvement than did the whole class discussions; peer group discussions for the CMC class involved fewer turns but were much more focused than in the regular class.

Gonzalez-Bueno (1998) reports on the voluntary use of e-mail as electronic dialogue journals by 50 first- and second-semester students of Spanish. The e-mail exchanges enhanced students' participation, in terms of the quantity and quality of their output, and allowed better management of time compared to paper-and-pencil dialogue journals. An additional benefit was the conversational nature of the language that students produced, as evidenced by the presence of discourse markers typically found in face-to-face interactions, the use of phatic questions to keep the conversation going, and the similarities of leave-taking formulas to those found in telephone conversations.

Blake (2000) investigated the negotiation of meaning of 50 students in two intermediate-level Spanish classes as they used synchronous CMC to complete three different task types: jigsaw activities, information gap activities, and decision-making tasks. He found that the students' negotiations followed the trigger-indicator-response-reaction pattern found in face-to-face interactions. The jigsaw tasks produced the majority of the negotiations; a chi-square comparison of jigsaw against all other tasks produced a significant result at an alpha level of .05. In addition, the majority of the negotiations between the students involved lexical confusions. Blake concludes that networked interactions produce the same benefits in terms of negotiations as face-to-face interactions.

Smith (2001) investigates the negotiation of meaning and communication strategy use that occurs when students encounter unknown lexical items during synchronous computer-mediated communication, and the effects on lexical acquisition of that interaction. Twenty-four ESL students, divided equally among two proficiency levels, completed two decision-making and two jigsaw-tasks, each of which contained eight unknown lexical items (as established on a pre-test). Acquisition of these items was tested on post-test administered immediately after the treatment and one week later. Smith found that learners used a wide range of communication strategies as they completed the tasks using CMC, that communication breakdowns led to negotiations of meaning that were similar to those found in face-to-face interactions, and that gain scores on both post-tests provide evidence of a link between negotiation of meaning and lexical acquisition. In a finding that contradicts Blake's (2000) results, Smith suggests that the interactions in his study suggest that the traditional model for negotiation of meaning be extended to include two extra phases which represent the need for explicit acknowledgement that understanding has occurred. Thus, learners may provide confirmation of their degree of understanding which elicits a brief reconfirmation from their interlocutors that the negotiation routine has been completed.

Pinto (1996) looked at patterns of interaction between 15 ESL students during four 90-minute sessions using a MOO. The majority of moves (70.6%) were initiating moves with relatively few continuing moves. Pinto concluded that conversations were not very fluent and lacked the give and take of regular face-to-face interaction, but he failed to consider that the large group size may have caused difficulty in following threads, especially for students lacking experience in this environment. In another study (Pellettieri, 2000), patterns of interaction between intermediate-level Spanish students in a chat room were found to be similar to those that occur during oral conversation in language classrooms: Negotiation of meaning occurred between students; this negotiation made the language input more comprehensible; and when combined with corrective feedback, it caused learners to attend to form and to modify output. Another study (Fernandez-Garcia & Martinez-Arbelaiz, 2002) of negotiation of meaning by third-year students of Spanish during two chat sessions showed that breakdown of meaning was mostly associated with lexical items, was usually indicated by an explicit statement of misunderstanding, and was resolved in most cases through supplying the L1 equivalent. The authors suggest that the written medium may limit the types of indicators which students can use. Finally, Davis and Thiede (2000) examined the participation of three recently arrived non-native speakers during on-line discussions in a graduate level course. The authors suggest that NNS modification of writing style to match their interlocutor's status showed 'cultural adjustment'.

2.8 CONCLUSION

The review of literature in section 2.7 shows that research on the use of CMC in language classrooms is, like the use of CMC itself, a dynamic area that is still in its infancy. There remains a clear need for further research in many areas. With regard to language pedagogy, much of the research to date has been too anecdotal and described in terms too general to guide teachers. There has been little systematic exploration of how CMC might be effectively used or the contextual factors which might contribute to its success. Research is needed to identify what types of CMC tasks and interactions facilitate language acquisition, with which students, and under what circumstances. Moreover, many of the claims about the success of CMC remain inadequately demonstrated. Among the many issues requiring further investigation are the extent to which CMC encourages oral proficiency, whether it aids in developing reading and writing skills, and whether its potential for allowing a post-communication focus on form actually facilitates language acquisition. Although studies have shown a role for CMC in promoting cultural understanding, there has been little systematic examination of the nature of that understanding or how it emerges.

With regard to the present study, the research offers few guidelines about the types of inferences we can make about oral performance on the basis of computer-mediated communication. The literature suggests that negotiation of meaning takes place in both environments, but the nature of that negotiation in CMC environments is unclear. Some studies find similar patterns of negotiation to those found in face-to-face conversations while others find slightly different patterns. The comparisons of face-to-face and computer-mediated interactions in Kern (1995) and Warschauer (1996) might lead us to surmise that on the CMC test, students may participate more equally and may produce more language in longer turns and involving greater complexity. Unfortunately, Kern's study, which looks at the greater number of morphosyntactic and functional variables, lacks the parametric tests that would allow us to generalize to other contexts, while Warschauer's study, which does conduct parametric tests on variables of linguistic complexity, examines very few variables.

An additional muddying of the waters is provided by the question of whether we can generalize from studies of classroom interactions to interactions which occur in assessment contexts. Skehan's model of language performance (described in section 2.2 above) includes *ability for use* as a means by which learners adjust their attention to performance conditions. This implies that different contexts may produce different types of performance as learners shift their attention to meet the perceived conditions ruling in those contexts. Of course, one consequence of this might be that learners perceive computer-mediated and face-to-face tests as different contexts and adjust their attention in ways that produce very different performances. However, there is an alternative possible consequence: that the key difference which defines learners' performance is not the modality of the test, per se, but the fact that it is a test rather than a classroom activity.

The bottom half of Skehan's model (i.e., all the model except for rater, scale criteria, and score) could be used to describe the interface between competence and performance in a non-testing environment such as an L2 classroom. Just as in a test, learners' performance is influenced by their perception of the performance conditions, that is, the combination of their interlocutors and the qualities and conditions of the task. However, in a testing context, the addition of a rater who will use scale criteria to assign a score to learners' performance fundamentally changes the performance conditions for learners. The learners' goal is no longer merely to use language to communicate with their peers but is, instead, to achieve the highest rating they possibly can by impressing the rater with their language ability. An additional potential difference is in the role of affect. Bachman and Palmer's (1996) model of language performance suggests that learners' previous testing experiences may produce affective responses which could also influence performance in ways that may not occur in the classroom environment. For the majority of students, tests are inherently more nerve-racking than classroom activities. In combination, the different purpose of testing and the potential for very different affective responses vis-à-vis classroom-based CMC activities mean that it is by no means clear whether the focus on performance in a classroom environment found in virtually all previous L2 CMC research provides adequate insights into performance on CMC tests. Moreover, since both of the previous studies of CMC in language testing (Jurkowitz, 2002; Kost & Jurkowitz, 2002) have been descriptive rather than comparative, they shed little light on these issues.

The present study will attempt to fill this gap in our knowledge by comparing students' performance on a group oral exam with their performance on a computer-mediated communicative test in order to ascertain the extent to which the two performances are similar enough to allow an examiner to make inferences about oral language ability from performance in computer-mediated communication.

CHAPTER 3

METHODOLOGY

3.1 INTRODUCTION

As stated in chapter 1, this study examines the validity of a computer-mediated communicative test and, more specifically, answers the following research questions:

(1) To what extent do measures of students' intelligibility in a group oral exam correlate with measures of other criteria on a computer-mediated communicative test?

(2) To what extent do students achieve similar scores in the group oral exam and the computer-mediated communicative test?

(3) In what ways is the language produced by students on the computer-mediated test similar to or different from that produced on the group oral exam?

(4) What are the differences in students' interactions on the two tests?

(5) What are students' perceptions of the two modes of testing?

This chapter describes the methodology used to collect and code data for this validation study. The first major section of this chapter describes data collection, including the research setting, the participants in the study, the development of the CMC and group oral tests, students' preparation for both tests, and data collection during test administration. The second section is concerned with scoring the test; it describes the development of the grading rubric and the conduct of the rating sessions. The third and fourth sections describe the coding of the linguistic and interactional features of students' performance during the test.

3.2 DATA COLLECTION

3.2.1 RESEARCH SETTING

This study took place in two third-semester French classes taught at the University of Arizona during the Spring 2002 semester. Since these classes met at very different times during the day, for convenience, they will be referred to as the day class and the evening class. The evening class met twice a week (Mondays and Wednesdays) for a ninety-minute lesson. The day class met four days a week for a fifty-minute lesson. For both classes one lesson per week was taught in a computer lab (on Monday for the evening class and on Thursday for the day class). Both classes were taught by the same instructor, who had not previously taught a computer-assisted course.

One of the software programs available in the computer lab is IRC Français, a chat program that automatically generates logs of discussions and that includes those features of the French language (e.g., accents) not available in other chat programs. IRC Français allows students to meet virtually in groups by assigning them to different channels, with each channel representing one group. Although IRC Français is a relatively simple program to use, it was thought likely that using CMC in a foreign language would be such a novel experience for many participants that they would need practice to feel comfortable doing so in a testing situation. Thus, the instructor planned to use the program as often as possible during the semester.

Each class followed a common syllabus based upon the textbook *Montage* (Baker et al., 1997). This syllabus was also used by the instructors of two other French 201 classes which were taught concurrently with the classes participating in this study. The need to give students in all four classes identical exams at approximately the same point in the semester meant that the instructor had little room either to deviate from the syllabus or to add additional activities that utilized the opportunities for computer-mediated communication. Further problems arose from the fact that (1) the participating classes met a different number of times each week, (2) due to the lab schedule, the classes were always at slightly different points in the syllabus when they met in the computer lab, and (3) the scheduling of written exams sometimes meant that a class was not able to use the chat software at all when in the computer lab. As a result, although the classes both used IRC Français eight times during the semester and chat sessions were of similar length, the two classes seldom completed identical tasks through computer-mediated communication because they were never at similar points in the syllabus when they met in the computer lab. This had implications for testing (and thus for the research) that are discussed later in this chapter. Table 3.1 shows how each class used IRC Français during the semester.

3.2.2 PARTICIPANTS

Although each participating class contained 21 students, not all students elected to participate in this study. In the day class, 16 students consented to participate, but two students later chose to withdraw from the study before participation began. Because participants and non-participants would take the same tests under the same conditions (i.e., in groups of three), it was necessary to place two potential participants in a group with a non-participant, thereby removing them from the study and leaving a total of 12 participants in the day class. A greater number of students (19) from the evening class agreed to participate in the study, but, as in the day class, one student had to be placed with non-participants. This student was excluded from the data, giving a total of 18 participants.

Table 3.1

Use of CMC During the Semester

	Topic	
Week	Day Class	Evening Class
2	Getting to know each other; swapping personal information	Getting to know each other; swapping personal information
3	Students' daily schedules	No CMC
4	Talking about their own families	No CMC
5	Describing pictures	Advantages and disadvantages of working parents
6	General chat with visiting high school students	No CMC
7	No CMC	The French Revolution
8	No CMC	Poverty issues
9	Retelling a story from a cartoon	An important event from their childhood
10	Favorite restaurants and meals	Important social problems
11	No CMC	Favorite restaurants and meals
12	Giving advice to others	General interview with each other

Both classes had the same number of male participants (five), but there were more female participants in the evening class (thirteen) than in the day class (seven). The vast majority of participants were native speakers of English, but the day class contained one native speaker of Arabic while the evening class contained one native speaker each of Ewe, Spanish, and German. The average age of participants for both classes was 21 with a range of 19 to 34 for the day class and 18 to 27 for the evening class. More of the students in the day class (seven) than in the evening class (five) had spent time in French-speaking countries. For most of these students, the period spent in French-speaking countries was less than two months; however, one female participant in the evening class had spent four months in France while a male participant in the day class had lived in Paris for over two years. Appendix 1 presents the characteristics of each student.

Participants were to be tested in groups of three, which meant that four testing groups from the day class and six testing groups from the evening class would participate in the study. Although the lab schedule precluded random assignment of classes to the CMC-test-first vs. FTF-test-first conditions, all other assignments were random. Within each class, the testing groups were randomly assigned to one of two topic conditions: topic 1-first or topic 2-first (the next section describes what those topics were and how they were selected). Then each participant was randomly assigned to one of the testing groups for his or her class. Since there is theoretical (McNamara 1996) and empirical (Berry 2000) support for the notion that group composition can influence performance, each group contained the same members for both testing sessions. Table 3.2 summarizes the assignments for each class.

Table 3.2

Summary of Research Design

Class	Group	Test-Format Order		Topic Order	
		CMC	FTF	Topic 1	Topic 2
Day	A	2	1	1	2
	B	2	1	1	2
	C	2	1	2	1
	D	2	1	2	1
Evening	E	1	2	1	2
	F	1	2	1	2
	G	1	2	2	1
	H	1	2	2	1
	I	1	2	2	1
	J	1	2	1	2

3.2.3 TEST DEVELOPMENT

In general, French 201 tests were identical across all sections; the one exception was the end-of-semester oral interview, whose format and content were controlled by the instructor. For the participating classes, it was agreed that this test would be in two parts which would contribute equally to students' course grade (5% each). Thus, while both tests were of equal importance to students and represented genuine testing situations, they were relatively low-stakes tests.

Following Morrison and Lee's early research (1985), the length of most group oral tests has typically been around twenty minutes. However, the instructor participating in this study raised two objections to this test length. The first reflected the difficulty of fitting multiple twenty-minute tests into an already crowded course schedule. The second was that students at the 201 level lacked the linguistic resources and the experience with group discussions necessary to maintain a conversation for such a long time. Thus, the length of both the face-to-face and the computer-mediated test was reduced to twelve minutes.

The decision to also allow twelve minutes for the computer-mediated test requires justification as interactions were typed rather than spoken. In their native language, few individuals can type as fast as they can speak, so allowing participants no extra time under the CMC condition might, at first glance, be seen as limiting the computer-mediated test's potential to elicit an adequate sample of participant's performance. Whether this is the case, however, is an empirical matter which will be answered by this study through analysis of the amount and nature of the language produced under both conditions. Thus, the tests were of equal length.

During the initial stages of this study, it was envisaged that all the students in both classes would have discussed similar topics during their IRC Français sessions. Test prompts would then have been drawn from these topics. However, this requirement was not met. As has already been discussed (and can be seen in Table 3.1), students in the two classes used the chat software in very different ways; while the day class tended to use chat to complete language learning tasks rather than for pure discussion about a topic, the evening class tended to discuss somewhat more "serious" topics. In fact, Table 3.1 reveals that the two classes discussed the same topic only twice ("getting to know you" in week 2 and "favorite restaurants and meals" in weeks 10 and 11).

The different ways the two classes used the chat software during the course meant that test prompts could not be derived from the topics they had discussed in class. Instead, the instructor decided to draw the test questions from topics that the students had written about for their in-class tests. All of these topics were derived thematically or lexically from the contents of the course textbook, and the resulting prompts are presented in Table 3.3. In consultation with the researcher, the instructor decided to use prompts 2 and 5 for the two tests since these topics appeared to be the most conducive to the group discussion format of the tests. Both prompts dealt with issues relating to marriage and the family, and they are referred to as topic 1 (two parents in the typical family) and topic 2 (marry only once) throughout this dissertation. From a research perspective, the use of two prompts dealing with similar topics reduced the likelihood of students' performance being limited by the topic of discussion while limiting the potential for a practice effect to influence scores on the second test. In addition the prompts were structured in very similar ways: a statement with which the students were invited to agree or disagree.

3.2.4 STUDENT PREPARATION FOR THE TESTS

During the thirteenth week of the semester, the instructor posted all six potential questions on the course web site so that students could review the relevant chapters of their textbook for vocabulary and structures. At the beginning of week 14, the instructor also announced the composition of the groups, which were constituted identically for both tests. To familiarize the students with the group oral test format, a practice session was held during week 14. This replicated the conditions of the real exam faithfully. Students were given a prompt—but not one that would be used for the real test—which they discussed for ten minutes while the teacher rated their performance.

Table 3.3

Final Exam Prompts

Question	Prompt
1	Quels problèmes sociaux est-ce que vous trouvez importants ? Pourquoi ? (*What social problems do you find important? Why?*)
2	Dans la famille typique, il y a deux parents. Oui ou non ? (*The typical family has two parents. Yes or no?*)
3	Quel est l'événement historique le plus important du monde ? Pourquoi ? (*What was the most important historical event in the world? Why?*)
4	Quelles sont vos préférences culinaires ? Qu'est-ce que vous n'aimez pas manger ? (*What do you like to eat? What don't you like to eat?*)
5	On devrait se marier seulement une fois dans la vie. Oui ou non ? (*You should only marry once in your life. Yes or no?*)
6	Est-ce que vous préférez la famille traditionnelle de l'époque de vos grandparents ou la famille moderne d'aujourd'hui ? Pourquoi ? (*Do you prefer the traditional family from your grandparents time or the modern family of today? Why?*)

Groups who were participating in this study were also audio- and video-taped during their practice session to familiarize the students with the presence of the recording technology. For recording the practice session, two tape recorders were used, each of which was connected to an external microphone placed about three feet from the students; a video camera with an internal microphone was also set up about six feet from the students. On review of the tapes from the practice session, it was discovered that neither the audio- nor the video-recording had captured participants' speech with a quality or volume sufficient for transcription or rating purposes. Thus, the researcher decided that each participant in a group would wear a clip-on microphone which would only be connected to the video camera. No separate audio recording would be made. As a result, less technology was present during the real testing session than during the practice session.

Since all the students had used IRC Français multiple times during the semester, a practice computer-mediated exam was not deemed necessary.

3.2.5 TEST ADMINISTRATION

All tests were administered during the fifteenth week of the semester, the week which the syllabus reserved for oral testing. For the computer-mediated test, all class members took the test simultaneously on the day that the classes regularly used the computer lab. The instructor gave each student a slip of paper which assigned the channel they were to join (each testing group had its own channel) and the topic they were to discuss with their group members. During the test, students were not allowed to use any outside resources (dictionaries, textbooks notes, etc.) and the instructor did not interact at all with the students. After 12 minutes, the instructor asked the students to finish whatever they were typing and to log out of their channel. At this point, the software automatically produced a transcript of their interactions. The transcripts of participants were collected and stored for later analysis.

The face-to-face test also occurred during the regular class period, with each group assigned a time to be tested. However, not all students took the test in their regular classroom. During the semester, the day class had needed to leave their room promptly once class had finished to allow the following class to start on time. The potential for groups to run late (e.g., because students arrive late or take longer to leave when finished) and the time required for the recording equipment to be dismantled at the end of the testing session meant that the day class needed a space that would be available a little longer than a standard class period. Thus, the testing session for the day class was moved to a different room, which was also used for the practice session.

The face-to-face test was administered slightly differently from the computer-mediated test. Each group was tested separately. When students were ready to start and the recording equipment had been turned on, the instructor spoke the prompt aloud and the students began their discussion. Unlike the computer-mediated test, there was some interaction between the instructor and the students. For the most part, this interaction consisted of the instructor asking questions to revitalize a flagging conversation, but occasionally the instructor also provided lexical or structural items. The instructor also timed the students, allowing them approximately twelve minutes. However, for at least one group, the instructor permitted the test to go significantly beyond twelve minutes.

After both testing sessions, each participant was given a survey to complete. The survey was adapted from an earlier study by Fulcher (1996) and asked 12 attitudinal questions arranged on a Likert-scale about the test format that the students had just experienced (see appendix 2). The survey administered after the second testing session was identical to that administered after the first session except that it contained an additional question (#13) about participants' preferences regarding test format.

One male student in the evening class missed the computer-mediated test for personal reasons, which meant that his group (group J) only had two students for the computer-mediated test. This same student was present for the face-to-face test, but due to a misunderstanding, he changed groups with another student in group I so that group I did not have identical members across both testing sessions. One of the important design features of this study was that group membership would remain constant across test formats. Since this requirement had been violated, it was decided not to analyze the data from groups I and J in the evening class.

3.3 SCORING THE TESTS

3.3.1 RUBRIC DEVELOPMENT

The teacher scored each students' performance on each of the tests using the departmental grading rubric. However, these scores were not used for this study because of concern about the rubrics used by the instructor. For both exams, the instructor used an adaptation of the rubric typically used for oral interviews in the French department. Like the department's rubric, the instructor's rubrics included a range of possible scores for each level of a criterion. Thus, for the face-to-face exam, a student whose responses placed them in the category of "Does not adequately respond to question" could receive from 10 to 13 points. Having a range of scores for each level of a criterion is problematic because of the possible effect on inter-rater reliability when two raters agree that a student's response placed them in a particular category but assign different scores, which may vary by as much as three points.

In addition, there were several differences between the instructor's rubrics for the computer-mediated and face-to-face tests. The same number of points (100) was available for each test, but they were divided among a higher number of criteria for the face-to-face test because the instructor wanted to rate students' pronunciation. As a result, each criterion was weighted differently in the two tests, as can be seen in Table 3.4. Further complications were caused by the criterion of *comprehensibility*, which had four levels in the rubric for the computer-mediated test but only three levels in the rubric for the face-to-face test, and the criterion of *vocabulary*, whose descriptions for the top band differed in the two rubrics. The combination of these factors was seen as limiting the ability to effectively answer the first and second research questions, both of which may be best answered by a rubric in which the descriptors for the criteria used to rate performance on each of the tests were identical in terms of number of levels, descriptors for the levels, and relative weighting. Thus, it was decided to abandon the instructor's rubrics in favor of one which met these requirements.

One obvious solution was to adapt a rubric developed for the group oral exam. The literature on group oral tests provides examples of both holistic and analytic rating scales. Hilsdon (1991) describes a five-band holistic scale used for group oral exams in

Zambia but cautions that this scale failed to discriminate between students in the middle bands where the majority of scores fell. Thus, she recommends a scale with ten bands. The holistic scale used for Israel's Bagrut exam (Shohamy, Reves et al., 1986) contains seven bands, ranging from unintelligible in the lowest band to near-native level in the highest band. Although Shohamy et al. report a high degree of inter-rater reliability for tests which used this scale, several factors mitigated against using this scale for the present study.

First, the Bagrut exam scale was used for a high-stakes national exam where examinees' scores followed a normal distribution along the whole scale of scores (Shohamy, Reves et al., 1986). The testing context in this study, however, was a low-stakes classroom-based assessment where students were likely to cluster around a particular point on the scale. Thus, the rubric may not be sensitive enough to adequately discriminate between the students in this study.

Table 3.4

Criteria Weights in Instructor's Rubrics

Criteria	Face-to-Face	Computer-Mediated
Content	20%	25%
Comprehensibility	20%	25%
Vocabulary	20%	25%
Grammar	20%	25%
Pronunciation	20%	

A second concern relates to the usefulness of the information contained in the score. While a holistic scale produces a single composite score which might be adequate for national high-stakes exams, it cannot provide the diagnostic information valued in most classroom assessment because composite scores may not be meaningful to either raters or students, differences in sub-abilities within an individual examinee cannot be captured with this approach, and important sub-abilities may be overlooked by raters (Cohen, 1994). Both Cohen (1994) and Genesee and Upshur (1996) identify analytical scales as being more likely to provide diagnostic feedback to teachers and students. Cohen presents two additional advantages for analytical scales: important sub-skills are less likely to be collapsed together; and the explicitness of category and band descriptions facilitates rater training. Thus, this study adapted the analytical rubric used by Venugopal (1992) to rate group oral tests at a university in Malaysia.

Venugopal rated her students according to six criteria: *accuracy*, *range*, *flexibility*, *contribution*, *intelligibility*, and *effectiveness*. She defines each of these criteria as follows (Venugopal, 1992, pp. 49-50):

1. Accuracy - primarily grammatical and lexical - extent to which structures are error-free and choice of lexis correct/appropriate to context; refers also to appropriate use of idiom;

2. Range - adequacy or sufficiency of repertoire of structure and lexis; available range of language use and expression;

3. Flexibility - interactive strategies used to communicate ideas and cope with breakdowns; ability to initiate, contribute and sustain interactions;

4. Contribution - size and substantiveness of contribution; ability to provide necessary and relevant input; can range from short and/or simple utterances to fairly complex, lengthy and developed discourse;

5. Intelligibility - phonological comprehensibility; extent to which understanding is not impeded by problems of pronunciation and intonation; expectations in keeping with language situation - need not be native speaker like; some allowance to be made for residual accents;

6. Effectiveness - effectiveness of communication in terms of global communicative ability and value; coherence subsumes fluency (task accomplishment when applicable).

Venugopal's criteria raise two issues for this study. The first concerns her definition of the *effectiveness* criterion to include fluency. Fluency is an important aspect of oral language production, but it is problematic for the current study because of uncertainty regarding an appropriate definition of fluency in computer-mediated written discourse. One potential proxy measure may be the amount of language produced on the CMC test, but this is likely to be confounded by typing ability. Students with poor typing skills may struggle to express their ideas quickly because they lack the ability to input them into the computer, while good typists may be less restricted in terms of the amount of language they can produce (i.e., type) in a given amount of time. Defining fluency in terms of the number of keystrokes entered by the student does not solve this problem because such a measure is also likely to be affected by typing ability (in addition, the chat software used in this study, IRC Français, does not record this information). Therefore, although Venugopal includes fluency in her definition, the descriptors developed in this study for the *effectiveness* criterion make no reference to fluency.

The second issue results from an element found in the models of communicative competence described in chapter 2. These models include the notion of sociolinguistic competence, that is, an understanding of the appropriateness of language for the context in which it is produced. Although Venugopal's rubric does not include the criterion of *appropriateness*, it was decided not to add it to her criteria for two reasons. The first concerns the mental workload of the raters. For the group oral exam, raters would have to view a 12-minute video and make six decisions for each of the three students in the group (i.e., a total of 17 decisions). Adding an extra criteria—and thus three extra decisions—may have overburdened the raters and decreased the reliability of the scores they assigned. However, this concern does not explain why *appropriateness* was not substituted for one of the criteria in Venugopal's rubric. The answer is found in the second reason for not including *appropriateness* as a criterion—the difficulty of defining appropriate behavior for a computer-mediated communicative environment.

The difficulty lies not in providing descriptors for possible levels of the criterion, but in how those descriptors should be interpreted for computer-mediated second language production. Since *appropriateness* is context-dependent, we first have to decide what context CMC represents. One possibility is to treat CMC as a form of written language and grade it according to the standards of written language. However, as Biber (1988, 1996) has shown, written language is not a monolithic entity but a multitude of different genres, each of which has different combinations of features. In addition, studies of L1 CMC discourse have shown that CMC is a hybrid form containing features of both written and spoken language, which suggests that using any other written genre as the benchmark against which to judge the appropriateness of language produced in CMC is problematic.

An alternative possibility would be to treat CMC as a genre with its own stylistic and sociolinguistic features and to judge the students' language production according to their approximation of those features. In fact, there have been calls in the literature for classrooms to focus on the development of electronic literacy skills through specific instruction of the features of CMC discourse (Shetzer & Warschauer, 2000). If such a focus had been a goal of the participating classes in this study, the use of CMC discourse norms in judging *appropriateness* might have been justified. Since development of electronic literacy skills was, however, not a goal of instruction and the tests were not designed to evaluate students' knowledge of associated norms, the use of CMC discourse norms in rating students' language production was deemed inappropriate for the classroom assessment that is the focus of this study.

A final logical possibility is to use the norms of conversation to judge the appropriateness of students' computer-mediated discourse. From one perspective, the use of conversational norms to evaluate *appropriateness* is attractive since it could ensure that not only the criteria, but also the interpretation of those criteria were similar (if not identical) for both tests; however, from a different perspective, it is a flawed solution since it assumes that computer-mediated discourse produced in test conditions is equivalent to face-to-face discourse in a test environment. Since the goal of this study is to test this assumption, incorporating an element into the rubric which relies on an untested assumption seems invalid. A better approach, and the one that is adopted in this study, is to first establish that performance on the two tests is equivalent before defining the context which will determine judgments of sociolinguistic competence. Thus, sociolinguistic competence was not included in the rubric used for this study.

Finally, it may be argued that the criteria of *effectiveness* subsumes that of *appropriateness*, as it may do other criteria such as *flexibility*, which measures the use of interactive strategies to communicate ideas and deal with breakdowns. In other words, *effectiveness* is, itself, a measure of the overall communicative competence whose components are being measured by other criteria in the rubric. While this argument is certainly plausible at a theoretical level, it also needs to be supported empirically. Examining the independence of scores on the sub-scales will allow us to determine whether, for example, *flexibility* is subsumed by *effectiveness*. If this is indeed the case, this criteria may safely be removed from the rubric.

Although Venugopal describes her criteria quite clearly, she provides no descriptors for the bands within each category because from personal experience she feels that “the descriptors may not match the description of the individual student in that some of the variables may be absent in any one descriptor while others may be present in the student's performance" (1992, p. 50). This argument seems more relevant to global scales which include multiple variables within a single band than to the type of analytical scale proposed by Venugopal. It is also likely that since Venugopal developed her scale to use with students in her own classes, she may already have developed some unspoken notion of what types of performance would be associated with each band within a category. Since the raters in this study would not have the advantage of this knowledge when making their decisions, this researcher felt that they would require more explicit guidance. Thus, descriptors were developed for each category.

The resulting rubric was piloted by two raters in a practice rating session which used the test data from groups I and J that had been discarded from the study. Both raters were experienced college-level teachers who had taught many of the courses in the French department at the University of Arizona, including the French 201 course from which participants were drawn. One rater had previously used IRC Français while teaching at the University of Arizona.

The trial rating session produced several results which affected both the rubric and the rating process. In addition to a number of minor wording revisions necessary to clearly differentiate levels within each category, two major revisions were made to the rubric. While the raters understood most of the categories well, they experienced difficulties with the focus of the *range* category, which originally emphasized the adequacy of students' language for their communicative needs. The raters were unable to use scores to differentiate between students who were equally effective in expressing their meanings but who used markedly and qualitatively different ranges of structures and lexical items to do so. The raters also had difficulty differentiating the category of *range* from that of *effectiveness*. Thus, it was decided to remove the emphasis on communicative needs from the *range* category, which would revert to a decision solely about the range of structures and lexical items that the students demonstrated. For example, the descriptor for the middle band of the *range* category was changed from "Range of lexis and structure adequate for communicative needs" to "Adequate range of structures and lexis."

The second major revision to the rubric resulted from the teacher's interventions during the face-to-face test. The raters felt that the teacher's contribution should be reflected in one of two categories: in *flexibility* when the teacher supplied lexical items or structures to the students so that they could express an idea; and in *contribution* when the teacher asked questions to further the conversation. The rubric was adjusted accordingly. The final version of the rubric is presented in appendix 3.

Finally, the trial session also significantly clarified the *flexibility* category. Using the test data, the researcher and the raters were able to generate a list of interactive strategies from which to evaluate the students. These strategies included answering and asking questions, rephrasing another's words, suggesting lexical terms and grammatical structures, expanding on a previous utterance, initiating a topic, expressing agreement or disagreement, and using opening and closing routines.

The process for rating face-to-face tests was also revised as a result of the trial. The raters were not confident of their ability to accurately rate students based on a single viewing of the videos of the face-to-face test. Several factors appeared to contribute to this unease: an unfamiliarity with the grading rubric; the number of decisions to be made (18 decisions for each 12-minute video, that is, one per rubric category per student); and lastly, the need for raters to familiarize themselves with the voices and accents of students whom they never previously encountered (this need was exacerbated by the relatively low proficiency of the majority of the students and by the presence in some of the groups of international students with unfamiliar accents). Thus, it was decided that each video would be viewed twice.

Rating the computer-mediated tests produced a different set of problems. The non-linear nature of interactions in computer-mediated communication meant that raters had to work harder to understand the flow of ideas in order to arrive at scores for *flexibility* and *effectiveness*. Suggesting that raters read each transcript a second time alleviated this problem by allowing them to focus on establishing the relationships between utterances in the first reading and on assigning scores in the second reading.

Two further problems arose because of the textual nature of the test data. First, both raters were doctoral candidates who had conducted research on discourse. Faced with transcripts where they could read a few words and stop, their natural tendency was to overanalyze the CMC transcripts by counting occurrences of structures, interactive strategies, errors, etc. While any rater is likely to act in this way to some extent when faced with written language, this researcher felt that the raters in this study were doing so more than would be typical. However, once their method of reading transcripts had been pointed out to the raters, they agreed to adopt a more holistic approach to reading the CMC transcripts.

A second problem was that the lack of voice and visual cues made identification of participants in the CMC transcripts harder than in the face-to-face videos. One potential solution was to provide separate printouts of each student's contribution; however, this was rejected as it would increase the difficulty of following the flow of ideas and thus of assigning a score for *flexibility*. Instead, it was decided to provide visual cues through color coding the CMC transcripts. Each student's contribution would be represented using a different color.

3.3.2 RATING SESSIONS

The face-to-face and computer-mediated tests were rated in separate sessions using the revised rubric produced after the trial session (see appendix 3). At the start of each session, raters used the relevant test data from groups I and J to norm themselves through examining the degree to which they agreed with both each other's scores and the scores generated during the trial session.

For the face-to-face scoring session, raters were given a scoring sheet which (a) contained a pseudonym for each participant, (b) indicated the position of each participant relative to the camera, and (c) provided a space for the rater to write a score for each student for each of the six categories on the grading rubric. Each video was viewed twice: once to familiarize the raters with participants' voices and a second time to rate their performance. Score sheets for a group were collected from raters before viewing the next video.

For the CMC rating session, both raters worked at their own pace. They were instructed to read each transcript twice: once to establish the flow of conversation and a second time to arrive at their final ratings. Participants' names on the transcripts were replaced by a pseudonym which differed from that used for the face-to-face test so as to prevent transfer of judgment. In addition, each students' postings in the chat room were color coded to help raters identify each students' contribution to the dialogue. Scoring sheets for the computer-mediated test were changed slightly from those used in the face-to-face rating session: There was no need to indicate the students' position relative to the camera; in addition, since students produced written rather than oral texts, raters were not asked to assign a score for the *intelligibility* category which dealt with phonological comprehensibility.

3.4 TRANSCRIPTION

To facilitate analysis, the group oral exams were transcribed by the researcher. However, transcription was somewhat problematic because oral French contains a number of homonyms, particularly with regard to conjugation of verbs. For example, the infinitive and the third person plural imperfect forms of the verb *to separate* are written differently—'séparer' and 'séparaient' respectively—but are pronounced identically. Where context provided an indicator of the suitable form, that form was used in the transcription. Thus, if a student were talking about something that happened in the past and used a verb form which could, among several choices, be interpreted as a past form, the past form was transcribed. All transcripts were checked for accuracy of representation and interpretation by a native speaker of French.

3.5 CODING LINGUISTIC FEATURES

The third research question concerns the similarity of the language produced in group oral tests and computer-mediated communicative tests. To answer this question, transcripts of the face-to-face and computer-mediated tests were analyzed for a number of linguistic features: quantity of language produced, type/token ratio, lexical density, functional use of language, structural complexity, and production of errors. The following sections describe in detail how each of these variables was defined and coded.

3.5.1 QUANTITY OF LANGUAGE PRODUCED

For each condition, the total number of words produced by each student was counted using the IRC Français Chat Transcript Analysis Tool developed by College of Humanities Instructional Computing staff at the University of Arizona. This program identifies each students' production and performs a number of analyses, one of which is a word count. The program identifies a word as any group of letters that are separated by a space, a hyphen, or an apostrophe. For example, the phrase 'j'ai' (*I have*) would be counted as two words by the program. However, French includes a number of semantic units which are separated by these features but should be treated as single words, such as 'vingt-et-un' (*twenty-one*), 'd'habitude' (*usually*) and 'parce que' (*because*); thus, the program also allows its user to specify such words as exceptions to be counted as single words.

3.5.2 TYPE/TOKEN RATIO

Type/token ratios are a measure of the lexical complexity of an individual's language production. They show the ratio of the total number of different words to the total number of words. For example, the sentence "The teacher ate the apple which the student gave her" would have a type/token ratio of 0.8 because there are eight different words out of a total of ten. Higher type/token ratios are generally considered to indicate a higher lexical complexity.

Type/token ratios are sensitive to the length of the passage from which they are calculated, so the type/token ratio for each student under each condition must be determined from a sample of equal length. The original intention was to set the length of the sample so that it equaled the lowest total number of words produced by any student on either of the tests, but this was not feasible because one student produced only nineteen words on the computer-mediated test, which would have provided an insufficient sample from which to determine the type/token ratio. Thus, the sample size was set at fifty words, and the data from the three students who produced fewer than fifty words on the computer-mediated test were excluded from this analysis.

For the remaining students, the first fifty words produced under both conditions were sampled. The samples were checked to ensure that multiple spellings of the same word—for example, 'famille', 'familles', and 'famile'—were standardized. Each sample was entered into the IRC Français Chat Transcript Analysis Tool, which counted the total number of different words in the sample and divided that by the total number of words to calculate the type/token ratio.

3.5.3 LEXICAL DENSITY

Halliday (1989) differentiates between lexical items—which function in open lexical sets—and grammatical items—a closed class containing function words such as determiners, pronouns, adverbs etc. Since the lexical density is the ratio of lexical words to the total number of words, the number of lexical items on each students' face-to-face and CMC transcript were counted.

3.5.4 LANGUAGE FUNCTIONS

This analysis focused on what Bachman and Palmer (1996) call "functional knowledge," which is the same as "illocutionary knowledge" in Bachman (1990) and consists of "knowledge of four categories of language functions: ideational, manipulative, instrumental, and imaginative" (p. 69). In the *ideational* use of language, speakers express and exchange information about ideas through such speech acts as description, classification, explanation, and expression of emotion. The *manipulative* function of language is invoked whenever language is used to affect the world around us. Bachman and Palmer identify three manipulative functions: instrumental, regulatory, and interpersonal. The *instrumental* function occurs whenever we try to get something done; giving commands, issuing warnings, and making requests, promises, threats, and offers are all associated with this function. The *regulatory* function is used to control the behavior of others. It includes, but is not limited to, the use of language in rules, regulations, and laws. The *interpersonal* function is used to establish, maintain, or change relationships and includes such acts as greetings, leave-taking, giving compliments, and making insults or apologies. The *heuristic* function is used to extend our knowledge

Table 3.5

Categories of Functional Knowledge

Knowledge Type	Definition	Examples of use
Ideational	enable us to exchange information about ideas, knowledge, or feelings	descriptions, classification, explanations, expressions of emotion
Manipulative	a) instrumental functions: get others to do things for us	requests, suggestions, commands, warnings
	b) regulatory functions: control what others do	rules, regulations, laws
	c) interpersonal functions: establish, maintain & change interpersonal relationships	greetings, leave-takings, compliments, insults, apologies
Heuristic	enable us to use language to extend our knowledge of the world around us	problem-solving
Imaginative	enable us to create an imaginary world or extend the world around us for humorous or esthetic purposes	jokes, use of figurative language

about the world as we engage in learning, teaching, problem-solving, and retention of information. Finally, the *imaginative* function involves the creative use of language to imagine new worlds or using language for humorous or esthetic purposes (e.g., jokes, figurative language, and poetry). Table 3.5 summarizes the taxonomy of functional knowledge.

Face-to-face and CMC transcripts were analyzed to identify examples of speech functions which were assigned to the four categories described above: ideational, manipulative (with the sub-categories of instrumental, regulatory, and interpersonal), heuristic, and imaginative. Bachman and Palmer caution that language use typically involves multiple functions in connected discourse rather than a single function mapping onto a single utterance; moreover, a single utterance may contain several functions. This poses a problem for the present study which seeks to quantify the functional use of language under two testing environments. If functions are found across connecteddiscourse, at what point do we separate one functional use of language from another? The following example taken from the CMC data exemplifies this problem.

Example 1

S3: D'habitude, je dirai que oui, on devrait se marier seulement une fois dans la vie. Mais, je crois que il y a des situations quand c'est mieux à divorcer. Mais c'est rare! (*Usually, I'd say yes, we should marry only once in our lives. But I think there are situations where it's better to get a divorce. But that's rare.*)

In Example 1, each of the three sentences contains a different idea but share a common function. Should this ideational use of language be counted as a single use or as three uses? It was decided that the unit of analysis would be the turn. No matter how long the turn, if the primary function did not change during that turn, then the turn would be counted a single instance of that function. Thus, Example 1 was counted as a single instance of the ideational function. Example 2 provides a different situation.

In Example 2, the student starts by stating their own opinion (i.e., an *ideational* function); however, she realizes that she lacks a lexical item necessary to express her meaning and asks for the word first in English and then in French (a *heuristic* function). This turn was coded as a single instance of both the *ideational* function and the *heuristic* function.

Example 2

S15: C'est la même chose pour les enfants aussi. Ils… Seulement une parent, ce n'est pas... ce n'est pas, how do you say "enough"? comment dit-on enough? (*It's the same for the children too. They… only one parent, it's not… it's not,' how do you say enough' How do you say 'enough'?*)

One key issue here is how a turn is defined. At what point does a turn begin and end? For the computer-mediated interactions, the end of a turn was indicated by the student hitting "enter" on their keyboard, and the turn consisted of all the words that appeared on their group members' screens as a result of hitting this key. Defining turn boundaries for the face-to-face interactions was a little trickier, as the following examples illustrate.

Example 3

S1: Tout est changé. (*Everything has changed*)

S3: Oui. Tout est changé. Cela...Je crois que c'est peut-être... cinquante pour cent des [gens du—] (*Yes, everything has changed. That... I think it's perhaps fifty percent of the people of—*)

S1: [Je crois] que c'est moins. (*I think that it's less*)

S2: Moins? (*Less?*)

S1: Oui. (*Yes*)

Example 4

S1: Oui. Ça continue—(*Yes, it's continuing—*)

S3: [Oui]. (*Yes*)

S1: [jusqu'a] ce moment-là. (*right up to the present moment*)

Both of these examples illustrate a typical feature of conversation—people talking simultaneously—but the effect is different in each case. In Example 3, S1's comment *I think that it's less* interrupts S3, leaving his comment unfinished as S2 questions S1's assertion. Each comment in this example is a single turn. However, in Example 4, S3's *Yes* does not prevent S1 from finishing his idea. Although in the transcript this interaction appears to be two separate turns separated by S3's comment, this was counted as a single turn for S1 because she is able to continue her idea.

Examples 3 and 4 also illustrate a distinction that was important for classification of functions in this study. In both examples, one of the students has a turn consisting of a single word 'Oui' (*Yes*). However, these words serve different functions. In Example 3, the turn is a response to a question and represents the student's idea about the question; thus, it is classified as an *ideational* function. In Example 4, however, the student is using 'Oui' in a supportive way, which has the function of maintaining relationships among speakers; thus, it is an example of an *interpersonal* function.

While classification was mostly straightforward, a few special cases need to be discussed. In many of the groups during the computer-mediated test, one student would type the prompt. Since this typically occurred after students had taken several turns greeting each other, it appeared that the purpose for typing the question was to focus group members' attention on the topic to be discussed and to start the discussion, that is, it served a *manipulative* function. The issue was whether this use was *instrumental* or *regulatory* in nature. Bachman and Palmer point out that the categories in their taxonomy are not mutually exclusive, and there is clearly overlap between these two categories (for example, a command from a military officer may change behavior in a way that coincides with army regulations). However, from the examples that Bachman and Palmer give, a clear difference is apparent. When language is used with a *regulatory* function, the control over behavior that is exercised seems to apply to any individual who enters a particular environment so that, for example, the rules of a club apply—at least in theory—to all members of that club at all times they are present (and sometimes extend outside of the club's physical boundaries). When language is used for an *instrumental* purpose, however, the effect may be more spatially and temporally localized or more selective in its target, and it may involve greater cooperation from others. This appears to be the case here. Typing the prompt can be seen as an indirect suggestion that it is time to turn the discussion to the discussion topic. As such, it is classified in this study as an *instrumental* use of language.

Another special case concerned students' use of English at the end of turns as has already been seen in Example 2. In this example, the classification was quite easy: the student posed the question in French, leaving only the unknown lexical item in English and was thus, employing language with its *heuristic* function. Examples 5 and 6, however, illustrate different uses of English.

Example 5

S9: Nous parents est meme lawyer (*We parents is even 'lawyer'*)

S8: Avocat (*lawyer*)

Example 6

S19: Et le femme est toujours... Happy? Happy? (*And the wife is always 'happy?' 'happy?'*)

In Example 5, S9's use of an English word prompts S8 to provide the French equivalent. However, while the data contains similar examples where use of L1 elicited the L2 equivalent from either a fellow student or from the instructor, it also contains several instances of students completing a turn with an L1 word or phrase which was not supplied by other participants (as is the case with Example 6). How should the use of L1 at the end of a turn be coded? The issue here is intent. In some cases, the student provided a clear indication of their desire to solicit the necessary lexical item, either by asking for it directly as in Example 2 or by using the questioning intonation indicated in Example 6. However, the student in Example 5 does not show the same intent to solicit the necessary

L2 item. The student's utterance did not end with the intonation typical of question, and, as a result, S9's turn was not categorized as containing a *heuristic* function even though it elicited the necessary item from S8 (whose turn is heuristic in nature since she teaches S9 a new word). This can be contrasted with Example 6 where S19's turn was coded as containing a *heuristic* function, even though no L2 item was forthcoming, because the student's intonation indicates his attempt to use language to extend his knowledge of the L2. In fact, this coding scheme also applied when students used a questioning intonation to indicate uncertainty regarding the correct usage of L2 lexical items as can be seen in Example 7.

Example 7

S5: Maintenant, les gens... deux parents travaillent, mais... dix ou vingt ans depuis. Hier? (*Now, people... two parents work, but... ten or twenty years since. Yesterday?*)

T: il y a (*ago*)

Here the student is unsure how to express the idea of *ago* and expresses this uncertainty in his articulation of 'hier' which prompts the teacher to provide the necessary item. Thus, the student's usage was coded as *heuristic*.

3.5.5 STRUCTURAL COMPLEXITY

The structures of the clauses that students produced in French were coded as either phrases or as coordinating, subordinating, relative, complementary, and simple clauses. If the clause consisted of just a few words and did not contain a verb, it was classified as a *phrase*. Phrases were often associated with greetings ('bonjour' *hello*), leave-taking ('au revoir' *goodbye*), expressions of agreement and disagreement ('oui' *yes*, 'non' *no*, 'd'accord' *agreed*, and 'bien sûr' *of course*), and comments on other students' utterances ('formidable' *great*, 'très interessant' *very interesting*). Clauses which occurred in combination with other clauses were coded as *coordinating* if they contained a coordinating conjunction, *subordinating* if they contained a subordinating conjunction, *relative* if they contained a relative pronoun, and *complementary* if they were a complementary clause which followed 'que'. All independent clauses were coded as *simple* clauses. In this study, any multi-clause utterance is referred to as a *complex* sentence.

Several points should be clarified about this coding system. The first issue is how to code clauses which were relative or complementary in nature, but which lacked the appropriate marker. For instance, the utterance in Example 8 should contain a 'que' after 'Je ne pense pas'.

Example 8

S8: Je ne pense pas il y a une famille typique aujoud'hui. (*I don't think there is a typical family nowadays*)

How should this utterance be coded? Should it be a complex sentence containing two simple clauses? Or a complex sentence containing a simple clause with a complementary clause? The difficulty arises because of how English constructs sentences containing complementary and relative clauses. A null marker is possible and grammatically correct in English as can be seen in the translation in Example 8 which omits *that*. This is not the case in French, which requires the presence of the marker. In deciding how to code this utterance, we need to decide between three logically possible options:

(1) The student does not know that complementary clauses are introduced by 'que'.

(2) The student knows that complementary clauses are introduced by 'que' but omitted the marker due to a performance factor.

(3) The student knows that complementary clauses are introduced by 'que' but mistakenly believes that this can be deleted as is the case in English.

While it is impossible to tell from the transcripts which of these options apply in this case, we can see that the second clause in Example 8 functions as a complementary clause. Thus, wherever a clause functioned clearly as a relative or complementary clause but lacked the necessary marker, the clause was coded as if the marker were present so that Example 8, for example, was coded as a complex sentence containing a simple and a complementary clause.

The data also contained cases where, for example, the student used a relative clause but the relative pronoun was incorrect. In such cases, coding proceeded according to the function of the clause rather than the presence of a correct marker. This principle also applied in a single case where a student incorrectly used a relative pronoun in a place where a relative clause was both ungrammatical and functionally incorrect. The clause was not coded as a relative clause. Finally, in spoken language, and to a lesser extent in computer-mediated discourse, speakers often begin utterances with a coordinating conjunction, as in Example 9, or with a subordinating conjunction. Single clause utterances which contain coordinating or subordinating conjunctions were coded as simple sentences in this study.

Example 9

S15: C'est d'accord avec un [parent]? (*Is it OK with one parent?*)

S13: [Oui] oui. Mais ... Mais j'aime une vie avec deux parents. (*Yes, yes. But... but I like a life with two parents.*)

3.5.6 ERRORS

A native-French speaking instructor from the University of Arizona was paid to examine CMC and face-to-face transcripts for the presence of errors made in the 11 features presented in Table 3.6. All the features identified in this analysis were identified by the instructor of the participating class as having been the focus of prior instruction yet still likely to occur in the language of students in 200-level classes.

The rater's coding was checked by this researcher. This researcher agreed completely with the rater's identification of errors but disagreed with the rater's coding, in particular the classification of errors with possessive pronouns, which the rater typically coded either as incorrect lexical choice (LC) or incorrect agreement of adjectives with nouns (AG). This researcher felt, however, that these were more properly classified as pronoun errors (PRO) and changed the coding accordingly. Approximately eight percent of the errors (50/611 errors) were recoded by this researcher, of which 46 involved possessive pronouns.

Table 3.6

Features Coded for Error

Code	Error
WO	Incorrect word order
VT	Incorrect verb tense
VF	Incorrect verb form (passive vs. active; helping verb + infinitive; incorrect use of avoir/etre with the passé composé)
G	Incorrect gender of noun
SV	Subject and verb don't agree
AG	Adjectives don't agree with noun
ART	Articles are omitted; wrong choice between definite/ indefinite articles
PRO	Incorrect or missing pronoun (includes personal, reflexive, and possessive pronouns)
N	Number (e.g., singular vs. plural nouns)
PREP	Incorrect or missing preposition
LC	Incorrect lexical choice (does not include use of L1)

Section 3.4 described the problem of transcribing verbs in a language such as French which has many homonyms. Clearly, such a problem may also affect the identification and coding of verb tense errors. In checking the rater's coding, all possible transcriptions of a particular verb were considered. In two cases, a homonym resulted in an utterance that was grammatically correct and semantically logical for the context in

which the verb occurred. These cases were not coded as errors. In all other cases involving possible homonyms, it was clear that none of the possible homonyms resulted in a grammatically correct utterance, and the error was coded as a verb tense error (VT).

The number of each type of error was counted. One issue was how to count multiple errors which resulted from a single initial mistake, as is the case in Example 10 where the student has not realized that she needs a plural noun 'amis' (*friends*) instead of the singular noun 'ami' that she uses. However, while the students' choice of a singular noun is incorrect, the decisions that she makes as a result of this choice—i.e., her agreement of "tout" (*all*) and the article "le" (*the*)—are consistent with the use of a singular noun.

<u>Example 10</u>

S22 : <u>Tout</u> <u>le</u> <u>ami</u> que je connais… (*All* [singular, masculine] *the* [singular, masculine] *friend* [singular] *who I know…*)

Where multiple errors resulted from a single initial error, as is the case in Example 10, only the first error was counted. Example 11 shows a different situation where the student uses a plural noun 'parents' (*parents*) with a plural verb 'sont' (*are*), but the possessive pronoun 'ton' (*your*) and the adjective 'content' (*happy*) are both incorrectly in the singular form. Because the student has correctly produced some plural forms in this utterance, both errors were counted (as PRO and AG, respectively).

Example 11

S18: Est-ce que tu pense que ton parents sont plus content maintentant? (*Do you think that your* [singular] *parents*[plural] *are* [plural] *more happy* [singular] *now?*)

3.6 CODING INTERACTIONAL FEATURES

The fourth research question concerns the similarity of interactions on the two tests. To answer this question, transcripts of the face-to-face and CMC sessions were analyzed for patterns of turn-taking, examples of language related episodes, and use of communication strategies.

3.6.1 TURN TAKING

The number of turns that each participant took was counted for each test condition following the guidelines described in section 3.5.4. That is for the computer-mediated discourse, each posting to the chat room counted as a turn. For the face-to-face discourse, turn boundaries were typically signaled by a change in speaker unless speaker was able to continue his or her idea, as occurs in Example 4 where the two students effectively speak simultaneously.

The length of turn was also investigated. For both testing conditions, the average length of turn for each participant was calculated by dividing the total number of words produced by the number of turns that a participant took.

3.6.2 LANGUAGE RELATED EPISODES

Swain (2001) defines language related episodes (LREs) as "any part of a dialogue where students talk about the language they are producing, question their language use, or other- or self-correct their language production" (pp. 286-87). Examples of LREs in the transcripts were identified and classified as either lexis-based or form-based.

Lexis-based LREs occur whenever students focus on the meaning by searching for vocabulary or by choosing from two or more alternative words. Examples 12 and 13 both illustrate lexis-based LREs: in Example 12, the student is questioning whether the preposition 'sur' (*on*) may be used with 'le weekend' (*the weekend*) while in Example 13, the student substitutes an alternative lexical item which also means 'year'.

Example 12 (Lexical-based LRE)

S22: Pour moi je habite avec ma mère. Mais—is it?...sur le weekend, je visitais ma père pour beaucoup de ma vie. (*Me, I live with my mother. But is it? ... on the weekend, I visited my father.*)

Example 13 (Lexical-based LRE)

S22: Mon grand-parents aussi est marié, I think cinquante ans... années. (*My grandparents also is married, 'I think', fifty years... years*)

Also included as lexical-based LREs are occasions where the student uses an English word or phrase which then either prompts him/her to remember its French equivalent, as in Example 14, or elicits the correct L2 word/phrase from an interlocutor, as in Example 15. In Example 15, it is one of the other students who provides the necessary vocabulary item; however, the teacher also frequently served in this role.

Example 14 (Lexical-based LRE)

S23: C'est difficult. Difficile. (*It's 'difficult'. Difficult*)

Example 15 (Lexical-based LRE)

S9: Nous parents est même lawyer. (*We parents are even lawyer*)

S8: Avocat (*lawyer*)

Form-based LREs occur whenever students focus on the pronunciation, spelling, morphology, or syntax of the language they are producing. Examples 16, 17, and 18 illustrate form-based LREs. In Example 16, the student initially uses a masculine form of the first person singular possessive pronoun but then changes it to the correct feminine form. In Example 17, student S23 initially uses a masculine form of the adjective 'important'. After she indicates her uncertainty about the appropriateness of this form, other students discuss it briefly. Example 18 shows a form-based LRE with a syntactical focus where the student realizes her error in using the English word order of 'adjective + noun' rather than the usual French order of 'noun + adjective' and corrects herself.

Example 16 (Form-based LRE)

S16: Moi aussi. Mon... mon ma famille deux parents. (*Me too. My* [masculine form] ... *my* [masculine form] *my* [feminine form] *family two parents*)

Example 17 (Form-based LRE)

S23: Oui. Est Important. Important? (*Yes. Is important. Important?*)

S22: Importante (*Important* [feminine form]

S23: Important. (*Important* [masculine form])

S24: Important. Oui (*Important* [masculine form]. *Yes*)

Example 18 (Form-based LRE)

S2: Évidemment, mais la typique… la famille typique, ils n'ont pas deux parents. (*Obviously, but the typical… the typical family, they don't have two parents.*)

3.6.3 USE OF COMMUNICATION STRATEGIES

Communication strategies are systematic strategies that are used by individuals when they become aware that linguistic shortcomings will prevent them from expressing their intended meaning. Thus, communication strategies are compensatory in nature. Although several taxonomies of communication strategies have been proposed (Tarone 1977; Corder 1983; Faerch and Kasper 1983; Tarone 1983; Tarone, Cohen et al. 1983; Poulisse 1990), this study employs Yoshida-Morise's (1998) synthesis of several other researcher's taxonomies. Her taxonomy divides communication strategies into three broad categories: reduction, achievement, and other (see Table 3.7).

Reduction strategies are used when learners cannot represent their intended meanings and instead opt to abandon or reduce their meanings by remaining silent or changing an intended goal (topic avoidance), by abandoning the message completely (message abandonment), or by changing their intending meaning (semantic avoidance).

Table 3.7

Taxonomy of Communication Strategies

Reduction Strategies	Achievement Strategies	Other Strategies
1. Topic Avoidance	1. Approximation	1. Repair Strategies
2. Message Abandonment	• Lexical Substitution	2. Telegraphic Strategies
3. Semantic Avoidance	• Generalization	3. Fillers
	• Exemplification	4. Change of Role
	2. Paraphrase	
	• Circumlocution	
	• Word Coinage	
	• Morphological Creativity	
	3. Restructuring	
	4. Interlingual Strategies	
	• Borrowing	
	• Foreignizing	
	• Literal Translation	
	5. Cooperative Strategies	
	6. Non-Linguistic Strategies	

From Yoshida-Morise (1998), pp. 208-215.

Achievement strategies are typically used to compensate for a disparity between learners' interlanguage knowledge and the linguistic competence necessary to achieve communicative goals. Such compensation may be realized in several ways: replacement of unknown lexical items with ones believed to be semantically related to their goal (approximation); use of circumlocution, word coinage, or morphological creativity (paraphrase); construction of an alternative plan in mid-sentence (restructuring); transfer from learners' L1 through borrowing, adapting L1 words, or direct translation (interlingual strategies); direct or indirect requests for help from their interlocutor (cooperative strategies); and use of mime, gestures, and sound-imitations (non-linguistic strategies).

The third category is a catch-all category which includes learner-initiated attempts to improve communication in response to the perception that initial utterances failed to convey intended meanings (repair strategies), successful attempts to convey meaning despite message reduction (telegraphic strategies), use of fillers (fillers) and changing the role of participants, such as reverting from a respondent role to that of a questioner (change in role).

Example 19

S3: Tu as raison S2, les gens se marient trop jeunes, ou il ne pensent pas à ce qu'ils font. (*You're right, S2, people marry too young, or he don't think about what they're doing*)

S3: oops, "ILS ne pensent pas" (*oops, THEY don't think*)

For each testing condition, transcripts were examined and uses of communicative strategies were identified, coded, and counted. Several decisions were made concerning the coding process. One issue for the computer-mediated test was how to classify the two cases of self-correction, one of which is presented in Example 19. The correction of 'il' (*he*) to 'ils' (*they*) in Example 19 can be seen in two ways: as either correction of a minor typographical error or as repair of a grammatical error which the student felt may have interfered with his meaning. It is important to note that 'repair' as it applies to this analysis is not used in the sense of 'fixing a grammatical error' but rather with the meaning of 'improving communication in response to a perceived failure to convey meaning'. The CMC transcripts offer no clues here that the student is either trying to improve communication or feels that his intended message failed. His comment 'oops' may equally reflect a slight embarrassment at making such a small slip—he was a very strong student who had lived for two years in Paris—or a perception of ineffective communication. Without additional evidence, there is no way of unambiguously interpreting this comment. The other case of self-correction in the CMC data offers similar problems of interpretation. Thus, both cases were identified as ambiguous and were excluded from the count of communicative strategies.

Usage of L1 was coded in a number of ways. In Example 20, the student's use of *out of* is clearly an example of the interlingual strategy of borrowing. However, the second L1 usage here is slightly different because the student supplies the correct L2 usage immediately afterwards. It was decided to treat such cases as borrowing since the primary attempt at reaching the intended communicative goal involved the use of L1.

Example 20

S9: Je ne tombais out of amour, et then... then need to get divorced. Besoin divorcer. (*I not fall out of love, and then.. then need to get divorced. Need to divorce.*)

A more complicated issue is how to classify the L2 usage which followed the borrowing from L1. If any of the students' interlocutors had been native speakers of French, this could be classified as a repair strategy since the student cannot be assured that a native speaker of French would share her meaning if that were expressed solely in English; thus, she would supply the French equivalent to improve communication. In the interaction in Example 20, however, all the interlocutors were native speakers of English, for whom S9's intended meaning needed little further clarification once the idea had been expressed in English. Supplying the French equivalent served no communicative purpose but, instead, may have resulted from the students' perception that they should use French as much as possible since this was a testing situation in which their knowledge of the language was being assessed. For this reason, the immediate translation from L1 to L2 was not classified as a communication strategy.

The data also contain examples of students using English to translate for their interlocutors, as can be seen in Examples 21 and 22.

Example 21

S4 J'ai... une soeur jumelle. Jumelle... twin. (*I have a twin sister. Twin... 'twin'*)

Example 22

S22 C'est très... vieux. Vieux. Vieux's old. (*It's very old. Old. Vieux 'is old'.*)

In both examples, the students use the appropriate lexical item for their intended meaning, but the other group members indicate their lack of comprehension non-verbally, which results in the student repeating the problematic lexical item and then supplying the

L1 equivalent. The cause of the lack of comprehension, however, is different in each case; in Example 21, the student's pronunciation is non-standard so the other interlocutors do not understand her while the problem in Example 22 seems to occur because of the other students' unfamiliarity with the word 'vieux' (*old*). The use of L1 as a result of production difficulties such as non-standard pronunciation, as in Example 21, was classified as a borrowing since it was clear that the student could not achieve her intended goal with her available L2 resources. Even the teacher did not understand the students' pronunciation of 'jumelle' and commented 'Ah oui. Une jumelle. Oui.' (*Ah, yes, A twin. Yes.*) when the student's use of the English term allowed her to comprehend the word.

The use of L1 illustrated in Example 22 is more difficult to classify because the source of the problem is not the speaker's interlanguage. The learner has used an appropriate lexical item for her intended meaning with correct pronunciation. The misunderstanding appears to occur because the word 'vieux' (*old*) was not present in the lexicon of one of the student's interlocutors, or if it was present, it could not be retrieved during the comprehension process. Strategies such as this, which resulted from deficiencies in the audience's interlanguage, were discarded since the focus of the present study is on communication strategies that are used to compensate for deficiencies in the speaker's interlanguage.

Finally, borrowing from L1 by students could also become a cooperative strategy, as seen in Example 23, if the result of such borrowing was that interlocutors supplied the L2 equivalent.

Example 23

S9 Nous parents est même lawyer. (*We parents are even lawyer*)

S8 Avocat (*lawyer*)

When all communication strategies had been coded and counted, an index of communication strategy use was calculated for each individual for both testing conditions by dividing the number of each type of communication strategy by the total number of words produced by the individual.

3.7 CONCLUSION

This chapter outlined the data collection and coding methods used to gather the evidence necessary to answer the research questions that were discussed in chapter 1 and restated at the beginning of this chapter. The next chapter describes the statistical and qualitative analyses that were conducted on the data and presents the results of those analyses.

CHAPTER 4

DATA ANALYSIS AND RESULTS

4.1 INTRODUCTION

This chapter describes the analyses that were completed on the data and presents the results of those analyses. Although the chapter also includes preliminary discussion of the results wherever it appears timely, the focus here is primarily on the analysis and results. Chapter 5 contains a detailed discussion of the implications of the results for the research questions and for language testing.

4.2 STATISTICAL PROCEDURES

Many of the comparisons described in this chapter require the use of Analysis of Variance (ANOVA). Since this study compares some aspect of students' performance on a computer-mediated test with the same aspect of their performance on a face-to-face test, it would be appropriate to make the comparisons with a Repeated Measures ANOVA. However, such an analysis cannot be performed using the statistical package (SPPS) that was available to this researcher. SPSS was not flexible enough to complete the Repeated-Measures ANOVA analyses required by this study's design because the software requires a fully repeated design in which each student received each test prompt in each test condition. This was not a feasible option for this study since data collection was integrated into a regular class which had limited time for testing. As a result of this software limitation, all *F* ratios were calculated using a Between-Subjects ANOVA with test method (computer-mediated versus face-to-face) as an independent variable. In other words, data was analyzed as if the students who took the computer-mediated test were different from those who took the face-to-face test.

Finally, one problem whenever a study conducts multiple statistical tests is the inflated risk of committing a Type I error. To control for this risk in the present study, the Bonferroni technique was used to calculate an alpha level of .005 (.25/52 rounded to three decimal places).

4.3 TEST SCORE ANALYSIS

As described in chapter 3, two raters scored the computer-mediated and face-to-face tests using a rating scale which contained five criteria: *accuracy* (of grammar and lexis), *range* (the adequacy of structure and lexis), *flexibility* (the use of interactive strategies), *contribution* (in terms of size and substantiveness), and *effectiveness* (global communicative ability). On the face-to-face test, the raters also assigned a score for the additional criteria of *intelligibility* (the extent to which phonology interferes with comprehension). Total score on the CMC test (hereafter referred to as CMC1) was calculated by summing the scores on the five sub-scales used to rate student performance on the test (i.e., *accuracy*, *range*, *flexibility*, *contribution*, and *effectiveness*). For the face-to-face test, two total scores were calculated, which will be referred to as FTF1 and FTF2. FTF1 summed the five sub-scales common to the face-to-face and the computer-mediated test (i.e., *accuracy*, *range*, *flexibility*, *contribution*, and *effectiveness*) while FTF2 included the additional criteria of *intelligibility*.

4.3.1 INTER-RATER RELIABILITY

Table 4.1

Inter-Rater Reliability Coefficients

		Correlation Coefficient (*r*)	
		Computer-Mediated	Face-to-Face
Total Score [a]		.57*	.68*/ .73*
Sub-Scales:	Accuracy	.48*	.76*
	Range	.44	.80*
	Flexibility	.46	.27
	Contribution	.59*	.65*
	Effectiveness	.47	.75*
	Intelligibility	N/A	.42

Note: [a] Correlations for both FTF1 and FTF2 are provided (FTF1/FTF2).
* p< .005

The extent to which the two raters agreed in their assignment of total scores and individual sub-scale scores was determined using a Pearson correlation. The reliability coefficients are presented in Table 4.1.

Table 4.1 indicates that while the raters attained a reasonable level of agreement on the total score for FTF1 ($r = .73$), correlations on the two other total scores were lower. For FTF1, the correlation coefficient, *r*, was .68 while the correlation between the total scores assigned for the computer-mediated test (CMC1) was even lower ($r = .57$). This pattern of greater agreement between raters for the face-to-face test over the computer-mediated test is repeated for all but one of the sub-scales. The exception is the *flexibility* sub-scale where the raters' scores on the face-to-face test showed a low correlation (r = .27) that was much lower than the correlation achieved on the computer-mediated test (r = .46).

4.3.2 INDEPENDENCE OF SUB-SCALES

The independence of the sub-scales was investigated for each test by combining each raters' scores and calculating Pearson correlations between scores on each sub-scale. Correlation coefficients for the computer-mediated (CMC) and face-to-face (FTF) tests are presented in Table 4.2.

It is to be expected that the scores for the most global subscale, *effectiveness*, would show high correlations with other sub-scales. This is supported by the correlation coefficients in Table 4.2 which show that scores for *effectiveness* have a very strong correlation with scores for *range* and *contribution* on both the face-to-face and the computer-mediated test. However, the scores for *effectiveness* have only a moderately strong correlation with scores for *accuracy* and *contribution.* All other correlations between sub-scales are either moderate (defined here as falling in the range, r = .3 – .5) or moderately strong (defined as r =.5 – .8). These correlations appear to justify the use of a multi-dimensional rating system—i.e., one that includes several sub-scales—although the strong correlations of *effectiveness* with *range* and *contribution* suggest that the latter two sub-scales could perhaps be eliminated to ease rater workload without resulting in construct under-representation.

Table 4.2

Inter-Subscale Correlations

	Range	Flexibility	Contribution	Effectiveness	Intelligibility
CMC Test					
Accuracy	.57*	.52	.42	.64*	N/A
Range		.47	.74*	.82*	N/A
Flexibility			.61*	.59*	N/A
Contribution				.89*	N/A
FTF Test					
Accuracy	.60*	.43	.41	.63*	.65*
Range		.60*	.72*	.89*	.55*
Flexibility			.75*	.65*	.39
Contribution				.82*	.43
Effectiveness					.59*

Note: * p < .005

4.3.3. RESEARCH QUESTION 1

The first research question asked the extent to which measures of students' intelligibility on the group oral exam correlate with scores on the computer-mediated communicative test. To answer this question, Pearson correlations were calculated between the *intelligibility* scores assigned on the face-to-face test and the total score assigned to the computer-mediated test performance as well as each of the other five sub-scales used to compute the total score on the computer-mediated test. Because of the relatively low inter-rater reliability reported in section 4.3.1, each rater's scores were analyzed separately. Table 4.3 presents the Pearson correlations that were obtained from this analysis. Table 4.3 indicates that the correlations between *intelligibility* and the scores on the CMC test are quite low for rater 1 and extremely low for rater 2. For rater 1, the highest correlation was between *intelligibility* and *overall effectiveness* (*r* = .51), while for rater 2, scores on the *flexibility* (*r* = .20) and *range* (*r* = .20) sub-scales achieved the highest levels of correlation with *intelligibility*. These low correlations suggest that *intelligibility* cannot be predicted with a high degree of certainty from scores attained on a computer-mediated test and must be measured separately if the goal of testing is to make inferences about learners' oral language ability.

Table 4.3

Correlation Between Intelligibility and CMC Sub-Scale Scores

	Intelligibility	
	Rater 1 (*r*)	Rater 2 (*r*)
CMC - Accuracy	.47*	.13
CMC - Range	.34	.20
CMC - Flexibility	.22	.20
CMC - Contribution	.40	.05
CMC - Overall Effectiveness	.51	.00
CMC – Total Score (CMC1)	.45	.15

Note: All correlations were non-significant at the .005 level.

4.3.4 RESEARCH QUESTION 2

Table 4.4

Analysis of Variance for Test Scores

Source	*df*	SS	MS	*F*
Rater 1				
Test Method (M)	1	20.02	20.02	0.50
Prompt (P)	1	7.52	7.52	1.33
M x P	1	6.02	6.02	0.40
Error	44	660.25	15.00	
Rater 2				
Test Method (M)	1	1.69	1.69	0.01
Prompt (P)	1	0.02	0.02	0.10
M x P	1	38.52	38.52	2.22
Error	44	762.75	17.34	

Note: All *F* values were non-significant at the .005 level.

The second research question asked: To what extent do students achieve similar scores in the group oral exam and the computer-mediated communicative test? Total score on the computer-mediated (CMC1) and the face-to-face test (FTF1) were compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables. Because of the low level of inter-rater reliability, each rater's scores were run as a separate analysis. Table 4.4 presents the results of the ANOVA analyses.

Neither analysis found statistically significant differences for the main effects of test method and test prompt, or for a first-order interaction effect. The main effect of test method yielded non-significant *F* ratios for rater 1 (*F* (1, 44) = 1.33, *p* = .25) and for rater 2 (*F* (1, 44) = 0.10, *p* = .78). The main effect of test prompt also yielded non-significant *F* ratios for rater 1 (*F* (1, 44) = .50, *p* = .48) and for rater 2 (*F* (1, 44) = 0.01, *p* = .97). Finally, the interaction between test prompt and test method also yielded statistically non-significant results for both rater 1 (*F* (1, 44) = 0.40, *p* = .53) and for rater 2 (*F* (1, 44 = 2.22, *p* = .14).

4.4 LINGUISTIC FEATURE ANALYSIS

The third research question concerns the ways in which the language produced by students on the computer-mediated test is similar to or different from that produced on the group oral exam. To answer this question, several variables were compared across the two tests: the amount of language produced, individual type/token ratios, lexical density, language functions, structural complexity, and the level of errors. The results of these comparisons are presented in sections 4.4.1 to 4.4.6. Individual data for all of these variables except for language functions are presented in appendix 4.

4.4.1 QUANTITY OF LANGUAGE PRODUCED

Two separate measures of language production were compared. The first examined the total number of words produced regardless of the language chosen while the second limited itself to the total number of French words produced during the tests.

Table 4.5

Analysis of Variance for Word Production

Source	*df*	SS	MS	*F*
	Total Number of Words			
Test Method (M)	1	344424.08	344424.08	17.93*
Prompt (P)	1	4524.0	4524.0	0.24
M x P	1	23056.33	23056.33	1.20
Error	44	845442.17	19214.60	
	Number of French Words			
Test Method (M)	1	299094.18	299094.18	16.06*
Prompt (P)	1	6888.02	6888.02	0.37
M x P	1	27408.52	27408.52	1.47
Error	44	1152743.81	24526.4	

Note: * p < .005.

Both measures were compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables. The results of the ANOVA analyses are presented in Table 4.5.

Table 4.5 shows that for the total number of words produced, the main effect of test method, *F* (1, 44) = 17.93, *p* < .001, was significant at an alpha level of .005. Students produced significantly more language on the face-to-face test (M = 261.38, SD = 190.56) than they did on the computer-mediated test (M = 91.96, SD = 40.55). Neither test prompt, *F* (1,44) = 0.24, p = .63, nor the interaction effect, *F* (1, 44) = 1.20, *p* = .28, yielded significant *F* ratios at the .005 level.

Table 4.5 shows a similar pattern in the results for the total number of L2 words produced. The main effect of test method, *F* (1, 44) = 16.06, *p* < .001, was significant at an alpha level of .005. In other words, students produced significantly more French on the face-to-face test (M = 249.38, SD = 188.30) than they did on the computer-mediated test (M = 91.50, SD = 40.75). Neither test prompt, *F* (1,44) = 0.37, *p* = .55, nor the interaction effect, *F* (1, 44) = 1.47, *p* = .23, yielded significant *F* ratios at the .005 level.

Table 4.6 presents descriptive statistics for the word production data. It is clear from Table 4.6 that the range for L2 word production on the face-to-face test is a lot higher (874) than that for the computer-mediated test (164). In fact, the difference in magnitude between the ranges is much larger than the difference in magnitude between the means for the two tests. The range for the face-to-face test is over five times larger than that of the computer-mediated test while the mean for the face-to-face test is only 2.7 times larger than the mean for the computer-mediated test. Examining the individual data presented in appendix 4, it is clear that the student who produced the most L2 words on the face-to-face (S3) test is an outlier. This student's exceptional level of L2 oral production is probably the result of having spent over two years living in France, which gave him a fluency that no-one else in the class could match (the second highest L2 production was 592 words). However, even if this student's data is excluded from the analysis, the difference in mean L2 production between the computer-mediated test (M =

87.78, SD = 37.27) and the face-to-face test (M = 220.70, SD = 128.18) remains statistically significant, *F* (1, 42) = 21.88, p < .001.

Finally, the individual data presented in appendix 4 shows that over half of the participants (15 out of 24) failed to produce 100 words on the computer-mediated test; in comparison, just three students produced fewer than 100 words on the face-to-face test. Such universally low levels of language production have important implications for this study, which will be discussed in chapter 5.

Table 4.6

Descriptive Statistics for Word Production

	Mean	SD	Min	Max
	Total Words Produced			
CMC	91.96	40.55	19	183
FTF	261.38	190.56	41	912
	L2 Words Produced			
CMC	91.50	40.75	19	183
FTF	249.38	188.30	35	909

4.4.2 TYPE/TOKEN RATIO

Type/token ratios are measures of lexical complexity which show the ratio of the total number of different words to the total number of words based on a sample of uniform length (50 words). The type/token ratio was compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables. Table 4.7 presents the results of this analysis.

The main effect of test method, *F* (1, 38) = 9.46, *p* = .004, was significant at an alpha level of .005. Students had a higher type/token ratio on the computer-mediated test (M = 0.72, SD = 0.07) than they did on the face-to-face test (M = 0.65, SD = 0.08). In other words, students produced lexically more complex language on the computer-mediated test than they did on the face-to-face test. Neither test prompt, *F* (1, 38) = 1.03, *p* = .31, nor the interaction effect, *F* (1, 38) = 2.48, *p* = .12, yielded significant *F* ratios at the .005 level.

Table 4.7

Analysis of Variance for Type/Token Ratio

Source	*df*	SS	MS	*F*
Test Method (M)	1	0.045	0.045	9.46*
Prompt (P)	1	0.005	0.005	1.03
M x P	1	0.012	0.012	2.48
Error	38	0.181	0.005	

Note: * p < .005.

Table 4.8

Descriptive Statistics for Type-Token Ratio

	Mean	SD	Min	Max
CMC	0.72	0.07	0.52	0.80
FTF	0.65	0.08	0.48	0.78

Table 4.8 presents the descriptive statistics for the type-token ratio (TTR). The higher mean for the computer-mediated test data is reflected in consistently higher values for individuals. Nineteen of the twenty-four participants achieved higher type-token ratios on the computer-mediated test than they did on the face-to-face test, with three individuals achieving the same TTR on the two tests, and just two participants achieving higher TTRs on the face-to-face test. Finally, one student's type-token ratio on the face-to-face test (TTR = 0.48) was a lot lower than that on the computer-mediated test (TTR = 0.78). This appears to be due to the presence of two false starts in the oral data where the student started an utterance and repeated the first few words. Since the TTRs in this study are based on a smaller sample (50 words) than that traditionally used (100 words), the impact of these false starts on the value for the type-token ratio is magnified.

4.4.3 LEXICAL DENSITY

Lexical density is a measure of the degree of orality versus literacy in a text, where higher lexical densities are associated with written texts. Lexical density is calculated as the ratio of lexical items—i.e., those words that occur in open classes such as verbs and nouns—to the total number of words. The lexical density of the language produced by students was compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables. The results of this analysis are presented in Table 4.9.

Table 4.9

Analysis of Variance for Lexical Density

Source	*df*	SS	MS	*F*
Test Method (M)	1	0.869	0.869	144.24*
Prompt (P)	1	0.010	0.010	1.59
M x P	1	0.001	0.001	0.09
Error	44	0.265	0.006	

Note: * p < .005.

The main effect of test method, *F* (1, 44) = 144.24, *p* < .001, was significant at an alpha level of .005. The language students produced on the computer-mediated test had a higher lexical density (M = 0.44, SD = 0.07) than the language produced on the face-to-face test (M = 0.16, SD = 0.09). That is, the language produced by students on the computer-mediated test tended to be more literate than that produced on the face-to-face test in the sense that the former contained a higher number of lexical items. Neither test prompt, *F* (1, 44) = 1.59, *p* = .21, nor the interaction effect, *F* (1, 44) = 0.09, *p* = .77, yielded significant *F* ratios at the .005 level. Table 4.10 presents descriptive statistics for lexical density. The data for each individual in appendix 4 indicates that the computer-mediated discourse of all participants demonstrated greater use of lexical items than did the face-to-face discourse.

Table 4.10

Descriptive Statistics for Lexical Density

	Mean	SD	Min	Max
CMC	0.44	0.07	0.20	0.54
FTF	0.16	0.09	0.03	0.36

4.4.4 LANGUAGE FUNCTIONS

Speech functions were assigned to four categories: *ideational*, *heuristic*, *imaginative*, and *manipulative* (with sub-categories of *instrumental*, *regulatory*, and *interpersonal*). Table 4.11, which presents the relative occurrence of each category and sub-category, shows that the most frequently found function on both tests was the *ideational* function. On the computer-mediated test, over 70% of functions were *ideational*, while an even higher proportion of functions (77.6%) were of this type on the face-to-face test. This high proportion is not surprising. The *ideational* function occurs whenever speakers exchange and express information about ideas, which is the task that students were set on both tests where they discussed their opinions about marriage and the family.

Table 4.11 also shows the almost complete absence of examples of the *heuristic* function in the computer-mediated data. Only on two occasions do students on the computer-mediated test appear to use language to extend their knowledge of the world through learning, teaching, retention of information etc. In comparison, 7.3% of the functions found on the face-to-face test were *heuristic*. From examining transcripts of the face-to-face tests, the majority of these instances of the *heuristic* function occurred when students lacked the necessary French lexical item. The student either directly solicited the French term or used the equivalent English term, which an interlocutor translated and, thus, took on the role of teacher of the French term. It would appear that students on the computer-mediated test did not engage in this type of behavior. They did not ask others for unknown French lexical items, and on the few occasions when they were forced to resort to using English, their interlocutors did not provide French equivalents.

The computer-mediated test produced a higher proportion of *interpersonal* functions than did the face-to-face test. One quarter of the functions found on the face-to-face test were *interpersonal* compared with about 15% of the functions on the face-to-face test. However, these proportions are a little misleading for two reasons. First, on the computer-mediated test, seven of the eight groups engaged in the act of greeting while three of the groups also engaged in leave-taking at the end of the test. Such acts fall under the category of *interpersonal* and account for 52 of the 67 instances of *interpersonal* functions in the computer-mediated data (40 for greetings and 12 for leave-taking). However, students did not engage in the acts of greetings and leave-taking during the face-to-face test because these acts occurred outside of the testing context while students were waiting for their group's turn or as they were leaving the room after the test had finished.

Table 4.11

Occurrence of Speech Functions

	Number (%^{a})	
Function	Computer-Mediated	Face-to-Face
Ideational	185 (70.3%)	412 (77.6%)
Heuristic	2 (0.1%)	39 (7.3%)
Imaginative	0	0
Manipulative		
Instrumental	9 (3.4%)	1 (0.1%)
Regulatory	0	0
Interpersonal	67 (25.5%)	79 (14.9%)

Note: [a] Percentages do not sum to 100% because of rounding.

The second reason that the results for the *interpersonal* data should be treated with caution arises from the significantly lower amount of language produced on the computer-mediated test. In section 4.4.1, we saw that participants produced, on average, 2.75 times as much language during the face-to-face test as they did during the computer-mediated test. As a result of the lower levels of overall language production, students used about half as many functions in total on the computer-mediated test (263) as they did on the face-to-face test (531). Unlike other examples of acts that are classified as *interpersonal* (e.g., giving compliments, making insults, and making apologies), examples of greeting and leave-taking are not likely to increase as the amount of language produced increases. These acts will typically occur only once in a test, but they will seem to be more important when students' levels of language production are relatively low, as is the case here with the computer-mediated test. Thus, what is interesting in the data is not the fact that the computer-mediated test produced proportionally higher numbers of the *interpersonal* function, but that those examples included the acts of greeting and leave-taking, which were not found in the face-to-face test.

Given the misleading nature of the count of *interpersonal* functions in the computer-mediated data, it was decided to re-calculate the proportion of *interpersonal* functions when functions involving greetings and leave-taking are excluded. The results of this re-calculation reversed the result of the previous analysis. Now, *interpersonal* functions occur twice as frequently in the face-to-face test, where they account for 14.9% of all functions, as the do in computer-mediated test, where they occur for only 7.1% of all functions.

Finally, the results in Table 4.11 indicate that students used the *instrumental* function much more on the computer-mediated test (nine instances) than on the face-to-face test (one instance). Of the nine examples of the *instrumental* function on the computer-mediated test, eight appear to be the result of the testing context. We have already seen that during the computer-mediated test, the majority of students engaged in greeting routines. These routines ended when one of the students specifically forced the group as a whole to turn their attention to the test prompt by asking whether they should start and/or by typing the prompt in its entirety. This did not happen during the face-to-face test because the instructor indicated that the test should start by asking the prompt question herself.

4.4.5 STRUCTURAL COMPLEXITY

Each utterance produced by a student was classified as either a phrase, a simple sentences or a complex sentence (i.e., a multi-clause utterance containing at least one relative, complementary, subordinate, or coordinate clause). Two indicators of structural complexity were determined for each test: (1) the Coordination Index (CI), and (2) the Complexity Ratio (CR).

The Coordination Index measures the degree of complexity found in each students' complex sentences and was calculated by dividing the total number of clauses a student used in complex sentences by the total number of complex sentences the student produced. Given the definition of a complex sentence as a multi-clause utterance, this indicator is always equal to at least 2.0; however, values greater than 2.0 indicate that participants produced at least one utterance with more than two clauses. Thus, higher values for the Coordination Index are associated with greater sentence complexity.

The Coordination Index was compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables. The results of this analysis (Table 4.12) showed no significant *F* values for the main effects of test method (F (1, 44) = 2.30, p = .14) and test prompt (F (1, 44) = 1.16, p = .29) , or for the interaction effect (F (1, 44) = 3.95, p = .05) at the .005 level. Thus, neither test method nor test prompt affected the number of clauses contained in the complex sentences that the students produced.

The Complexity Ratio measures the ratio of complex to non-complex sentences and is calculated by summing the number of phrases, simple sentences, and complex sentences and dividing the result by the number of complex sentences. Lower values on this measure indicates a greater proportion of complex sentences in a student's language production.

Table 4.12

Analysis of Variance for Measures of Structural Complexity

Source	*df*	SS	MS	*F*
		Coordination Index		
Test Method (M)	1	0.572	0.572	2.30
Prompt (P)	1	0.288	0.288	1.16
M x P	1	0.980	0.980	3.95
Error	44	10.926	0.248	
		Complexity Ratio		
Test Method (M)	1	0.112	0.112	0.03
Prompt (P)	1	0.010	0.010	0.01
M x P	1	14.520	14.520	3.95
Error	42	154.410	3.676	

Note: All *F* values were non-significant at the .005 level

The Complexity Ratio was compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables. The results of this analysis (Table 4.12) showed no significant *F* values for the main effects of test method (*F* (1, 42) = 0.03, *p* = .86) and test prompt (*F* (1, 42) = 0.01, *p* = .96), or for the interaction effect (*F* (1, 42) = 3.95, *p* = .05) at the .005 level. Thus, students produced single-clause and multi-clause utterances in similar proportions regardless of test method or test prompt.

Table 4.13 presents the descriptive statistics for the two measures of structural complexity. The individual data presented in appendix 4 shows that while every student produced a complex sentence in the face-to-face test, two students failed to do on the computer-mediated test. However, the low production of both students on the computer-mediated test (19 and 44 words, respectively) suggests that they may have lacked an opportunity to use complex sentences. The individual values for the Coordination Index also show that the majority of students produced at least one sentence that was more complex than the basic two-clause combination; in other words, their Coordination Index exceeded 2.0. Thirteen students had a Coordination Index greater than 2.0 for the computer-mediated test while on the face-to-face test, sixteen students achieved this. Thus, it would appear that students were able to construct rather complex utterances regardless of the medium in which they were tested.

Table 4.13

Descriptive Statistics for Structural Complexity Measures

	Coordination Index				Complexity Ratio			
	M	SD	Min	Max	M	SD	Min	Max
CMC	2.03	0.68	0	3.0	3.11	1.96	0	8.0
FTF	2.25	0.26	2.0	2.78	3.28	2.15	1.29	11.25

Table 4.14

Frequency of Clause Types in Complex Sentences

	Number (%)	
	Computer-Mediated	Face-to-Face
Simple Clauses	86 (44.6%)	198 (42.8%)
Relative Clauses	15 (7.8%)	40 (8.6%)
Complementary Clauses	36 (18.7%)	72 (15.6%)
Coordinating Clauses	38 (19.7%)	91 (19.7%)
Subordinating Clauses	18 (9.2%)	62 (13.3%)
Total	193 (100%)	463 (100%)

Although the test methods produced no statistically significant differences in either the relative numbers of complex versus non-complex sentences or in the average number of clauses in complex sentences, it is possible that students tended to produce complex sentences containing more of a particular type of clause (i.e., simple, relative, complementary, subordinate, or coordinate clauses). Table 4.14, which examines the frequency with which five types of clauses were produced on each test, indicates that in general, the frequency of each type of clause was very similar across the two tests.

4.4.6 ERRORS

The total number of errors produced by each student was determined by summing the number of each of the eleven error types described in section 3.5.6: Word order, verb tense, verb form, noun gender, subject/verb agreement, noun/adjective agreement, article choice, pronoun use, number, preposition, and lexical choice.

Since the number of errors committed by participants is likely to increase as a function of the amount of language produced, a direct comparison of the total number of errors without normalizing the error counts to allow for differential levels of language production, both between individuals and between tests, would return a false result. The results presented in section 4.4.1 showed that participants produced, on average, over two and a half times as much French on the face-to-face test (M = 249.38) than they did on the computer-mediated test (M = 91.50). Thus, for each student, an error ratio was calculated to show the number of errors per ten L2 words produced.

Error ratios were compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables. The results of this analysis are presented in Table 4.15.

Table 4.15

Analysis of Variance for Error Ratio

Source	*df*	SS	MS	*F*
Test Method (M)	1	2.54	2.54	10.42*
Prompt (P)	1	0.12	0.12	0.49
M x P	1	0.01	0.01	0.06
Error	44	10.75	0.24	

Note: * p < .005.

The main effect of test method, *F* (1, 44) = 10.42, *p* = .002, was significant at an alpha level of .005. The error ratio for the computer-mediated test was higher (M = 1.18, SD = 0.60) than the error ratio for the face-to-face test (M = 0.72, SD = 0.33). That is, the language produced by students on the computer-mediated test tended to contain more errors than that produced on the face-to-face test. Neither test prompt, *F* (1, 44) = 0.49, *p* = .49, nor the interaction effect, *F* (1, 44) = 0.06, *p* = .81, yielded significant *F* ratios at the .005 level.

Table 4.16

Descriptive Statistics for Error Ratio

	Mean	SD	Min	Max
CMC	1.18	0.60	0.22	2.55
FTF	0.72	0.33	0.28	1.47

Table 4.17

Comparison of Error Type Frequencies

	Computer-Mediated [a]	Face-to-Face [a]
Word Order	0.50	0.14
Verb Tense	1.36	1.24
Verb Form	0.63	0.53
Gender	0.91	0.81
Subject/Verb Agreement	1.09	0.53
Adjective Agreement	1.04	0.19
Articles	0.73	0.45
Pronouns	0.95	0.91
Lexical Choice	2.27	0.96
Number	0.45	0.19
Preposition	0.40	0.19

Note: [a] Numbers represent the frequency of each type of error per 100 words produced by all students on the test.

One question which the preceding analysis does not answer is whether students tended to produce more of each type of error in the computer-mediated test or whether particular types of error were more prevalent in the computer-mediated discourse than in the face-to-face discourse. This question can be answered by converting the number of each type of error to an Error Type Ratio which takes into account differing amounts of language production on the two tests. Thus, for each test method (computer-mediated and face-to-face), the total number of each type of error was expressed in terms of its occurrence per 100 words produced: Since students produced a total of 2207 words on the computer-mediated test, the total number of instances of word order errors on that test (11) was divided by 22.07, giving a ratio of 0.50. Error Type Ratios for each of the error types on both tests are given in Table 4.17.

Although all types of errors have higher Error Type Ratios for the computer-mediated test than for the face-to-face test, the differences are not particularly large for errors involving verb tense, verb form, gender, articles, and pronouns. However, errors involving lexical choice or number occur more than twice as often in the computer-mediated test as in the face-to-face test while word order errors occur almost three times as often in computer-mediated discourse as in the face-to-face discourse. The greatest difference is found for errors involving incorrect agreement of adjectives with nouns, which occur more than five times as often in the computer-mediated data than in the face-to-face data.

4.5 INTERACTION ANALYSIS

The third research question examined the ways in which students' interactions differed on the computer-mediated and face-to-face tests. In answering this question, turn-taking, language-related episodes, and the use of communication strategies were examined. Sections 4.5.1 to 4.5.3 examine each of these variables in turn. Individual data for turn-taking and use of communication strategies is found in appendix 5.

4.5.1 TURN TAKING

Turn taking was examined in terms of two variables: the number of turns taken on each test and the average length in number of words of those turns. Both variables were compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables. Table 4.18 presents the results of both analyses.

Table 4.18

Analysis of Variance for Turn-Taking

Source	*df*	SS	MS	*F*
		Number of Turns		
Test Method (M)	1	1131.02	1131.02	31.79*
Prompt (P)	1	67.69	67.69	1.90
M x P	1	38.52	38.52	1.08
Error	44	1565.25	35.57	
		Average Turn Length		
Test Method (M)	1	185.75	185.75	6.03
Prompt (P)	1	34.10	34.10	1.11
M x P	1	116.53	116.53	3.78
Error	44	1355.91	30.82	

Note: * p < .005.

Table 4.19

Descriptive Statistics for Turn-Taking

	Number of Turns				Turn Length			
	M	SD	Min	Max	M	SD	Min	Max
CMC	11.04	3.48	4	18	8.68	3.99	4.23	19.67
FTF	20.75	7.78	9	37	12.61	7.04	3.42	32.57

For number of turns, the main effect of test method, *F* (1, 44) = 31.79, *p* < .001, was significant at an alpha level of .005. Students had almost twice as many turns on the face-to-face test (M = 20.75, SD = 7.78) as they did on the computer-mediated test (M = 11.04, SD = 3.48). Neither test prompt, *F* (1, 44) = 1.90, *p* = .18, nor the interaction effect, *F* (1, 44) = 1.08, *p* = .30, yielded significant *F* ratios at the .005 level.

Table 4.18 also presents the results for average length of turn. The main effect of test method, *F* (1, 44) = 6.03, *p* = .02, was not significant at an alpha level of .005. Neither test prompt, *F* (1, 44) = 1.11, *p* = .29, nor the interaction effect, *F* (1, 44) = 3.78, *p* = .06, yielded significant *F* ratios at the .005 level.

4.5.2 LANGUAGE RELATED EPISODES

A language-related episode (LRE) occurs whenever students focus on the language they produce. Thus self- or other-correction, discussion of the language being produced, or questioning of language are all counted as language-related episodes and are classified according to whether the focus is on vocabulary (lexis-based LREs) or on form (form-based LREs).

Table 4.20

Occurrence of Language-Related Episodes

	Computer-Mediated [a]	Face-to-Face [a]
LRE (Form-Based)	4	35
LRE (Lexis-Based)	1	52
Total LRE	5	87

Note: [a] Numbers represent actual occurrence of LREs.

A language-related episode may involve multiple interlocutors as is the case when a student is unsure of the correct lexical item to use in a sentence and solicits the item from one or more of his/her interlocutors. Thus, the appropriate level of analysis is not the individual, but the group. Since it is reasonable to assume that the number of LREs may be a function of the amount of language produced, two ratios were calculated. The lexis-ratio represented the number of lexis-based LREs which occurred per 100 words produced by the group as a whole. The form-ratio represented the number of form-based LREs which occurred per 100 words produced by the group. Summing the form-ratio and the lexis-ratio produced the LRE-ratio, an overall measure of the number of LREs per 100 words produced by the group.

The original plan to conduct comparisons on all three ratios had to be modified because only one group engaged in lexical-based LREs on the computer-mediated test (see Table 4.20). Thus, only the LRE-ratio was compared using a correlated groups *t*-test at an alpha level of .005. This test was found to be statistically significant, *t* (7) = -4.25, p = .004, suggesting that participants engaged in more language-related episodes in the face-to-face test (M = 1.55, SD = 0.77) than they did in the computer-mediated test (M = 0.22, SD = 0.28). In fact, the face-to-face test contained seven times as many LREs as the computer-mediated test.

4.5.3 USE OF COMMUNICATION STRATEGIES

In order to compare use of communication strategies where students produced differing amounts of language, the total number of communication strategies used by a student was divided by the number of words that the student produced. Separate measures were obtained for the computer-mediated and face-to-face tests. Use of communication strategies was compared using a Two-Way Between-Subjects ANOVA with two levels of test method (computer-mediated and face-to-face) and two levels of test prompt (prompt 1 and prompt 2) as independent variables (see Tables 4.21 and 4.22). The main effect of test method, *F* (1, 44) = 21.62, *p* < .001, was statistically significant at an alpha level of .005. Students used over four times as many communication strategies in the face-to-face test (M = 0.46, SD = 0.32) as they did in the computer-mediated test (M = 0.10, SD = 0.19). In fact, fifteen of the twenty-four participants in this study showed no evidence of using communication strategies during the computer-mediated test. Neither test prompt, *F* (1, 44) = 1.47, *p* = .23, nor the interaction effect, *F* (1, 44) = 0.00, *p* = .97, yielded significant *F* ratios at the .005 level.

Table 4.21

Analysis of Variance for Communication Strategy Use

Source	*df*	SS	MS	*F*
Test Method (M)	1	1.55	1.55	21.62*
Prompt (P)	1	0.11	0.11	1.47
M x P	1	0.00	0.00	0.00
Error	44	3.16	0.07	

Note: *p < .005.

Table 4.22

Descriptive Statistics for Communication Strategy Use

	Mean	SD	Min	Max
CMC	0.10	0.19	0.00	0.68
FTF	0.46	0.32	0.04	1.34

Table 4.23 presents a breakdown of the use of communication strategies in both tests. An important difference between the two tests is that the computer-mediated test contained no examples of *reduction* strategies or of the achievement strategy of *restructuring*. This does not mean that neither strategy occurred during the computer-mediated test. It is completely feasible that students started to type a message and either abandoned or restructured it once they realized that they lacked the linguistic resources to express their intended meaning. However, the nature of computer-mediated communication means that the other students see only the product of the students'

Table 4.23

Occurrence of Communication Strategies

	Number (%)	
	Computer-Mediated	Face-to-Face
Reduction Strategies		
Topic Avoidance	0	8 (3.3%)
Message Abandonment	0	2 (0.8%)
Achievement Strategies		
Lexical Substitution	2 (12.5%)	4 (1.6%)
Word Coinage	0	1 (0.4%)
Creativity	1 (6.3%)	0
Restructuring	0	59 (24.0%)
Borrowing	10 (62.5%)	107 (43.50%)
Foreignizing	0	2 (0.8%)
Literal Translation	0	2 (0.8%)
Cooperative Strategies	1 (6.3%)	25 (10.2%)
Other Strategies		
Repair	2 (12.5%)	9 (3.7%)
Telegraphic Strategies	0	2 (0.8%)
Fillers	0	25 (10.2%)

abandonment or restructuring rather than the process. While software exists which can capture each keystroke and revision, the software used in this study, IRC Français, does not have this feature and, thus, certain types of communication strategy are hidden from the rater and this researcher, both of whom lack access to the process of language production in which these strategies would be revealed. The ability to revise messages invisibly would also explain why the computer-mediated data contains no examples of *fillers*, which occur when students talk to themselves in their first language as they try to form their message in the second language. Since students only send fully-formed messages to the chat room, the computer-mediated data does not contain any examples of *fillers*.

The final difference between the two tests arises in the use of *cooperative strategies,* in which students directly or indirectly ask for help from their interlocutors. On the computer-mediated test, this occurred only once, while there were 25 uses of this strategy on the face-to-face test. Again, it would appear that the reduced pressure on the computer-mediated test to produce language immediately may allow students to find ways to express their message without relying on the aid of others. Alternatively, in a computer-mediated test, students may be more ready to abandon their message in the face of difficulties because they have not yet claimed the floor or shared any part of their message.

4.6 QUESTIONNAIRE ANALYSIS

The questionnaire elicited students' perceptions of the group oral and computer-mediated tests (see appendix 2). Student responses on each of the 12 Likert-scale items were compared across the two tests using a Multivariate Analysis of Variance (MANOVA), a procedure which allows multiple dependent variables to be analyzed using common independent variables. In this analysis, each Likert-scale item was a dependent variable and test method (computer-mediated vs. face-to-face) and test prompt were independent variables. The results of the MANOVA for all twelve items, presented in Table 4.24 show that, except for items 2, 4, and 7, there was no significant differences in responses to the two tests. These results are discussed in sections 4.6.1 to 4.6.4.

4.6.1 STUDENT ANXIETY

Question 2 (" I felt nervous before the _____ test") and Question 7 (" I felt nervous during the _____ test") asked students to report their levels of anxiety before and during the test. For question 2, the results of the MANOVA (Table 4.24) showed a statistically significant ($\alpha = .005$) main effect for test method, $F\,(1,44) = 8.86$, $p < .001$. Question 7 also showed a statistically significant ($\alpha = .005$) main effect for test method, $F\,(1,44) = 18.0$, $p < .001$. Neither question produced a statistically significant result for the main effect of test prompt or for the interaction between test prompt and test method.

Table 4.24

Multivariate Analysis of Variance of Likert-Scale Responses

	F Ratios		
	Test Method	Prompt (P)	Interaction
Item	(M)		(M x P)
1. Test gives examiner accurate idea of French ability	0.23	0.03	0.03
2. Felt nervous before test.	**8.86***	0.42	1.36
3. Discussion topic was interesting.	1.28	0.72	1.36
4. Time was too short.	**14.61***	0.98	0.02
5. Test related to class work.	0.45	0.45	0.05
6. Could demonstrate French ability.	0.57	0.57	0.57
7. Felt nervous during test.	**18.0***	0.32	2.88
8. Liked taking test.	0.30	0.07	0.07
9. Perform better on another occasion.	1.09	1.09	0.20
10. Test was too difficult.	1.93	0.53	0.04
11. Did well on test.	0.03	0.83	3.05
12. Perform better with different group members.	0.50	1.04	0.18

Note: * $p < .005$.

Table 4.25 presents a breakdown of how students responded to these questions. Question 2 asked students to agree or disagree with the statement "I felt nervous before the _____ test." It is clear from their responses that the face-to-face test provoked more pre-test anxiety than did the computer-mediated test. Two-thirds of the students (66.8%) agreed that they were nervous before the face-to-face test, but only 29.2 % of students claimed they were anxious before the computer-mediated test. In fact, well over half of the students (58.3%) disagreed with the statement in reference to the computer-mediated test.

Table 4.25

Responses to Questions 2 and 7

Question	Response	FTF (%)	CMC (%)
2. I felt nervous before the test.	Strongly Agree	5 (20.8%)	1 (4.2%)
	Agree	11 (45.8%)	6 (25.0%)
	No Opinion	3 (12.5%)	3 (12.5%)
	Disagree	4 (16.7%)	12 (50.0%)
	Strongly Disagree	1 (4.2%)	2 (8.3%)
7. I felt nervous during the test.	Strongly Agree	5 (20.8%)	0
	Agree	14 (58.3%)	7 (29.2%)
	No Opinion	2 (8.3%)	3 (12.5%)
	Disagree	2 (8.3%)	11 (45.8%)
	Strongly Disagree	1 (4.2%)	3 (12.5%)

The results from Question 7 suggest that the level of anxiety exhibited prior to the face-to-face test was maintained once the test began because 79.1% of respondents agreed that they felt nervous during the test. In fact, since more students (19) report being nervous during the test than before the test (16), there appears to have been an overall increase in anxiety once the face-to-face test began. This is supported by examining the individual responses to these questions, which show that three students who appeared not to be nervous before the face-to-face test (i.e., they disagreed with the statement in Question 2) did feel nervous during the test (i.e., they agreed with the statement in Question 7). In contrast, the same number of students (7) report being nervous before the computer-mediated test as were nervous during it.

4.6.2 TEST LENGTH

Question 4 asked students whether the duration of the test was adequate. The results of the MANOVA (Table 4.24) show a statistically significant ($\alpha = .005$) main effect for test method, $F(1, 44) = 14.61$, $p < .001$; however, no statistically significant differences were found for the main effect of test prompt or for the interaction between prompt and method.

Table 4.26 shows the breakdown of students' responses to Question 4. Ten of the twenty-four students (41.7%) felt that the computer-mediated test had been too short compared with just a single student who wanted more time on the face-to-face test. This desire for more time on the computer-mediated test may be the result of the test being shorter than the time allotted to computer-mediated discussions in class, which typically lasted about twenty minutes.

Table 4.26

Responses to Question 4

Question	Response	FTF (%)	CMC (%)
4. The time was too short.	Strongly Agree	1 (4.2%)	3 (12.5%)
	Agree	0	7 (29.2%)
	No Opinion	0	3 (12.5%)
	Disagree	18 (75.0%)	10 (41.7%)
	Strongly Disagree	5 (20.8%)	1 (4.2%)

Another explanation may be that the students felt they were not able to adequately demonstrate their ability in the time allotted. However, if this were the case, we would expect these ten students to have responded negatively to Question 1 ("I believe that the IRC Français test provides an examiner with an accurate idea of my ability in French") and Question 6 ("I believe that the IRC Français test provided me with an adequate opportunity to demonstrate my ability in French"). In fact, of the ten students who wanted more time on the computer-mediated test, only four of the students felt they had not had an adequate opportunity to demonstrate their ability in French, and of those four, only three felt that an examiner could not get an accurate idea of their ability. In short, though the data shows a clear preference by a large number of students for a longer computer-mediated test, it does not suggest a single underlying reason for this preference.

4.6.3 STUDENTS' ENJOYMENT OF THE TESTS

Question 3 ("I thought the discussion topic for the ____ test was interesting"), Question 8 ("I liked doing the ____ test), and Question 10 ("I thought the _____ test was difficult") elicited the extent to which students had enjoyed the test. The results of the MANOVA (Table 4.24) showed no statistically significant differences ($\alpha = .005$) on any of these items for the main effects of test method and test prompt or for the interaction effect between prompt and method.

Looking closer at the results presented in Table 4.27, we find that students as a group found both discussion topics to be interesting. While 17 participants (70.8%) agreed or strongly agreed that topic 1 (*The typical family has two parents*) was interesting, slightly fewer students—14, or 58.3%—felt similarly about topic 2 (*You should only marry once in your life*). Answers to Question 8 ("I liked doing the ____ test) showed a mixed reaction to the tests. For both of the tests, fourteen participants (58.3%) indicated that they had enjoyed the tests, but five (20.8%) claimed they had no opinion about whether the computer-mediated test had been enjoyable while six participants (25%) made the same claim about the face-to-face test. Finally, the vast majority of students found neither test to be difficult. Nineteen participants (81.2%) disagreed or strongly disagreed with the statement "I thought the group oral test was too difficult". A slightly higher number of participants (21, or 87.5%) responded in this way for the computer-mediated test.

Table 4.27

Responses to Questions 3, 8, and 10

Question	Response	FTF (%)	CMC (%)
3. The discussion topic was interesting.	Strongly Agree	2 (8.3%)	2 (8.3%)
	Agree	15 (62.5%)	12 (50.0%)
	No Opinion	4 (16.7%)	1 (4.2%)
	Disagree	2 (8.3%)	9 (37.5%)
	Strongly Disagree	1 (4.2%)	0
8. I liked doing the test.	Strongly Agree	2 (8.3%)	3 (12.5%)
	Agree	11 (45.8%)	11 (45.8%)
	No Opinion	5 (20.8%)	6 (25.0%)
	Disagree	5 (20.8%)	3 (12.5%)
	Strongly Disagree	1 (4.2%)	1 (4.2%)
10. The test was too difficult.	Strongly Agree	1 (4.2%)	0
	Agree	0	1 (4.2%)
	No Opinion	4 (16.7%)	2 (8.4%)
	Disagree	18 (75.0%)	16 (66.7%)
	Strongly Disagree	1 (4.2%)	5 (20.8%)

4.6.4 STUDENT PERCEPTIONS OF VALIDITY

Questions 1, 5, and 6 asked students' perceptions of the validity of each test in terms of how well the test related to what they had learned in class (#5) and how well the test provided (a) the student with an opportunity to demonstrate their French ability (#6) and (b) the examiner with an accurate idea of that ability (#1). Again, the results of the MANOVA (Table 4.24) showed no statistically significant differences ($\alpha = .005$) on these items for the main effects of test method and test prompt or for the interaction effect between prompt and method.

Table 4.28 presents a breakdown of the results for these items. It shows that 70.9% of the participants felt that they had been able to adequately demonstrate their French ability on the computer-mediated test; in addition, three-quarters of the participants (75%) felt that the computer-mediated test had provided the examiner with an accurate idea of their ability. Responses to these two questions for the face-to-face test were slightly higher with 83.3% (20) of participants indicating both that this test allowed them to demonstrate their abilities and that the test had provided an examiner with an accurate idea of those abilities. Finally, a very high number of participants—87.5% (21 participants) for the computer-mediated test and 95.8% (23 participants) for the face-to-face test—indicated that the test had related to their class work.

Table 4.28

Responses to Questions 1, 5, and 6

Question	Response	FTF (%)	CMC (%)
1. The test gave the examiner an accurate idea of my ability.	Strongly Agree	2 (8.3%)	3 (12.5%)
	Agree	18 (75.0%)	15 (62.5%)
	No Opinion	1 (4.2%)	1 (4.2%)
	Disagree	3 (12.5%)	5 (20.8%)
	Strongly Disagree	0	0
5. The test was related to what I learned in class.	Strongly Agree	5 (20.8%)	4 (16.7%)
	Agree	18 (75.0%)	17 (70.8%)
	No Opinion	0	2 (8.3%)
	Disagree	1 (4.2%)	1 (4.2%)
	Strongly Disagree	0	0
6. The test gave me an adequate opportunity to demonstrate French ability.	Strongly Agree	2 (8.3%)	1 (4.2%)
	Agree	18 (75.0%)	16 (66.7%)
	No Opinion	0	1 (4.2%)
	Disagree	3 (12.5%)	6 (25.0%)
	Strongly Disagree	1 (4.2%)	0

Table 4.29

Responses to Questions 9, 11, and 12

Question	Response	FTF (%)	CMC (%)
9. If I had done the test on another day, I would have done better.	Strongly Agree	1 (4.2%)	0
	Agree	4 (16.7%)	5 (20.8%)
	No Opinion	9 (37.5%)	3 (12.5%)
	Disagree	8 (33.3%)	14 (58.3%)
	Strongly Disagree	2 (8.3%)	2 (8.3%)
11. I believe I did well on the test.	Strongly Agree	3 (12.5%)	1 (4.2%)
	Agree	13 (54.2%)	13 (54.2%)
	No Opinion	3 912.5%)	7 (29.2%)
	Disagree	4 (16.7%)	3 (12.5%)
	Strongly Disagree	1 (4.2%)	0
12. I would have performed better on the test with different students in my group.	Strongly Agree	2 (8.3%)	0
	Agree	1 (4.2%)	1 (4.2%)
	No Opinion	3 (12.5%)	4 (16.7%)
	Disagree	11 (45.8%)	12 (50.0%)
	Strongly Disagree	7 (29.2%)	7 (29.2%)

4.6.5 STUDENT PERFORMANCE

A number of questions asked students to describe the quality of their own performance on the test (#11) and to explore factors which may have changed that performance such as the test being administered on a different day (#9) or with a different group composition (#12). No statistically significant differences ($\alpha = .005$) were found on any of these items for the main effects of test method and test prompt or for the interaction effect between prompt and method (Table 4.24).

Table 4.29 presents a breakdown of the students' responses. For both tests, only five students (20.8%) indicated that they would have performed better if they had taken the test on another day. Even fewer students felt that a different group composition would have allowed them to perform better (Question 12). Only one student felt this way about the computer-mediated test while three students expressed this opinion about the face-to-face test. Finally, for both tests a majority of students—16 for the face-to-face test and 14 for the computer-mediated test—felt that they had done well.

4.6.6 TEST PREFERENCE

After the second test, all students responded to an additional question which asked them to indicate their preferred test method (computer-mediated, face-to-face, or no preference). Responses to this question are presented in Table 4.30, which shows a clear preference by two-thirds of the students for the face-to-face test.

The students who preferred the face-to-face test gave several reasons for doing so, which could be grouped into four themes. The first theme included comments from four students who compared typing and speaking. A typical comment is that of S24, who said

"I'm not able to type as fast as some other students and I'm not sure of spellings." The other three students also focused on the ability to produce language faster on the face-to-face test, the need to concern oneself with spelling on the computer-mediated test, and their own poor typing ability. The last point may be related to the second theme found in the students' responses—that they can demonstrate their abilities better through speaking than through writing. Three other students made comments similar to that of S8, who said "I just prefer to speak. I feel I can produce French better."

The third theme pertains to comments about the different nature of interaction in face-to-face versus computer-mediated communication. Two students commented about the role of paralinguistic and non-verbal factors in aiding communication: Thus, S13 said "the tone of voice helps to understand better" while S14 stated "It's easier for me to react when I'm looking at the person I'm talking to." Another student (S19) was unhappy at the lack of direction that computer-mediated communication can engender: "Sometimes talking on the computer doesn't make for well directed conversations. There are too many things going on at once to focus."

Table 4.30

Students' Preferred Test Method

Method	No. Responses (%)
Computer-Mediated	7 (29.2%)
Face-to-Face	16 (66.7%)
Either	1 (4.2%)

The fourth theme is perhaps the most interesting because it focuses on the potential for learning from face-to-face interaction. One student suggested that she had enjoyed the challenge of speaking because it was something that she was not used to andtherefore she had to make a greater effort: “The face-to-face requires a lot more thought. When you write in French, you are just doing what you have done since the beginning of French, but there is not a lot of speaking practice in class.” This comment seems to imply that the act of speaking in a testing situation may in fact be a learning experience. Such an idea is echoed in another student’s comment concerning the potential for improved thinking and expression to be found in face-to-face interaction: “Parce que ça nous aide à improver notre pensée et tout ce qu'on a à dire” (*It helps us to improve our thoughts and everything we have to say*).

Two main themes emerged from the responses given by those who expressed a preference for the computer-mediated test. These are exemplified by S20’s comment: “I have more time to think about what I'm going to say. I also can look at what I'm writing to make sure it's correct.” Four of the seven students who preferred the computer-mediated test gave reduced time pressure as a reason for their preference, with two of them adding that they appreciated being able to correct errors. One student specifically stated that this reduction in pressure reduced anxiety: “I personally have a tendency to get nervous when speaking face-to-face; on the computer I had more time to formulate my thoughts and didn't feel as rushed or pressured.” Another student commented that expression of ideas was easier when you could see them being formulated on the screen. Finally, two students commented that they preferred the computer-mediated test because they had more experience in the class of interacting through CMC than face-to-face. For one of these students, the fact that the test also went faster was a benefit.

4.7 CONCLUSION

This chapter has described the analyses that were performed on the data and has presented the results of those analyses. In the following chapter, I discuss the implications of these results for the research questions described in chapter 1 and draw conclusions about the potential uses of computer-mediated communicative tests.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

5.1 INTRODUCTION

Chapter 1 presented a methodology for test validation which aimed to construct an argument around evidence concerning the weakest assumptions underlying any inferences made from a test score. Inferences made about spoken language ability on the basis of a written computer-mediated test were seen to rest, in particular, on five assumptions, which then determined the research questions examined in this study. In this chapter, each of those assumptions is discussed in light of the results presented in chapter 4 in order to examine the validity of using computer-mediated communicative tests as measures of spoken proficiency. The chapter ends with recommendations for the future use of computer-mediated communicative tests, discussion of the limitations of this study, and suggestions for future research.

5.2 ASSUMPTION OF PRONUNCIATION IRRELEVANCE

The first assumption was that pronunciation does not need not be measured in order to gain an accurate measure of speaking ability. That is, while pronunciation may be psychologically important because successful communication may be inhibited by unclear pronunciation, it may be psychometrically unimportant if scores on a scale which measures pronunciation can be shown to correlate highly with other sub-scales used to evaluate an individual's language production.

The results presented in section 4.2 showed low correlation rates between a measure of pronunciation—the *intelligibility* sub-scale on the face-to-face test—and either the total score or individual sub-scales scores achieved on the computer-mediated test. For rater 1, the highest correlation—between *intelligibility* and *overall effectiveness*—was moderate ($r = 0.51$) and was certainly not high enough to justify the systematic non-measurement of pronunciation. The possibility of excluding measurements of pronunciation deteriorates further when we look at the results for rater 2, where the highest correlation between *intelligibility* and a score on the computer-mediated test is even lower ($r = 0.20$ for both *range* and *flexibility*). For this rater, there was also a zero correlation between *intelligibility* and *effectiveness.*

The fact that pronunciation correlates so little with any of the measures obtained from the computer-mediated test does not, in itself, necessarily mean that computer-mediated tests could not provide valid information about a language learner's oral ability. Should the other assumptions be supported by evidence from this study, it would be relatively easy to build a battery of computer-mediated tests which incorporated both a computer-mediated communicative test and a computer-recorded sample of spoken language. Such samples may be recorded at the beginning or end of the same testing session in which students participate in a computer-mediated communicative test. The oral language samples elicited in this way can be rated for intelligibility, thereby providing a score for pronunciation which can be combined with the scores on other criteria as determined from performance on the computer-mediated communicative test. Thus, the latter test's lack of information concerning pronunciation may be overcome, though probably at the cost of lower test efficiency since the inclusion of an additional task into the testing situation will, of course, have a negative impact on the efficiency of test administration and scoring. Whether this impact is excessive in terms of additional resources when compared to the benefits of making this change is an empirical matter which the present study cannot address but which could be investigated and evaluated in further validation studies.

One final result that should be discussed is the zero correlation between *intelligibility* and *effectiveness* for rater 2. An examination of the distribution of scores by this rater suggests that she was less stringent in her interpretation and application of the *intelligibility* sub-scale than was the other rater. The average score awarded by rater 2 for *intelligibility* was 4.63 (out of 5) with two-thirds of the students receiving full points (i.e., 5/5). In comparison, rater 1's average score for *intelligibility* was 4.0, and she awarded full points to only one in five students. The majority of students (15 out of 24) received a rating of 4 from rater 1. It is important to note that while the descriptors for band four and band five of the *intelligibility* subscale both assume that any phonological errors present do not render the speech incomprehensible, they differ in terms of frequency of phonological errors. Band five is characterized by the presence of "few phonological errors" while band four contains "many phonological errors." Such wording immediately presents problems for raters for whom the boundary between "few" and "many" may be imprecise and ill-defined even after training. Added to the confusion from the rubric's wording is the possible role of physiological and psychological differences between the raters, which may result in their noticing different numbers of phonological errors. Thus, it is not clear at this time why the rater awarded such different scores on the *intelligibility* sub-scale.

In summary, the first assumption—that measurement of pronunciation was not necessary to achieve an accurate measure of speaking ability—was not supported by the evidence in this study. However, we have seen that if the other assumptions hold, technological developments may offer potential methods for evaluating pronunciation that could complement the results of a computer-mediated communicative test.

5.3 TEST SCORE EQUIVALENCE ASSUMPTION

According to assumption 2, if the computer-mediated test were tapping a similar construct, the scores achieved on the face-to-face and computer-mediated tests would be similar. The evidence from this study supports this assumption. When compared using Analysis of Variance, no statistically significant differences were found between the computer-mediated and the face-to-face test on total scores derived from the five sub-scales in common for either rater 1 (CMC mean = 18.21; FTF mean = 16.92) or rater 2 (CMC mean = 17.79; FTF mean = 18.17). Thus, the test score evidence suggests that the two tests appear to be measuring the same construct.

While the evidence for similarity of test scores is certainly persuasive, the low inter-rater reliability of scores on the computer-mediated test is a concern. If a computer-mediated test is to be used as a substitute for an oral test, one would hope that it were as reliable as the test which it replaces. This was not the case here. The inter-rater reliability of the two raters for total score on the computer-mediated test was quite low ($r = 0.57$) and certainly much lower than the correlation coefficient achieved on the face-to-face test ($r = 0.68$). Several possible factors may explain the lower reliability coefficient achieved on the computer-mediated test.

One possible reason for this lack of reliability may be the amount of language that students produced on the computer-mediated test. The raters based their evaluations for the computer-mediated test on a much smaller sample of French (M = 91.50, range = 164) than was available for evaluating performance on the face-to-face test (M = 249.38, range = 874). As was noted in chapter 3, three of the students failed to produce a sufficiently large language sample to conduct a type/token analysis. Even excluding these students, the sample used to calculate the type/token ratio had to be smaller—at fifty words—than the 100-word sample typically used for this analysis because of the low levels of production across all students. In fact, only nine of the twenty-four students produced 100 words or more on the computer-mediated test. In contrast, only two students failed to produce 100 words on the face-to-face test. Thus, it may be the case that the language samples provided by the computer-mediated test were inadequate for raters to make a reliable estimate of learners' abilities.

Another possible source of the low inter-rater reliability is the grading rubric. As reported in chapter 3, the rubric used in this study was based on Venugopal's (1992) rubric that had previously been developed for group oral exams. While great care was taken in writing the descriptors for each criteria so that they were sufficiently generic to apply equally to both oral and computer-mediated discourse (with the obvious exception of the criteria of *intelligibility*), it is possible that the resultant rubric was more suited to rating oral rather than written language samples. In other words, the problem may have arisen because the rubric cannot be interpreted as easily and as consistently when used to evaluate non-spoken language, even if the latter is interactive in nature. Lack of a clear, common interpretation among the raters may be causing differences in their scores.

There is, however, another possible cause of different rater interpretations of a rubric. One way in which reliability can potentially be increased is through extensive rater training. Although this study piloted the rubric prior to the rating sessions and started each rating session by norming the raters, it may be that this training was inadequate to ensure either that the raters had similar interpretations of the rubric as it applied to the computer-mediated test data or that they employed similar strategies when rating. During the pilot session, raters' initial inclination was to analyze the language samples in detail by reading the transcripts multiple times. This researcher requested that the raters adopt a more global, impressionistic approach and limit themselves to just two readings of the transcript. However, it is not certain that the raters received adequate training and practice to ensure that they would fulfill this request in the same way. As a result, the two raters may have read the CMC transcripts in different ways and with differing levels of analysis. Doing so would surely have reduced their level of agreement.

An additional cause of the lower inter-rater reliability on the computer-mediated test may be the raters' different levels of experience with computer-mediated communication. Both raters had a lot of experience teaching French 201 students and administering oral exams. Only one of the raters, however, had previously used computer-mediated communication in the classroom and, thus, had extensive experience with reading CMC transcripts. The other rater's lack of experience with the non-linear nature of interaction in computer-mediated communication may have caused her greater difficulties in reading the transcripts, which could translate into less consistency in the scores she assigned. In the same way that the rubric interpretation problem discussed in the previous paragraph could result from the training process, this problem of inexperienced CMC transcript readers may be attributable to inadequate or insufficient training.

Finally, it should be noted that the low inter-rater reliabilities achieved for both tests raise important ethical issues. It is clear that the highest reliability coefficient achieved here ($r = 0.68$) would be unsatisfactory for a high-stakes test, but low inter-rater reliabilities are no less problematic for low-stakes tests such as the classroom assessment found in the present study. The course grades of participants were determined in part by their performance on both the computer-mediated and face-to-face tests. In the majority of cases (13 for the computer-mediated test and 15 for the face-to-face test), the raters either agreed completely or differed by two points or less. While it may be suggested that the tests in this study represented a relatively small part of the students' course grade (10%), and that, therefore, such differences are not of concern, for one student, the raters disagreed in their evaluation by ten points on the computer-mediated test and by nine points on the face-to-face test, with the higher scores being assigned by rater 2 on both tests. If the student's grade depended on the evaluation of rater 1, the lower scores assigned by that rater would be enough to lower the student's grade if she or he were on the border between two grades. Such a low evaluation may also lower the student's levels of confidence about his or her second language ability with potential negative implications for motivation to study French, especially if rater 2's evaluation was closer to the student's actual ability level. Conversely, if the student's ability was actually closer to rater 1's evaluation, receiving rater 2's evaluation may raise confidence and increase motivation to learn languages. If teachers wanted to use these test scores diagnostically, the impression of this student's strengths and weaknesses they would receive would depend on the source of the scores, as would the accuracy of the diagnosis made about the student.

The fact that the scores used in this study were not those that contributed to students' actual final grade—which was determined by the course instructor using a different rubric—does not negate this problem. Language testers have a responsibility to ensure that all evaluations of performance, whether in high-stakes situations or not, are reliable because of their potential effect on stakeholders such as students, teachers, administrators, etc. If the reliability of scores cannot be assured, the use of those scores is ethically unjustified. This is true of the face-to-face test, but it is especially true of the computer-mediated test. The discussion in this section has suggested several factors which may have contributed to the low levels of inter-rater reliability for the computer-mediated test. Until research shows that computer-mediated tests can achieve acceptable levels of reliability, we should use them with caution.

To summarize, the assumption of similarity of test scores was supported by the test data; however, the low inter-rater reliability on both tests, and especially on the computer-mediated test, is a concern which should be addressed in future research.

5.4 SIMILAR LANGUAGE ASSUMPTION

The third assumption underlying the validity of using performance on a computer-mediated communicative test to make inferences about oral language ability is that the language produced on a computer-mediated test is similar to that produced when students are tested orally.

Although the language produced on the face-to-face and computer-mediated tests was found to be similar with respect to structural complexity—as expressed in the length of multi-clause utterances and the relative frequencies of multi- and single-clause utterances—overall the data did not support this assumption. Compared to the face-to-face test, students' language production on the computer-mediated test tended to be more lexically complex (i.e., it had higher type-token ratios) and literate (i.e., it had higher lexical densities), to involve the use of a wider range of functions, and to be more error-prone.

5.4.1 DISCUSSION OF QUANTITY OF LANGUAGE PRODUCED

Given the short duration of the test and the fact that few people can type as fast as they can speak in their first language let alone in a second language, it is not surprising that students produced much less language on the computer-mediated than on the face-to-face test. However, this finding does not replicate Kern's (1995) results which showed greater levels of production using the CMC program, *Interchange,* compared to oral discussions. Why should the present study find such different results? Three reasons suggest themselves.

First, in Kern's study, computer-mediated discussion in one class period was always followed by an oral discussion the following period. Moreover, the topic was identical for both discussions, which leads Kern to suggest that "it may be that some students felt 'talked out' by the time they began the oral discussion" (pp. 463-64). Such feelings may have contributed to the lower levels of oral production in Kern's study. In contrast, although the topics used in the present study were similar, they were not identical, so students may not have felt 'talked out' in the same way. Moreover, since this study's design was counter-balanced for test method and test prompt, if students did react in this way and produce less language in the second testing session, it would not affect the study's overall results.

The second possible explanation arises from the different participation levels of the teacher in this study compared to those in Kern's study. Of the two teachers who participated in Kern's study, one did not join the computer-mediated discussion at all while the other only contributed 4% of the total number of computer-mediated turns. Such an absence of participation in the computer-mediated interaction mirrors that of this study where the instructor did not participate at all in the IRC Français test session. Where the two studies differ, however, is instructor participation in the oral discussions. Kern states that both instructors in his study took 45% of the total number of turns in the oral discussions held in their classes. In contrast, the instructor in this study participated much less in the oral discussions. In every group, the instructor took at least two turns that were related to test administration—giving the prompt to be discussed and informing students that the testing session had ended. For half of the groups, the instructor took only one or two turns beyond those necessary for test administration. The greatest number of turns taken by the instructor was in group 1B, where she took 24% of the turns. For all the other test groups, she took no more than 15% of the turns. Thus, the instructor's lower participation level in the face-to-face interaction in this study may have allowed participants to produce more language.

The third explanation concerns a possible effect due to the time allotted for the computer-mediated test, which almost certainly restricted the amount of language that students were able to produce. If more time had been available, production on the computer-mediated test would have been much higher. In fact, the results from the survey that are presented in section 4.6.5 show that 10 of the students wanted the computer-mediated test to last longer. In contrast, only one student wanted the face-to-face to last longer. Every group in the face-to-face test relied on the teacher to revive the conversation at least once, and for two of the groups, the instructor had to restart the discussion on four different occasions. In his study, Kern also found that instructors asked a lot of questions during the oral discussions, which, like the computer-mediated discussions, lasted for a whole 50-minute class period. Kern's findings in combination with those of this study suggest that oral discussions may run out of steam faster in terms of time than do computer-mediated discussions. Whether students have produced more language once the oral discussion has run its course is another matter which neither study can answer. It may, however, be possible to address this issue by completely eliminating any teacher or examiner involvement and allowing students to continue producing language until they felt they had said all they possibly could about a topic. In this way claims about relative levels of language production could be made that were independent of confounding variables such as examiner interaction, typing speed, or time constraints.

5.4.2 DISCUSSION OF TYPE/TOKEN RATIOS

In chapter 4, it was noted that one student had a much lower type/token ratio (TTR) on the group oral exam than on the computer-mediated test because she had two false starts in the oral data. Given the decision to use a small sample—the first 50 words produced—is it possible that there is warm-up effect where students type/token ratios are affected by the fact that they have not fully made the transition from interacting in their first language to interacting in French? At first glance, this seems an unlikely occurrence since the TTR for both tests was determined under the same conditions: the first fifty words produced. However, what this argument ignores is the fact that language production on the computer-mediated test occurred in a different way. Students produced language at a much slower rate and had more time to compose and revise their messages. Additionally, the short greetings routines discussed in section 4.4 may have served to warm students up in a way that was not available in the face-to-face test where students did not engage in these routines before discussing the prompt. Thus, although the results reported here match those reported in Warschauer (1996), they should be treated with some caution.

5.4.3 DISCUSSION OF LANGUAGE FUNCTIONS

The wider range of discourse functions found in the computer-mediated discourse echoes the results of Kern's (1995) comparison of computer-mediated and face-to-face interactions (though Kern's study used a different classification system). However, the discussion in section 4.4.4 suggested that the results in this study may have been a little misleading. In particular, the presence in the computer-mediated data of higher frequencies of *interpersonal* and *instrumental* functions seems to be an artifact of test administration. The higher numbers of *interpersonal* functions were caused (a) by the exchange of greetings as students entered the virtual testing space after the test had started and (b) by the lower levels of language production on the computer-mediated test, which tended to inflate the importance of the greetings routines as a percentage of the total number of functions. Since students were already present in the testing space when the face-to-face test formally began, they did not engage in greetings routines. Once such routines were excluded from the data, the computer-mediated discourse contained fewer instances of *interpersonal* functions (7.1% of total functions) than did the face-to-face discourse (14.9%).

The second artifact caused by the testing environment is the presence of nine instances of the *instrumental* function in the computer-mediated discourse compared to just a single instance in the face-to-face discourse. As was discussed in chapter 4, eight of the *instrumental* uses of language occurred as transitions from greeting routines to discussion of the prompt. Such a transition was necessary because the simultaneous testing of all groups required that the participating instructor could not simply announce the prompt as she had done for the face-to-face test where each group was tested individually. Instead, the instructor handed each student a sheet of paper with instructions about the chat room to join and the topic to discuss. Under different circumstances, where the students were not participating in a research study and could, therefore, all be given an identical prompt for the computer-mediated test, such *instrumental* uses of language may not be present.

In spite of the caveats described above, one important difference regarding the functional use of language in the two tests remains robust. The results in chapter 4 show that students used language with the *heuristic* function very infrequently on the computer-mediated test when compared to their functional use of language on the face-to-face test. Only two examples of the *heuristic* function, representing less than one percent of total functions, were found in the computer-mediated data. In contrast, 39 instances of the heuristic function (7.3% of total functions) were found in the face-to-face data. The implications of this finding are identical to those concerning the relative infrequency in the computer-mediated data of *cooperative* strategies, which is discussed in detail in section 5.4.4.

5.4.4 DISCUSSION OF LANGUAGE COMPLEXITY

The measures of linguistic complexity examined in this study—type/token ratio and structural complexity—were also compared for face-to-face versus computer-mediated discourse in Warschauer (1996). As in this study, Warschauer found that computer-mediated discourse tended to be more complex lexically; unlike this study, however, the computer-mediated discourse produced by the students in Warschauer's study was also more complex structurally. These results inspire the question: Why might the students in this study show no statistically significant difference across the test methods in terms of the structural complexity of their language when the students in Warschauer's study showed a difference?

There is no simple answer to this question. A partial explanation may be found in the fact that the students in Warschauer's study were drawn from a different proficiency level. Since they were advanced writers, their written language may generally have a contained a high level of subordination which may have been transferred, at least partially, to the discourse produced during their chat session. Their ability to produce more complex sentences would almost certainly be reinforced by the more relaxed pace at which chat room interactions occur, which would give the students in Warschauer's study enough time to formulate more complex sentences. In contrast, the students in the present study were low-intermediate students whose level of language acquisition may have prevented them from producing more complex language in the computer-mediated test. They simply lacked the tools to do so.

Of course, the argument in the preceding paragraph does not fully explain why the students in this study were able to produce equally complex utterances in both tests in spite of having less time for reflection and planning on the face-to-face test. The context in which data was collected may be the cause of this apparent discrepancy. Warschauer's study examined the classroom use of CMC, in which students might be less pressured to perform to the best of their ability than would be the case in the testing context which was the focus of this study. In Warschauer's study, the extra time for reflection permitted in chat rooms may have allowed students to produce more complex sentences than they did in the face-to-face discussion, where the focus may have been on communication without concern for demonstrating the full extent of their linguistic knowledge. In this study, however, students' awareness that their language would be evaluated may have led them to try to demonstrate the full range of their linguistic knowledge in both the computer-mediated and the face-to-face tests i.e., the students tried to use structures in the face-to-face test that they might have used less frequently in a classroom oral discussion. This possibility cannot be addressed by the present study but could easily be investigated in the future.

Finally, there is one other possible explanation for the similar levels of sentence complexity on both tests. Previous research by Beauvois (1997; 1998) and by Payne and Whitney (2002) offers tentative evidence that computer-mediated interaction may have a positive influence on the development of oral language proficiency. Since the students in the present study had many opportunities to engage in computer-mediated discussion over the course of the semester, they may have become more proficient orally as a result of these interactions. On the face-to-face test, this increased proficiency may have resulted in a level of structural complexity similar to that found on the computer-mediated test. Thus, this study's finding of equal structural complexity in the computer-mediated and face-to-face discourse may result from a complex interaction between the context of data collection, the influence of that context on student motivation, and the possible existence of transfer effects from computer-mediated interaction to oral proficiency. Since the data in this study neither supports nor refutes this hypothesis, further research may be justified.

5.4.5 DISCUSSION OF ERRORS

The greater number of errors in the computer-mediated language is, at first glance, rather surprising. The Monitor Hypothesis (Krashen & Terrell, 1983) posits that second language production is initiated by acquired language and that consciously learned language only plays a role in editing, or monitoring, language that has already been generated. Krashen and Terrell argue that three requirements must be met to permit successful use of the monitor. They suggest that the first requirement—that students have enough time to think about the rules—rarely occurs in conversation. However, since students type their messages in computer-mediated communication at a slower rate than they would speak them, it can be argued that they do have more time in which to monitor their language production, especially since students are also able to edit and revise their contributions before posting them to their group (indeed, the ability to do this was cited by two of the students who expressed an overall preference on the post-test questionnaire for the computer-mediated test over the face-to-face test). Thus, the requirement of sufficient time to think about the rules is, in theory, more likely be met for computer-mediated communication than for face-to-face interaction.

The second requirement—that students be focused on form—cannot be demonstrated from the data. However, it is intuitive that students will focus on form to some degree when they are in a testing situation and they are aware that part of their score will be determined by their grammatical accuracy as was the case in the present study. Of course, students may also have focused on form in the face-to-face test but may have been less successful due to the time constraints of an oral interaction which moves faster and is less amenable to delays in communication.

The third requirement is that students know the rule. Logically, this cannot account for differential frequencies of errors across the two tests. If students do not know a particular rule, then they should make the error resulting from violations of that rule at similar rates in both tests, regardless of whether they are monitoring language production more on the computer-mediated test. If students do know the rule, however, one might expect the computer-mediated language to be less error-prone than the face-to-face language because the former is produced in a less pressured environment in which students may have greater time for use of the monitor and in which the written nature of interactions may make errors more salient. Therefore, the fact that the computer-mediated language contains more errors than the face-to-face language needs to be explained.

The breakdown of frequencies of error types in Table 4.17 shows that while all error types were more prevalent in the computer-mediated test, errors of lexical choice, number, word order, and adjective agreement showed the greatest difference in frequencies across the two tests. A possible explanation for this trend is that certain types of errors become more salient to observers in a computer-mediated environment because the interactions are written. For example, in French, adjectives can agree with the noun in terms of gender (masculine = ‘petit’, feminine = ‘petite’) and number (singular = ‘petit’/’petite’, plural = ‘petits/petites’). While both types of agreement are marked for the majority of adjectives in written language, only gender agreement is marked in spoken language because the word final –s that marks plurality is silent (in fact, even gender agreement is not always marked in spoken language as in 'normal'/'normale'). A student may be blissfully unaware of the need for adjective agreement in French, but a rater would only be aware of these missed agreements in oral discourse in the case of feminine nouns. In contrast, the written nature of computer-mediated communication means that raters can notice every adjective agreement error, which may have resulted in the higher numbers of such errors found in this study. A similar situation occurs with errors involving singular and plural nouns (*number* errors in this study) which are also marked in the majority of cases with a word final –s.

One question that arises is whether the results of the error analysis would change if the same criteria were used to identify errors on both test. The computer-mediated data were examined to determine which of the *number* and *adjective agreement* errors would not have been identified as errors in the oral data. Of the 23 *adjective agreement* errors found in the CMC data, 12 (52.2%) involve such incorrect plural or feminine forms while the percentage of *number* errors involving word final -s (70.0%) is even higher. In other words, the perceived numbers of *adjective agreement* and *number* errors increase enormously when the medium for producing and receiving the message is visual rather than auditory. Individual's error ratios on the computer-mediated test—i.e., the frequency that errors occurred per 10 words—were also recalculated using the new criteria and were compared to their error ratios on the group oral exam using a Between-Subjects ANOVA. No statistically significant differences between the two tests were found at the .005 level (F (1, 44) = 7.34, p = .01). In other words, the perceived numbers of *adjective agreement* and *number* errors increase enormously when the medium for producing and receiving the message is visual rather than auditory and result in a statistically significant higher total number of errors on the computer-mediated test.

Increased saliency does not, however, explain why *lexical choice* errors should be more prevalent in the computer-mediated tests. The data on communication strategies suggests a plausible explanation for the greater number of these errors. It was noted in chapter 4 that while there were 25 examples of *cooperative* strategies in the face-to-face data, there was only a single instance in the computer-mediated data. Cooperative strategies occur when a learner's interlanguage is inadequate to express his or her intended meaning and the learner either asks an interlocutor for the L2 usage or indicates that they cannot explain the meaning. What might have happened in the computer-mediated test is that when a message required an unknown lexical item, the learner chose not to ask his or her interlocutors for the necessary L2 item and tried to use available linguistic tools. Such a suggestion is supported by the very low number of *heuristic* functions in the computer-mediated data where learners used language to extend their knowledge of the target language only twice.

The decision not to solicit help on the computer-mediated test—and thus the low use of the *heuristic* function—may arise from the different ways language is produced in oral versus computer-mediated communication. In oral communication, typically only a single individual has the floor at any one time. If students run into difficulties expressing their message, they can ask for assistance, receive it immediately, and complete their message without fear that the conversation has moved on. This is not necessarily true with computer-mediated communication, where several individuals may be formulating and typing their turns simultaneously. An individual who requests unknown vocabulary probably will not receive a reply until others have noted the request, completed (or abandoned) their current message, and typed their response. The response will certainly not be received with the same immediacy as occurs with oral communication. In the meantime, the student who requests aid may not only have to delay his or her message until a point where it may be less timely or even irrelevant, but will also have to start it afresh since the composing area in chat programs does not allow the storage of draft messages while another message—the request for assistance—is typed. Where computer-mediated communication is subject to the time constraints of a testing situation, students may choose not to solicit unknown lexical items and to take their chances with existing linguistic resources. However, these resources may be inadequate with the result that students commit more *lexical choice* errors.

Finally, it is also possible that the tendency discussed in the preceding paragraph may have been reinforced by the different roles that the teacher played in the two tests. For the face-to-face test, the teacher was physically present in the testing space and may, thus, have been seen by students as a potential resource when they experienced difficulties. In fact, it was sometimes unclear during the face-to-face test who was the intended recipient of a request for assistance: the teacher, the other students, or perhaps both. The teacher did provide assistance during the face-to-face tests of four of the eight test groups participating in this study. This aid consisted of a single instance of supplying the French needed to complete a message for one group, two instances for two of the groups, and five instances for the fourth group. In contrast, although the teacher was *physically* present in the computer lab during the computer-mediated test, she was not *virtually* present in the testing space because she did not enter any of the chat rooms where the examinees were holding their discussion. Knowing that they did not have an advanced speaker of French to draw upon may have reinforced students' belief that they should rely on their own linguistic resources for the computer-mediated test. Thus, they used fewer *cooperative* strategies.

5.5 SIMILAR INTERACTION ASSUMPTION

The fourth assumption is that students' interactions are similar across the two tests. This assumption was not supported by any of the variables analyzed in this study. Students produced a greater number of turns in the face-to-face test than in the computer-mediated test, and on average, the oral turns were longer. The face-to-face test also contained more examples of language-related episodes and greater use of communication strategies.

5.5.1 DISCUSSION OF INTERACTION FEATURES

Given the greater amount of language produced in the face-to-face test, it is not surprising that students also produced more turns in that test. What is interesting, however, is the much lower numbers on the computer-mediated test of language-related episodes and of certain types of communication strategies such as *reduction* strategies, *restructuring* and *fillers*. The infrequency of these features probably arises because of the nature of computer-mediated communication.

Language related episodes are "any part of a dialogue where students talk about the language they are producing, question their language use, or other- or self-correct their language production" (Swain, 2001, pp. 286-87). Of the four possible types of actions which would be classified as a language related episode, three potentially involve students focusing on their own language production (talking about their language production, questioning their language production, or self-correcting). These actions may occur during a computer-mediated test, but they would be hidden because the student could do all of these actions prior to hitting 'enter' to submit their text to the discussion. As was discussed in section 4.5.3, a similar case can be made for the infrequency of *reduction* strategies, *fillers*, and *restructuring* in the computer-mediated data. Students may start to type a message, realize they lack the linguistic resources to complete it, and either abandon, reduce, or restructure their message; however, these actions would be invisible to the other students, the instructor, or a rater, all of whom see only the result of revision in the face of difficulties (i.e., the message) rather than the process of revision. The ability to revise messages invisibly would also explain why the computer-mediated data contains no examples of *fillers*, which occur while students talk to themselves in their first language as they try to form their message in the second language. Since students only send fully-formed messages to the chat room, the computer-mediated data would not contain any examples of *fillers*.

This poses a problem for language testers because potentially important aspects of second language performance in face-to-face interactions are invisible to an examiner who evaluates language ability solely from performance on a computer-mediated test. If raters cannot see the process of L2 message revision as well as its result, can they assign an accurate score? And if the scores assigned lack reliability, how can test users make valid inferences from them? The results from this study show that there were no significant differences in the scores assigned to the two tests: Raters arrived at similar assessments of students' second language abilities even when they lacked information about the process of L2 production. However, it is not clear that a similar convergence would be achieved with students at other proficiency levels or with different test tasks. In addition, the low inter-rater reliabilities for the computer-mediated test suggest that raters were, in fact, not able to assign scores with the level of accuracy necessary for valid inferences. Section 5.3.1 suggested several reasons why this may have occurred, but it is possible that a lack of information about the process of message revision and production may have contributed to the raters' difficulties.

5.6 ASSUMPTION OF SIMILAR STUDENT REACTIONS

The final assumption is that learners perceive the computer-mediated and face-to-face tests in similar ways. Questionnaire responses showed no statistically significant differences on nine of the twelve Likert-scale questions. A clear majority of participants found both discussion topics interesting (58.3% for topic 2, 70.8% for topic 1), enjoyed taking the tests (58.3% for both tests), did not think the test were too difficult (81.2% for the FTF test and 87.5% for the CMC test), felt that they had been able to adequately demonstrate their French ability (83.3% for the FTF test and 70.9% for the CMC test) and that an examiner could accurately rate that ability (83.3% for the FTF test and 75.0% for the CMC test), thought that the test had related to course work (95.8% for the FTF test and 87.5% for the CMC test), believed that they had done well on the tests (66.7% for the FTF test and 58.4% for the CMC test), and didn't think that changing the group composition (75.0% for the FTF test and 79.2% for the CMC test) would have improved their performance.

Students did produce statistically different responses on the three remaining items. While almost all participants (95.8%) felt that the face-to-face test was not too short, opinion was divided about the length of the computer-mediated test, with 41.7% agreeing that it was too short and 45.9% disagreeing. More importantly, levels of nervousness prior to and during the two tests were significantly different with students reporting lower levels of anxiety on the computer-mediated test. Two-thirds of students were nervous before the face-to-face test compared to just 29.2% before the computer-mediated test while more participants were nervous during the face-to-face test (79.1%) than the computer-mediated test (29.2%). Finally, two-thirds of students showed an overall preference for the face-to-face test.

5.6.1 DISCUSSION OF SURVEY RESULTS

The lower anxiety levels reported for the computer-mediated test support prior claims that CMC lowers affective barriers (e.g., Sanchez, 1996). The reduced pressure that results from the greater time for reflection and planning in a computer-mediated environment led to lower test anxiety not only during the test, but also prior to the test. This was appreciated particularly by those students who expressed an overall preference for the computer-mediated test. Meunier (1998) found that traits such as personality, motivation, attitude, and gender influence students' participation in computer-mediated communication. While this study did not collect data about such individual traits, it is possible that those learners who expressed a preference for the computer-mediated test share traits such as shyness or a strong predilection for accuracy over fluency. Such a possibility is worthy of further investigation.

However, in spite of this potential advantage for the computer-mediated test, two-thirds of students indicated a preference for the face-to-face test. For these students, the lower levels of anxiety experienced on the computer-mediated test were outweighed by the fact that the face-to-face test allowed them to produce language faster, to demonstrate their abilities better, and to achieve more effective communication through the use of non-verbal and paralinguistic factors. In other words, the majority of students appear to have been less concerned with the stress of being in an assessment situation and more concerned with exploiting that situation by presenting the full range of their language abilities to the teacher.

Finally, it should be noted that test length may have influenced students' overall preferences. Four students commented on the questionnaire regarding their ability to produce language faster in oral tests. When tests are of equal length, as in this study, this may be an important factor in favoring one test over another because students will want to produce as much language as possible. Thus, the slower rates of language production on the computer-mediated test would be seen as a hindrance to successful completion of the test. However, if more time were allocated to the computer-mediated test so that students felt that they had the opportunity to produce equal amounts of language on the two tests, this preference for the face-to-face test may disappear because of the potential benefit from reduced time pressure and opportunities for revision referred to by four of the seven students who preferred the computer-mediated test. Thus, the relationship between test length and student preferences should be investigated further.

5.7 CONCLUSION

This study was motivated by one overarching question: Can you make inferences about spoken language ability on the basis of performance on a computer-mediated test? It was argued in chapter 1 that to do so with validity required that certain assumptions be met. The preceding sections of this chapter have discussed both the evidence and the limitations of that evidence in supporting each of the assumptions which provide the validity of making inferences about oral language ability from performance on computer-mediated communicative tests. We are now in a position where we can begin to synthesize the various parts of this validity argument.

This study found no significant difference in the total scores derived from the sub-scales shared by the two tests. While students' affective responses differed with respect to levels of anxiety, test length, and overall preferred testing method, they were similar with regards to enjoyment of the test, perceptions of test validity, and students' performance. However, it is clear from the linguistic and interactional data that students' performance on the two tests is dissimilar. The different testing procedures of the computer-mediated test produced functional uses of language not seen in the face-to-face test; the computer-mediated test produced language that was more literate and more lexically complex; the written nature of computer-mediated communication caused certain language features to become more salient than in oral interactions; and many interactional features occurred less frequently in the computer-mediated discourse. In addition, the computer-mediated test provided no method for measuring students' intelligibility, either directly or indirectly through correlations with scores on other sub-constructs. Thus, in spite of the similarities of test scores, the lack of evidence in support of the other assumptions—and in particular the assumptions of similarity of language and interactions—requires us to conclude that students' performance on the computer-mediated test in this study differed from that on the face-to-face test in a number of ways that preclude making inferences about their oral abilities on the basis of their computer-mediated interactions.

5.8 IMPLICATIONS FOR CMC IN LANGUAGE TESTING

Although this study has rejected the notion that computer-mediated performance can serve as an indicator of oral performance, it does not rule out a role for computer-mediated communicative second language testing. The arguments for better convergence and integration of instruction and assessment in classes that utilize CMC (Kost & Jurkowitz, 2002; Jurkowitz, 2002) have not been affected by the results of this study. Teachers should use assessment methods that reflect the types of activities experienced by students and, thus, should include computer-mediated assessment activities representative of computer-mediated classroom activities. A corollary of this argument is that, unless future research suggests otherwise, computer-mediated communicative tests should probably not be used unless students have prior second language experience with the classroom use of CMC.

Whether computer-mediated tests may totally replace oral language testing in contexts where CMC is an important medium for second language interaction will depend on the goals of the curriculum. Where the curriculum aims to develop oral proficiency, then assessment should clearly reflect that goal by including an oral component which, presumably, would correspond to equivalent face-to-face interactions in the classroom. The pursuit of alternative goals to oral proficiency—such as the development of 'electronic literacy' (Shetzer & Warschauer, 2000)—would allow examiners to rely solely on computer-mediated communicative tests.

The experiences gained during this study result in several suggestions for test administration and scoring when using a computer-mediated communicative test. First of all, computer-mediated tests should be longer than the equivalent face-to-face test would be, not only to compensate for the slower rate of language production caused by typing utterances rather than speaking them, but also to allow for the time that students may spend engaging in greeting routines. In fact, it might be preferable to have students log in to the chat room and complete such introductory and warm-up activities prior to formally starting the test so that no test time is wasted on language production that may be irrelevant to the goals of testing. Based on results of this study, computer-mediated communicative tests need to be about two and a half times the length of a face-to-face test if language samples of comparable sizes are to be elicited. Thus, for the assessment context in this study, the test should have lasted approximately 25-30 minutes to obtain a language sample equivalent in size to that produced during the face-to-face test.

Second, group size needs to be set so as to facilitate the rating process. In the present study, raters initially experienced some difficulty in separating an individual's contributions to the discussion from those of the other two group members. While color coding each turn to indicate the speaker probably helped alleviate this problem, the low inter-rater reliability for the computer-mediated test and the unexpectedly slow speed with which transcripts were read may be evidence that the issue was not totally resolved in spite of the addition of a time-consuming, extra step. Limiting group size to dyads may, thus, be necessary to ensure ease, efficiency, and reliability of rating.

Finally, the low inter-rater reliabilities on the computer-mediated test suggest an important and pressing need for extensive rater training if such tests are to become a regular assessment tool in language programs where they will be administered and scored by multiple examiners. The present study has suggested several possible factors which may contribute to low inter-rater reliabilities: multiple interpretations of the grading rubric; different approaches to reading and analyzing CMC transcripts; and diverse levels of experience and skill with deciphering the flow of ideas in CMC discourse and identifying inter-textual links between comments that may be separated by several turns. All of these factors may be influenced by a training program which (a) clearly defines the criteria to be evaluated, (b) exemplifies and clarifies the important differences between bands for each criteria, (c) identifies and practices acceptable strategies for reading transcripts, and (d) familiarizes raters with the interactional patterns of computer-mediated communication.

5.9 THEORETICAL AND PEDAGOGICAL IMPLICATIONS

Chapter 2 explained that many proponents of CMC write from an interactionist perspective. That is, they suggest that the use of CMC in the classroom can enhance language development and promote communicative competence. Studies of CMC in second language classrooms (Blake, 2000; Pellettieri, 2000; Smith, 2001; Fernandez-

Garcia & Martinez-Arbelaiz, 2002) have found evidence of the negotiation of meaning which is assumed to play an important role in language acquisition (Gass, 1997). However, none of these previous classroom-based CMC studies has directly compared students' levels of negotiation of meaning in computer-mediated versus face-to-face interaction. The present study allowed such a comparison. Although learners did negotiate meaning in their computer-mediated discussions, they did so much less than they did in the face-to-face discussions. Language-related episodes occurred seven times as often in the face-to-face discussions as they did in the computer-mediated discussions. In fact, the computer-mediated data contained just five instances of language-related episodes compared to 87 in the face-to-face data.

If negotiation of meaning is important to language acquisition, it follows that the more students engage in negotiating meanings and discussing language, the better their language acquisition will be. A corollary of this point is that if learners engage in negotiation of meaning less frequently in a computer-mediated activity than they do face-to-face—as is the case in this study—the use of CMC for an activity may result in lower levels of language acquisition than would be achieved though oral interaction. Of course, caution must be exercised when generalizing results from an assessment situation to a classroom context because adding an evaluative component to an activity can change behavior in important ways that may not be relevant to pedagogy. However, this study's findings concerning lower levels of negotiation of meaning do raise important questions about the use of computer-mediated communication to promote language acquisition. The presence of negotiation routines in transcripts is not sufficient to justify the use of

CMC if oral interaction can be shown to produce a better environment for language acquisition by offering more opportunities for negotiation of meaning. Thus, there is a clear need for further research about the comparative effects of computer-mediated and oral interactions on language acquisition.

This need for more comparative research is compounded when one takes into account the concept of washback, that is, the influence of tests on teaching and learning (Bailey, 1996). Notwithstanding the recommendation above that computer-mediated communicative tests only be used when CMC is a regular part of instruction, it is possible that program administrators may be attracted to the potential for fast, efficient testing of large numbers of students offered by computer-mediated communicative tests and may, thus, attempt to incorporate such tests into their programs even if CMC is not integrated into the curriculum. In order to prepare students for the computer-mediated tests, teachers may decide to include not only more computer-mediated tasks (at the expense of oral tasks) but also to allocate time to training students in the use of chat software. While the problem of generalizing from an assessment to a classroom context means that we do not know yet what actual impact these changes may have on learning and language acquisition, interactionist theories suggest that if the results of this study *were* replicated in the classroom, the impact would probably be negative, not only because of the lower levels of negotiation of meaning that would occur in the computer-mediated discourse, but also because of the lower levels of language production in CMC. Swain (1985) claims that input alone cannot be a sufficient condition for language acquisition to occur; students also need opportunities to produce ‘comprehensible output’. Thus, if language teachers wish to maximize opportunities for language acquisition, the data in this study suggest that, all other things being equal, teachers should prefer oral activities over computer-mediated activities because the former may allow greater levels of language production for a given time period, especially if teacher involvement is minimal.

5.10 LIMITATIONS OF THIS STUDY

This study has shown that one potential confounding variable—discussion topic—was not an influence because none of the analyses found a significant effect for test prompt either as a main effect or in interaction with test method. Since preceding sections of this chapter have discussed the potential confounding roles of the time limit imposed on the computer-mediated test, of raters' differential approaches to and experiences with computer-mediated data, and of teacher interventions during the face-to-face test, nothing more will be said about these limitations here. However, this study suffers from further limitations. The first is that both tests used a single type of task—discussion of a topic—to elicit second language data. Other CMC tasks may result in linguistic and interactional behavior that more closely resembles that found in oral tests. The second limitation is the participant pool. This study has a relatively low number of participants, all of whom were drawn from a single course and may, thus, be regarded as representing a very limited range of proficiencies.

5.11 DIRECTIONS FOR FUTURE RESEARCH

In many ways, the present study raises more questions about the use of computer-mediated communicative tests than it answers. The limitations described in the previous section suggest a need to explore the effect of a wider range of test tasks on students at a full range of proficiencies. What is the effect of using different tasks? Do other tasks produce linguistic and interactional behavior that more closely resembles that found in oral tests? If so, do they do so at all proficiency levels? Do personality and motivational factors play similar roles during computer-mediated tests to those found in studies of classroom-based CMC? What are the roles of levels of comfort and expertise with technology? How does typing ability affect performance on a computer-mediated test?

Perhaps the biggest question that the present study leaves unresolved is what sorts of inferences can validly be made from performance on a computer-mediated test. The results here suggest that the language produced in a computer-mediated test differs from that produced in a face-to-face test. What, then, does it resemble? Is it the language that students produce when asked to write in a particular genre? If so, which one(s)? Or does students' CMC discourse represent something completely different with its own unique characteristics? Clearly, these questions need to be resolved if computer-mediated tests are to be used with any validity.

This study has also suggested a need for further classroom-based CMC research in two areas. First, more studies should avail themselves of software which records keystrokes in order to investigate the use of communication strategies such as restructuring, message reduction, and message abandonment which typically cannot be seen in CMC transcripts because they occur prior to sending the message to the chat room. Second, future studies should directly compare levels of negotiation of meaning across face-to-face and computer-mediated interactions and explore the effects of that negotiation on language acquisition, taking into account other factors such as personality type and motivation level. Direct classroom-based comparisons are necessary to determine whether the lower levels of meaning negotiation found in this study are typical of CMC in general or are an artifact of its use for assessment.

Finally, it is a truism that technology is constantly changing. While this study has ruled out the use of text-based computer-mediated communication where test users need to infer students' oral abilities, current technological developments are beginning to allow spoken computer-mediated interactions, as chat software starts to integrate video conferencing capabilities. We are probably several years from the widespread adoption of 'video chat' for instructional or assessment purposes, but the potential for such systems to overcome many of the limitations of text-based computer-mediated communicative tests found in this study suggests that the latter will become an increasingly important area of study as researchers probe their usefulness for both classroom assessment and proficiency testing.

APPENDIX 1

CHARACTERISTICS OF PARTICIPANTS

DAY CLASS

Student	Gender	Age	First Language	Time in France (weeks)
1	female	19	Arabic	0
2	female	21	English	2
3	male	21	English	112
4	female	21	English	0
5	male	19	English	2
6	male	20	English	0
7	male	22	English	0
8	female	19	English	0
9	female	19	English	0
10	male	34	English	8
11	female	19	English	1
12	female	19	English	4

NIGHT CLASS

Student	Gender	Age	First Language	Time in France (weeks)
13	female	19	English	8
14	female	18	English	4
15	female	20	English	0
16	male	24	English	0
17	female	19	English	0
18	female	19	English	0
19	male	18	English	0
20	female	19	English	0
21	female	19	English	0
22	female	20	English	17
23	male	25	English	1
24	female	21	English	0
25*	male	25	Ewe	0
26*	female	21	English	0
27*	male	27	Spanish	0
28*	female	23	English	0
29*	female	20	English	2
30*	female	20	German	0

Note: * Participants' data was not analyzed.

APPENDIX 2

POST-TEST QUESTIONNAIRE

Thank you for participating in this research. Please answer all the questions below. Your answers will remain anonymous. However, you are being asked to supply your name so that your responses on this survey may be compared to your responses on a survey taken after your Oral Interview test.

Name: ____

Questions about the IRC Français test

Please complete the following by placing a circle around the most appropriate answer:

1. I believe that the IRC Français test provides an examiner with an accurate idea of my ability in French.

Strongly Agree Agree No Opinion Disagree Strongly Disagree

2. I felt nervous before the IRC Français test.

Strongly Agree Agree No Opinion Disagree Strongly Disagree

3. I thought the discussion topic for the IRC Français test was interesting.

Strongly Agree Agree No Opinion Disagree Strongly Disagree

4. The time allowed for the IRC Français test was too short.

Strongly Agree Agree No Opinion Disagree Strongly Disagree

5. I thought that the IRC Français test was related to what I learn in class.

Strongly Agree Agree No Opinion Disagree Strongly Disagree

6. I believe that the IRC Français test provided me with an adequate opportunity to demonstrate my ability in French.

Strongly Agree Agree No Opinion Disagree Strongly Disagree

7. I felt nervous while I was doing the IRC Français test.

Strongly Agree Agree No Opinion Disagree Strongly Disagree

8. I liked doing the IRC Français test.

Strongly Agree Agree No Opinion Disagree Strongly Disagree

9. If I had done the IRC Français test on another day, I would have done better.

Strongly Agree   Agree   No Opinion   Disagree   Strongly Disagree

10. I thought the IRC Français test was too difficult.

Strongly Agree   Agree   No Opinion   Disagree   Strongly Disagree

11. I believe I did well on the IRC Français test.

Strongly Agree   Agree   No Opinion   Disagree   Strongly Disagree

12. I would have performed better on the IRC Français test with different students in my group.

Strongly Agree   Agree   No Opinion   Disagree   Strongly Disagree

13. If you had to choose either the face-to-face or the computer-based test, which one would you prefer? Circle your answer:

Face-to-face   Computer-based   No Preference

Please explain why you made this choice:

APPENDIX 3

FINAL GRADING RUBRIC

Accuracy	5	Few or no grammatical/lexical errors; no interference with meaning.
	4	Many grammatical/lexical errors; no interference with meaning.
	3	Many grammatical/lexical errors; occasional interference with meaning.
	2	Many grammatical/lexical errors; frequent interference with meaning.
	1	Grammatical/lexical errors make language almost entirely incomprehensible.
Range	5	Excellent range of structure and lexis.
	4	Good range of structure and lexis.
	3	Adequate range of structure and lexis.
	2	Limited range of structure and lexis.
	1	No communication.
Flexibility	5	Wide range of interactive strategies; strategies are always effective.
(teacher intervention lowers grade)	4	Wide range of interactive strategies; strategies not always effective.
	3	Limited range of interactive strategies; strategies always effective.
	2	Limited range of interactive strategies, strategies not always effective.
	1	No evidence of interactive strategies.

Contribution (teacher intervention lowers grade)	5	Outstanding contribution in terms of size and substantiveness.
	4	Good contribution in terms of size and substantiveness.
	3	Adequate contribution in terms of size and substantiveness.
	2	Minimal contribution in terms of size and substantiveness.
	1	No contribution in terms of size and substantiveness.
Overall Effectiveness	5	Excellent ability to communicate.
	4	Good ability to communicate.
	3	Average ability to communicate.
	2	Limited ability to communicate.
	1	No ability to communicate.
Intelligibility (used only for FTF test)	5	Few phonological errors, mostly comprehensible.
	4	Many phonological errors, but mostly comprehensible.
	3	Many phonological errors, about half comprehensible.
	2	Many phonological errors, only occasional phrases comprehensible.
	1	Phonological errors make language incomprehensible.

APPENDIX 4

INDIVIDUAL DATA (LINGUISTIC ANALYSES)

	Total Words		L2 Words		Type/Token Ratio		Lexical Density	
Student	CMC	FTF	CMC	FTF	CMC	FTF	CMC	FTF
1	183	468	183	468	0.74	0.62	0.42	0.16
2	89	271	89	264	0.70	0.68	0.20	0.06
3	177	912	177	909	0.78	0.70	0.40	0.07
4	74	120	69	110	0.80	0.62	0.52	0.28
5	123	288	123	288	0.66	0.70	0.45	0.19
6	55	75	55	74	0.76	0.76	0.44	0.31
7	70	212	70	209	0.72	0.68	0.54	0.18
8	143	255	143	248	0.66	0.62	0.49	0.28
9	59	277	59	206	0.74	0.64	0.46	0.03
10	99	180	99	175	0.66	0.60	0.45	0.23
11	44	146	44	133	N/A	N/A	0.41	0.10
12	115	186	115	186	0.66	0.76	0.42	0.26
13	88	107	88	105	0.72	0.70	0.44	0.36
14	88	238	86	229	0.78	0.78	0.45	0.16
15	100	350	100	339	0.78	0.48	0.44	0.11
16	47	159	46	140	N/A	N/A	0.52	0.13
17	120	366	120	346	0.70	0.64	0.42	0.12
18	106	162	105	151	0.80	0.58	0.42	0.28
19	89	268	88	254	0.74	0.68	0.49	0.14
20	51	110	51	99	0.78	0.70	0.49	0.22
21	88	278	88	263	0.52	0.52	0.51	0.15
22	125	637	125	592	0.72	0.60	0.45	0.07
23	55	167	54	162	0.72	0.68	0.46	0.14
24	19	41	19	35	N/A	N/A	0.37	0.17

Note: CMC = computer-mediated test, FTF = group oral exam

	Structural Complexity				Errors/10 Words	
	Coordination Index		Complexity Ratio			
Student	CMC	FTF	CMC	FTF	CMC	FTF
1	2.38	2.40	1.37	2.36	0.22	0.13
2	2.50	2.43	2.00	2.76	1.46	0.33
3	2.75	2.44	1.50	1.67	0.51	0.36
4	2.00	2.00	1.90	2.38	1.89	1.17
5	2.20	2.27	2.36	1.29	0.65	0.83
6	2.00	2.33	3.00	2.71	1.82	1.47
7	2.00	2.00	2.75	2.20	1.00	0.47
8	2.25	2.70	2.78	2.33	0.63	0.90
9	2.00	2.00	8.00	3.71	1.02	0.43
10	2.00	2.14	1.75	2.33	0.91	0.67
11	0.00	2.33	0.00	2.29	0.91	0.75
12	3.00	2.22	2.33	2.20	1.13	0.59
13	2.00	2.00	2.83	3.63	1.59	0.28
14	2.20	2.78	1.64	1.84	1.25	0.71
15	2.25	2.71	2.78	2.74	0.70	0.40
16	2.00	2.00	5.50	4.50	2.55	0.69
17	2.43	2.31	1.41	2.13	0.75	0.57
18	2.33	2.13	1.71	1.53	1.13	0.86
19	2.00	2.00	3.83	6.60	0.67	0.71
20	2.00	2.00	7.00	6.25	1.18	0.82
21	2.50	2.60	2.80	3.31	0.45	0.94
22	2.00	2.24	2.07	3.24	2.08	0.88
23	2.00	2.00	7.00	11.25	1.64	0.78
24	0.00	2.00	0.00	3.50	2.11	1.46

Note: CMC = computer-mediated test, FTF = group oral exam

APPENDIX 5

INDIVIDUAL DATA (INTERACTION ANALYSIS)

	Number Turns		Turn Length		Communication Strategies	
Student	CMC	FTF	CMC	FTF	CMC	FTF
1	12	28	15.25	16.71	0.00	0.04
2	11	25	8.09	10.84	0.11	0.48
3	9	28	19.67	32.57	0.00	0.14
4	15	13	4.93	9.23	0.68	0.75
5	15	12	8.20	24.00	0.00	0.24
6	10	12	5.50	6.25	0.00	0.53
7	9	10	7.78	21.20	0.00	0.19
8	16	34	8.94	7.50	0.00	0.31
9	12	37	4.92	7.49	0.00	1.34
10	8	16	12.38	11.25	0.10	0.39
11	4	16	11.00	9.13	0.00	0.75
12	7	21	16.43	8.86	0.00	0.05
13	10	19	8.80	5.63	0.00	0.19
14	11	21	8.00	11.33	0.11	0.42
15	10	22	10.00	15.91	0.10	0.23
16	9	17	5.22	9.35	0.64	1.19
17	11	18	10.91	20.33	0.00	0.46
18	13	9	8.15	18.00	0.09	0.25
19	16	28	5.56	9.57	0.11	0.56
20	11	21	4.64	5.24	0.00	0.55
21	11	22	8.00	12.64	0.00	0.50
22	18	33	6.94	19.30	0.00	0.46
23	13	24	4.23	6.96	0.36	0.18
24	4	12	4.75	3.42	0.00	0.73

Note: CMC = computer-mediated test, FTF = group oral exam

REFERENCES

Abrahms, Z. I. (2001). Computer-mediated communication and group journals: Expanding the repertoire of participant roles. *System, 29*(4), 489-503.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

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

Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing, 13*(3), 257-279.

Baker, L. F., Allen-Bleuze, R., Border, L. L. B., Grace, C., Owen, J. B., Serratrice, M. A., et al. (1997). *Montage* (3rd. ed.). New York: McGraw-Hill.

Barson, J., & Debski, R. (1996). Calling back CALL: Technology in the service of foreign language learning based on creativity, contingency, and goal-oriented activity. In M. Warschauer (Ed.), *Telecollaboration in foreign language learning* (pp. 49-68). Manoa, Hawaii: Second Language Teaching and Curriculum Center.

Bauman, R. (2001). Genre. In A. Duranti (Ed.), *Key terms in language and culture.* London: Blackwell.

Beauvois, M. H. (1992). Computer-assisted classroom discussion in the foreign language classroom. *Foreign Language Annals, 25*, 455-464.

Beauvois, M. H. (1997). Computer-mediated communication: Technology for improving speaking and writing. In M. D. Bush & R. M. Terry (Eds.), *Technology enhanced language learning* (pp. 165-184). Lincolnwood, IL: National Textbook Company.

Beauvois, M. H. (1998). Write to speak: The effects of electronic communication on the oral achievement of fourth semester students. In J. A. Muyskens (Ed.), *New ways of learning and teaching: Focus on technology in foreign language education* (pp. 93-115). Boston, MA: Heinle & Heinle.

Berkoff, N. A. (1985). Testing oral proficiency: A new approach. In Y. P. Lee (Ed.), *New directions in language testing* (pp. 93-100). Oxford: Pergamon Institute of English.

Bernhardt, E., & Kamil, M. (1998). Enhancing foreign language culture learning through electronic discussion. In J. A. Muyskens (Ed.), *New ways of learning and teaching: Focus on technology in foreign language education* (pp. 39-55). Boston, MA: Heinle & Heinle.

Berry, V. (2000). *An investigation into how individual differences in personality affect the complexity of language test tasks*. Kings College, University of London, London.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1996). Investigating language use through corpus-based analyses of association patterns. *International Journal of Corpus Linguistics, 1*(2), 171-197.

Biber, D. (1999). A register perspective on grammar and discourse: Variability in the form and use of English complement clauses. *Discourse Studies, 1*(2), 131-150.

Biber, D., & Conrad, S. (2001). Register variation: A corpus approach. In D. Schiffrin, D. Tannen & H. Hamilton (Eds.), *Handbook of discourse analysis* (pp. 175-196). London: Blackwell.

Blake, R. (2000). Computer mediated communication: A window on L2 Spanish interlanguage. *Language Learning and Technology, 4*(1), 120-136.

Bonk, C. J., & Cunningham, D. J. (1998). Searching for learner-centered, constructivist, and sociocultural components of collaborative educational learning tools. In C. J. Bonk & K. S. King (Eds.), *Electronic collaborators: Learner centered technologies for literacy, apprenticeship, and discourse* (pp. 25-50). Mahwah, NJ: Lawrence Erlbaum Associates.

Brammerts, H. (1996). Language learning in tandem using the internet. In M. Warschauer (Ed.), *Telecollaboration in foreign language learning* (pp. 121-130). Honolulu: University of Hawaii Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81-105.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Carter, R., & McCarthy, M. (1997). Speech genres. In *Exploring spoken English.* Cambridge: Cambridge University Press.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language* (pp. 35-54). Norwood, NJ: Ablex.

Chafe, W. L. (1994). *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago: University of Chicago Press.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*(254-272).

Chapelle, C. A., & Douglas, D. (1993). Foundations and directions for a new decade of language testing. In D. Douglas & C. A. Chapelle (Eds.), *A new decade of language testing research.* Alexandria, VA: TESOL.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Chun, D. L. (1994). Using computer networking to facilitate the acquisition of interactive competence. *System, 22*(1), 17-31.

Clark, G. (1990). Discourse in dialogue; The social context of writing. In *Dialogue, dialectic, and conversation: A social perspective on the function of writing* (pp. 1-18). Carbondale: Southern Illinois University Press.

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Boston, MA: Heinle & Heinle.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-219). New York: Macmillan.

Collot, M., & Belmore, N. (1996). Electronic language: A new variety of English. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 13-28). Philadelphia: John Benjamins.

Colomb, G. G., & Simutis, J. A. (1996). Visible conversation and academic inquiry: CMC in a culturally diverse classroom. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 203-224). Philadelphia: John Benjamins Publishing Company.

Cononelos, T., & Oliva, M. (1993). Using computer networks to enhance foreign language/culture education. *Foreign Language Annals, 26*, 527-533.

Corder, S. P. (1983). Strategies of communication. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 15-19). New York: Longman.

Courchene, R. J., & de Bagheera, J. I. (1985). A theoretical framework for the development of performance tests. In P. C. Hauptman, R. LeBlanc & M. B. Wesche (Eds.), *Second language performance testing* (pp. 45-58). Ottawa: University of Ottawa Press.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum.

Cronbach, L. J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement theory and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.

Cumming, A. (1996). Introduction: The concept of validation in language testing. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 1-14). Clevedon: Multilingual Matters.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621-694). Washington D.C.: American Council on Education.

Darhower, M. (2002). Instructional features of synchronous computer-mediated communication in the intermediate L2 class: A sociocultural case study. *CALICO Journal, 19*(2), 249-277.

Davis, B., & Thiede, R. (2000). Writing into change: Style shifting in the foreign language curriculum. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 87-120). Cambridge: Cambridge University Press.

Duranti, A. (1997). Speech events: From functions of speech to social units. In *Linguistic anthropology* (pp. 284-294). Cambridge: Cambridge University Press.

Faerch, C., & Kasper, G. (1983). Plans and strategies in foreign language communication. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 20-60). New York: Longman.

Fernandez-Garcia, M., & Martinez-Arbelaiz, A. (2002). Negotiation of meaning in nonnative speaker-nonnative speaker synchronous discussions. *CALICO Journal, 19*(2), 279-294.

Ferrara, K., Brunner, H., & Whittemore, G. (1991). Interactive written discourse as an emergent register. *Written Communication, 8*(1), 8-34.

Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal, 30*, 156-167.

Frederiksen, J. R., & Collins, A. (1989). A system's approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Fulcher, G. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51.

Gaines, J. (1999). Electronic mail - a new style of communication or just a new medium?: An investigation into the text features of e-mail. *English for Specific Purposes, 18*, 81-101.

Gass, S. (1997). *Input, interaction, and the second language learner.* Mahwah, NJ: Lawrence Erlbaum.

Genesee, F., & Upshur, J. A. (1996). *Classroom-based evaluation in second language education.* Cambridge: Cambridge University Press.

Georgakopoulou, A., & Goutsos, D. (1997). Genres of discourse types: Spoken and written. In A. Georgakopoulou & D. Goutsos (Eds.), *Discourse analysis: An introduction*. Edinburgh: Edinburgh University Press.

Goffman, E. (1967). *Interaction ritual*. Garden City, NY: Anchor Books.

Gonzalez-Bueno, M. (1998). The effects of electronic mail on Spanish L2 discourse. *Language Learning and Technology, 1*(2), 55-70.

Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. *Language Testing, 3*, 159-185.

Halliday, M. A. K. (1989). *Spoken and written language* (2nd ed.). Oxford: Oxford University Press.

Hatch, E. (1978a). Acquisition of syntax in a second language. In J. C. Richards (Ed.), *Understanding second and foreign language learning: Issues and approaches* (pp. 34-69). Rowley, MA: Newbury House.

Hatch, E. (1978b). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 401-435). Rowley, MA: Newbury House.

Henning, G. (1987). *A guide to language testing*. Cambridge, MA: Newbury House.

Hilsdon, J. (1991). The group oral exam: Advantages and limitations. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 189-197). London: Modern English Publications.

Holec, H. (1985). You did say 'oral interactive discourse'? In P. Riley (Ed.), *Discourse and learning* (pp. 21-34). London: Longman.

Hymes, D. H. (1972a). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Middlesex: Penguin.

Hymes, D. (1972b). Models of interaction of language and social life. In J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: Ethnography of communication* (pp. 35-71). New York: Holt, Rinehart & Winston.

Jakobsen, R. L. (1960). The speech event and functions of language. In L. Waugh & M. Monville-Burston (Eds.), *On language* (pp. 72-79). Cambridge, MA: Harvard University Press.

Johnson, L. C. (1996). The keypal connection. In M. Warschauer (Ed.), *Telecollaboration in foreign language learning* (pp. 131-143). Honolulu: University of Hawaii Press.

Johnston, B. (1999). Theory and research: Audience, language use, and language learning. In J. Egbert & E. Hanson-Smith (Eds.), *CALL environments: Research, practice, and critical issues* (pp. 55-64). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Jurkowitz, L. (2002). C'est cool! French student engage in a collaborative computer-mediated final exam: A case study. *Arizona Working Papers in Second Language Acquisition and Teaching, 9*, 19-52.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6*, 125-160.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527-535.

Kelm, O. R. (1992). The use of synchronous computer networks in second language instruction: A preliminary report. *Foreign Language Annals, 25*(5), 441-454.

Kelm, O. R. (1996). The application of computer networking in foreign language education: Focusing on principles of foreign language education. In M. Warschauer (Ed.), *Telecollaboration in foreign language learning* (pp. 19-28). Honolulu: University of Hawaii Press.

Kenyon, D. M. (1992). *Introductory remarks at symposium on development and use of rating scales in language testing.* Paper presented at the Language Testing Research Colloquium, Vancouver, B.C.

Kern, R. (1995). Restructuring classroom interaction with networked computers: Effects on quantity and quality of language production. *Modern Language Journal, 79*(4).

Kern, R. (1996). Computer-mediated communication: Using e-mail exchanges to explore personal histories in two cultures. In M. Warschauer (Ed.), *Telecollaboration in foreign language learning* (pp. 105-119). Honolulu: University of Hawaii Press.

Kern, R. (1998). Technology, social interaction, and FL literacy. In J. A. Muyskens (Ed.), *New ways of learning and teaching: Focus on technology in foreign language education* (pp. 57-92). Boston, MA: Heinle & Heinle.

Kern, R., & Warschauer, M. (2000). Introduction: Theory and practice of network-based language teaching. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 1-19). Cambridge: Cambridge University Press.

Kost, C., & Jurkowitz, L. (2002). *Using computer-assisted classroom discussion (CACD) as an authentic assessment tool.* Paper presented at the CALICO, University of California, Davis.

Krashen, S., & Terrell, T. D. (1983). *The natural approach: Language acquisition in the classroom.* Harlow: Prentice Hall.

Kunnan, A. J. (Ed.). (1998). *Validation in language assessment.* Mahwah, NJ: Lawrence Erlbaum Associates.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* New York: McGraw-Hill.

Landy, F. J. (1986). Stamp collecting versus science: Validation as hypothesis testing. *American Psychologist, 41*, 1183-1192.

Lee, L. (1997). Using internet tools as an enhancement of L2 cultural teaching and learning. *Foreign Language Annals, 30*, 410-427.

Linn, R. L., Baker, E. L., & Dunbar, S., B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Lombardo, L. (1984). *Oral testing: Assessing the language learner's ability to process discourse.* Rome, Italy: Centro Informazione Stampa Univ.

Long, M. (1981). Input, interaction, and second language acquisition. In H. Winitz (Ed.), *Annals of the New York Academy of Sciences: Vol. 379. Native language and foreign language acquisition* (pp. 259-278). New York: New York Academy of Sciences.

Long, M. (1983). Linguistic and conversational adjustments to nonnative speakers. *Studies in Second Language Acquisition, 5*, 177-193.

Long, M. (1985). Input and second language acquisition theory. In S. Gass & D. Madden (Eds.), *Input in second language acquisition* (pp. 377-393). Rowley, MA: Newbury House.

Madsen, H. S. (1983). *Techniques in testing.* Oxford: Oxford University Press.

Maynor, N. (1994). The language of electronic mail: Written speech? In M. Montgomery & G. Little (Eds.), *Centennial usage studies* (pp. 48-54). Alabama: University of Alabamas Press.

McNamara, T. F. (1996). *Measuring second language performance.* New York: Addison Wesley.

McNamara, T. F. (1997). Interaction in second language performance assessment: Whose performance? *Applied Linguistics, 18*(1), 446-466.

Meagher, M. E., & Castanos, F. (1996). Perceptions of American culture: The impact of an electronically-mediated cultural exchange program on Mexican high school students. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social, and cross-cultural perspectives* (pp. 187-201). Philadelphia: John Benjamins Publishing Co.

Meskill, C., & Ranglova, K. (2000). Sociocollaborative language learning in Bulgaria. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 20-40). Cambridge: Cambridge University Press.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*, 1012-1027.

Messick, S. (1988). The once and future meanings of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (Third ed., pp. 13-104). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Meunier, L. E. (1998). Personality and motivational factors in computer-mediated foreign language communication (CMFLC). In J. A. Muyskens (Ed.), *New ways of learning and teaching: Focus on technology in foreign language education* (pp. 145-197). Boston, MA: Heinle & Heinle.

Morrison, D. M., & Lee, N. (1985). Simulating an academic tutorial: A test validation study. In Y. P. Lee (Ed.), *New directions in language testing* (pp. 85-92). Oxford: Pergamon Institute of English.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258.

O'Donnell, R. (1974). Syntactic differences between speech and writing. *American Speech, 49*, 102-110.

Oliva, M., & Pollastrini, Y. (1995). Internet resources and second language acquisition: An evaluation of virtual immersion. *Foreign Language Annals, 28*, 551-559.

Oller, J. W. (1979). *Language tests at school: A pragmatic approach.* London: Longman.

Olson, D. (1977). From utterance to text: The basis of language in speech and writing. *Harvard Educational Review, 47*, 257-281.

Palmer, A. S., & Groot, P. J. M. (1981). An introduction. In A. S. Palmer, P. J. M. Groot & G. A. Trosper (Eds.), *The construct validation of tests of communicative competence* (pp. 1-11). Washington D.C.: TESOL.

Pantelidis, V. (1995). Reasons to use virtual reality in the classroom. *VR in The Schools, 1*(1), 9.

Payne, J. S., & Whitney, P. J. (2002). Developing L2 oral proficiency through synchronous CMC: Output, working memory, and interlanguage development. *CALICO Journal, 20*(1), 7-32.

Pellettieri, J. (2000). Negotiation in cyberspace: The role of *chatting* in the development of grammatical competence. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 59-86). Cambridge: Cambridge University Press.

Peyton, J. K. (1999). Theory and research: Interaction via computers. In J. Egbert & E. Hanson-Smith (Eds.), *CALL environments: Research, practice, and critical issues* (pp. 17-26). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Pinto, D. (1996). What does 'schMOOze' mean?: Non-native speaker interactions on the internet. In M. Warschauer (Ed.), *Telecollaboration in foreign language learning* (pp. 165-184). Honolulu: University of Hawaii Press.

Popham, W. J. (1997). Consequential validity: Right concern - wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13.

Poulisse, N. (1990). *The use of compensatory strategies by Dutch learners of English.* Dordrecht: Foris.

Reves, T. (1980). The group oral test: An experiment. *English Teachers' Journal, 24*, 19-21.

Reves, T. (1991). From testing research to educational policy: A comprehensive test of oral proficiency. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 178-188). London: Modern English Publications.

Sanchez, B. (1996). MOOving to a new frontier in language learning. In M. Warschauer (Ed.), *Telecollaboration in foreign language learning* (pp. 145-163). Honolulu: University of Hawaii Press.

Shephard, L. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Shephard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8.

Shetzer, H., & Warschauer, M. (2000). An electronic literacy approach to networked-based language teaching. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 171-185). Cambridge: Cambridge University Press.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*, 99-123.

Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal, 40,* 212-220.

Shohamy, E., Gordon, C., Kenyon, D. M., & Stansfield, C. W. (1989). The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Higher Hebrew Education, 4.*

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

Smith, B. (2001). *Taking students to task: Task-based computer-mediated communication and negotiated interaction in the ESL classroom.* Unpublished Dissertation, University of Arizona, Arizona.

Soh, B.-L., & Soon, Y.-P. (1991). English by e-mail: Creating a global classroom via the medium of computer technology. *ELT Journal, 45*(4), 287-292.

Sotillo, S. M. (2000). Discourse functions and syntactic complexity in synchronous and asynchronous communication. *Language Learning and Technology, 4*(1), 82-119.

Spitzer, M. (1986). Writing style in computer conferences. *IEEE Transactions of Professional Communication, PC-29*(1), 19-22.

Stevenson, D. K. (1985). Authenticity, validity, and a tea party. *Language Testing, 2*, 41-47.

Sullivan, N., & Pratt, E. (1996). A comparative study of two ESL writing environments: A computer-assisted classroom and a traditional oral classroom. *System, 24*(4), 490-501.

Swain, M. (1985). Communicative competence: Some roles for comprehensible input and comprehensible output in its development. In S. Gass & D. Madden (Eds.), *Input in second language acquisition* (pp. 235-253). Rowley, MA: Newbury House.

Swain, M. (2001). Examining dialogue: Another approach to content specificiation and to validating inferences drawn from test scores. *Language Testing, 18*(3), 275-302.

Swales, J. (1990). The concept of genre. In *Genre analysis: English in academic and research settings* (pp. 33-67). Cambridge: Cambridge University Press.

Tannen, D. (1982). The oral/literate continuum in discourse. In D. Tannen (Ed.), *Spoken and written language* (pp. 1-16). Norwood, NJ: Ablex.

Tarone, E. (1977). Conscious communication strategies in interlanguage. In H. D. Brown (Ed.), *On TESOL '77* (pp. 195-203). Washington, D.C.: TESOL.

Tarone, E. (1983). Some thoughts on the notion of 'communication strategy'. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication.* New York: Longman.

Tarone, E., Cohen, A., & Dumas, G. (1983). A closer look at some interlanguage terminology: A framework for communication strategies. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 4-14). New York: Longman.

Tella, S. (1992). The adoption of international communications networks and electronic mail into foreign language education. *Scandinavian Journal of Educational Research, 36*(4), 303-312.

Turbee, L. (1999). Classroom practice: MOO, WOO, and more - language learning in virtual environments. In J. Egbert & E. Hanson-Smith (Eds.), *CALL environments: Research, practice, and critical issues* (pp. 346-361). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Van Handle, D., & Corl, K. A. (1998). Extending the dialogue: Using electronic mail and the internet to promote conversation and writing in intermediate level German classes. *CALICO Journal, 13*, 129-144.

Venugopal, S. N. (1992). The group discussion as a measure of oral interaction. *Guidelines, 14*(1), 45-54.

Wales, K. (2001). Genre. In K. Wales (Ed.), *A dictionary of stylistics* (pp. 176-178). London: Longman.

Warschauer, M. (1996). Comparing face-to-face and electronic communication in the second language classroom. *CALICO Journal, 13*(2), 7-26.

Warschauer, M. (2000). On-line learning in second language classrooms. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 41-58). Cambridge: Cambridge University Press.

Warschauer, M., Turbee, L., & Roberts, B. (1996). Computer learning networks and student empowerment. *System, 24*(1), 1-14.

Wilkins, H. (1991). Computer talk. *Written Communication, 8*(1), 56-78.

Wold, A. H. (1992). Oral and written language: Arguments against a simple dichotomy. In *The dialogical alternative* (pp. 175-194). Oslo: Scandinavian University Press.

Yaguello, M. (1998). What is language for? In *Language through the looking glass: Exploring language and linguistics*. Oxford: Oxford University Press.

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 29-46). Philadelphia: John Benjamins.

Yoshida-Morise, Y. (1998). The use of communication strategies in language proficiency interviews. In R. F. Young & A. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 205-238). Amsterdam, PA: John Benjamins.