The Influence of the Social Interactional Context on Test Performance: A Sociocultural View

by

Youyi Sun, BA

A thesis submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree of

Master of Arts

School of Linguistics and Applied Language Studies

Carleton University, Ottawa, Ontario, Canada

2005

Your file Votre référence
ISBN: 0-494-00728-1
Our file Notre référence
ISBN: 0-494-00728-1

Abstract

In recent years there has been increasing interest in the language testing community in incorporating a sociocultural perspective in second language performance assessment. Sociocultural theory provides language testers with interesting insights, but it also presents challenges for both test development and test validation enquiry. Some scholars (e.g. Swain, 2001) suggest, for example, that co-constructed tasks may bias for the best in language testing. This study is an attempt to explore the necessity, usefulness and feasibility of incorporating a sociocultural perspective in a small group oral language test in the English for academic purposes (EAP) context by examining the influence of the social interactional context on test performance from a sociocultural point of view. Two tasks from the Oral Language Test of the Canadian Academic English Language Assessment were used and parallel task versions were developed by changing test method facets. Results of this study show analysis of the influence of the social interactional context on test performance from a sociocultural perspective will offer language test developers and language testing researchers useful information about test development and test validation inquiry. Practical, methodological and theoretical implications are considered.

Acknowledgement

I would like to thank my thesis supervisor Professor Janna Fox for her encouragement, help and support. Without her mediation and scaffolding, this thesis would never have existed. Professor Fox was on sabbatical while I was working on this thesis, but she was always an email-away to offer me most immediate help, whichever part of the world she was in. She is the most understanding teacher I have ever met. The courses she taught during my graduate studies have intrigued my interest in language testing and our personal communications have made the process of my socialization to the language testing community so enjoyable.

I would especially like to thank Professor Devon Woods, graduate supervisor of School of Linguistics and Applied Language Studies (SLALS) and internal examiner of Examination Board, who has given me full support throughout my graduate studies at Carleton University. The questions he asked about my study have been most thought provoking.

I would like to express my sincere appreciation to other members on Examination Board: Professor Desmond Allison, director of SLALS and Chair of Examination Board, and Professor Marjorie Wesche, external examiner, Second Language Institute, University of Ottawa. Their insightful comments on my study not only contributed much to the final version of this thesis, but also have broadened my view for future research.

My thanks go to Wendy Fraser and Hong Zhao for their help and friendship in the data collection process of my study.

My thanks also go to Stuart Hannay for his technical support in the computer lab.

My special thanks go to my family ---- my wife Shuzhen and my son Haoyu ---- for their love and sacrifice.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

AP: the Audio-Pal (Task)

CAEL: (the) Canadian Academic English Language (Assessment)

CLA: communicative language ability

EAP: English for academic purposes

GC: group context

IA: interactive ability

IC: individual context

L2: second language

LPI: language proficiency interview

OET: the Occupational English Test (in Australia)

OLT: Oral Language Test (of the CAEL Assessment)

OPI: oral proficiency interview

SFL: systemic functional linguistics

SLA: second language acquisition

SR: Story Retell (Task)

TLU: target language use

ZPD: the zone of proximal development

Introduction

Mainstream second language acquisition (SLA) and language testing researchers have adopted, by and large, a highly cognitive view on language, language acquisition and language use, focusing on conceptualizing, modeling and explaining the cognitive mechanism regarding the inner-part of the individual mind which underlies language performance. While more socially oriented views have been voiced from time to time and have been well documented in other fields of language studies, they have remained relatively marginal to SLA and language testing overall.

In an individual-focused cognitive approach to language acquisition and language testing, *interaction* is an important concept, but it is generally referred to as the interaction inside the mind; the external social interaction is considered important in this tradition of mainstream SLA and language testing only to the extent that it provides input for language acquisition, or helps as an elicitation device to draw out those abilities of the candidate that are to be tested.

Accordingly, researchers in SLA and language testing have generally adopted psychometric statistical approaches in their analysis, attempting "to gain objective data by controlling human and other extraneous variables and thus gain what they consider to be reliable, hard data and replicable findings" (Davis, 1995, p. 428). As a result, the influence on performance of the social interactional context is usually treated as unwanted variance in language testing where the primary interest is the

individual's stable core abilities, which are used as the basis for making inferences about his or her performance beyond the test setting. However, the inseparability of the ability to be tested and the situational context whereby the ability is tested has presented a fundamental dilemma for language testers (Bachman, 1990).

Some researchers have attempted to solve the fundamental dilemma in a cognitive approach by integrating task characteristics and ability requirements. For example, Skehan (1998) suggests from a cognitive perspective the possibility of modeling selective effects on ability of task characteristics in terms of task difficulty, which is to be used as a basis for making generalizations from performance on one task to likely performances on tasks with related difficulty sources. Norris, Brown, Hudson, and Yoshioka (1998) and Brown, Hudson, Norris and Bonk (2002) have applied this approach in criterion-referenced language testing practice. Their research has proved the feasibility and usefulness of Skehan's (1998) approach. However, their findings also show that without a better understanding of how tasks are actually accomplished in both cognitive and "real-world" terms and what makes a given task more or less "difficult" for different examinees, the validity of inferences about language learners' abilities (of whatever sort) will remain in question.

In recent years mainstream SLA has been increasingly criticized by scholars who have argued against an exclusive individual-focused cognitive approach to language learning for its denaturalizing feature (Atkinson, 2002). Instead of seeing the language learner as a radically autonomous language *acquirer,* a growing number of researchers have viewed the language learner "as someone who experiences

productive *participation* [added emphasis] in joint activity" (Lantolf and Appel, 1994, p.11). This *participation* view of language learning is rooted in sociocultural theory, originally conceived of by Vygotsky (1987).

The major theme of sociocultural theoretical framework is that social interaction plays a fundamental role in the development and function of cognition including the development and use of language (Lantolf, 2000a).

The sociocultural view of language development has attracted increasing interest in the field of second language (L2) education in recent years. It has informed research on L2 learning as a mediated process (Lantolf, 2000b) and studies on the collaborative dialogues among students in the L2 learning context (Swain and Lapkin, 1998; Swain, 2000; Swain 2001; Swain and Lapkin, 2001). Recently, the influence of sociocultural theory is also beginning to be felt in the language testing community (e.g. Young, 2000; Swain, 2001; Fox 2002; Chalhoub-Deville, 2003).

A fundamental notion of the sociocultural approach to language testing is that language ability is local, context-bound and co-constructed. This is captured in Chalhoub-Deville's (2003) "ability-in language user-in context" representation of the L2 construct and He and Young's (1998) definition of interactional competence:

> [I]nteractional competence is fundamentally different from communicative competence. Whereas communicative competence has been interpreted in the testing literature as a trait or bundle of traits that can be assessed in a given individual, interactional competence... is co-constructed by all participants in an interactive practice and is specific to that practice. Participants' knowledge and interactive skills are local: they apply to a given interactive practice and either do not apply or apply in a different configuration to different practices. (p. 7)

The sociocultural perspective offers language testers interesting insights and has significant theoretical and practical implications for language testing. For example, since ability, context and performance are inextricably meshed in language use, performance will not be seen as simply the manifestation of the individual's ability. The more dynamic aspect of social interaction will not be considered simply as a source of measurement error. Rather, it will be seen as part of what we are trying to measure. As such, in test validation enquiry we need to describe and interpret the influences on test performance of the dynamic social interactional context from a sociocultural perspective.

The sociocultural perspective also presents challenges for language testing research. For example, the inseparability of ability and context, the dynamics of context and the co-constructedness of interactional events are difficult to reconcile with the tester's need to generalize the score assigned to the individual's test performance to make inferences about his/her performance in target language use beyond the test. With regard to validation enquiry, researchers have generally used quantitative methods, most often correlational analysis to provide test validity evidence. From a sociocultural perspective, however, quantitative analysis provides necessary but not sufficient evidence for test validity. To describe and interpret the influences on test performance of the dynamic social interactional context from a sociocultural perspective entails qualitative turn-by-turn analysis of dialogues or conversations as a source of validity evidence to reveal the way context is co-constructed by participating individuals, and the way context generates cognitive

and strategic processes. This analysis will differ from discourse analysis which focuses on linguistic and interactional features of speaking (Swain, 2001).

Therefore, while sociocultural theory offers language testers interesting insights, these insights are not fully developed and much work is needed before practical benefits might be obtained. With this consideration, a study was conducted in an English for academic purposes (EAP) program at Carleton University in Ottawa. The purpose of this study was to explore the usefulness and feasibility of incorporating a sociocultural approach in a small group oral language test in the EAP context by considering the influence on test performance of the social interactional context.

The present thesis is a report on this study.

In Chapter 1, two perspectives on interaction in SLA and language testing are reviewed. This review suggests while SLA and language testing researchers have generally taken an individual-focused cognitive perspective on interaction, focusing on the interaction inside the individual's mind – the way the cognitive mechanism processes information coming inside the mind, there is growing interest in incorporating a sociocultural approach to interaction. A sociocultural approach emphasizes both interaction in the mind and in the world. Thus it embraces a broader definition of the L2 construct. From a sociocultural point of view, language ability and the social interactional context are intricately connected. At the end of this chapter, the social interactional context is defined in terms of *genre* and *register* in a systemic functional linguistic (SFL) approach.

Chapter 2 discusses the necessity and usefulness of incorporating a sociocultural perspective in language performance assessment. Performance assessment contrasts with the traditional paper-and-pencil test in that it is more authentic and direct in terms of real-world criterion and thus has higher face validity in terms of predicting the performance of the candidate in target language use (TLU) domain. However, it is not acceptable to take the view that performance assessments are in and of themselves valid due to their authentic or direct natures, or their higher face validity. Most of the research on performance assessment validation has taken an individual-focused cognitive approach. Recent developments in task-based approaches to language testing have cast new light on performance assessment validation research by integrating task characteristics and ability requirements; however, this approach presupposes global and static features of tasks. To fully understand the validity of performance assessment we need to examine how a particular task is accomplished in real-world terms. For this it is necessary and useful to incorporate a sociocultural perspective in language performance assessment.

On the basis of the literature review, the research gap was suggested and the research question was formed. That is:

How does the social interactional context influence test taker performance?

To address this research question, a study was designed. Chapter 3 and Chapter 4 present this study.

Chapter 3 introduces the methodology of the current study. First, the Canadian Academic English Language (CAEL) Assessment and the use of its scores at Carleton

University are introduced from a sociocultural perspective. Then details of the methodology of the current study are presented. The participants of this study included one EAP teacher, three raters (the EAP teacher and other two raters) and twenty-three students enrolled in an EAP support program. Two tasks from the Oral Language Test (OLT) of the CAEL Assessment were used. Parallel task versions were developed by changing test method facets. A questionnaire was designed to elicit information on the students' perception of the test. Procedures of the study involved administration of the test and post-test questionnaire, rating of the students' performance on the test and transcribing of the recordings from the test administration. Data generated from these procedures were analyzed both quantitatively and qualitatively.

In Chapter 4, results of the analysis are presented and the current study's research question is discussed in relation to the results.

In the last chapter, some conclusions are drawn and theoretical and practical implications for language testing of incorporating a sociocultural approach are discussed in relation to findings of this study.

It is expected that the current study will provide some useful information on the development of a small group oral language performance test in the EAP context and its validation enquiry.

However, given the limitations of this study, especially in terms of small sample size and time limit of data collection, results of this study are only suggestive. To fully discuss implications of a sociocultural perspective for language testing, more

empirical studies need to be conducted.

Chapter 1

Two Perspectives on Interaction

One common interest to both SLA and language testing researchers is the definition of the L2 construct. Chapelle (1998) distinguishes among three perspectives on construct definition: a construct may be defined as a trait, as a behaviour, or as some combination of these two, which is referred to as the interactionalist definition of construct. Young (2000, p. 3) argues that "neither trait nor behaviourist definitions are satisfactory for theories of language in use" because language use involves both "knowledge, or competence, and the capacity for implementing, or executing that competence" (Bachman, 1990, p. 84) in specific contexts of use. Therefore, it is desirable to consider the interactionalist definition of L2 construct in both SLA research and communicative language testing.

Interaction is a term of long-standing discussion in both SLA and language testing research. Generally, researchers have discussed interaction from two perspectives: a cognitive perspective and a sociocultural perspective. In this chapter, first, discussions on interaction from these two perspectives will be reviewed, and then, the social interactional context will be defined from a sociocultural perspective.

1.1 Individual-focused Cognitive Perspective on Interaction

In the field of SLA research, researchers have generally adopted a cognitive perspective, which places SLA mainly within individual heads and sees individuals as radically autonomous language acquirers (Pennycook, 1997). In this tradition of mainstream SLA research, "interaction" is studied "mainly for the sake of understanding its conditioning effect on input" and on language acquisition, the focus of work being on "language-in-the-head" (Atkinson, 2002, p. 535). This is made clear by Gass (1998):

> The goal of my work (and the work of others within the input/interaction framework...) has never been to understand language use per se... but rather to understand what types of interaction might bring about what types of changes in linguistic knowledge. (p. 84)

In the field of language testing, Bachman and Palmer (1996) state that

> [l]anguage use involves complex and multiple interactions among the various individual characteristics of language users, on the one hand, and between these characteristics and the characteristics of the language use or testing situation, on the other. Because of the complexity of these interactions, we believe that language ability must be considered within an interactional framework of language use. (p. 62)

To express visually some of the major interactions that they assume to be involved in language use, Bachman and Palmer (1996, p. 63) present some components of language use and language test performance, as shown in Figure 1:

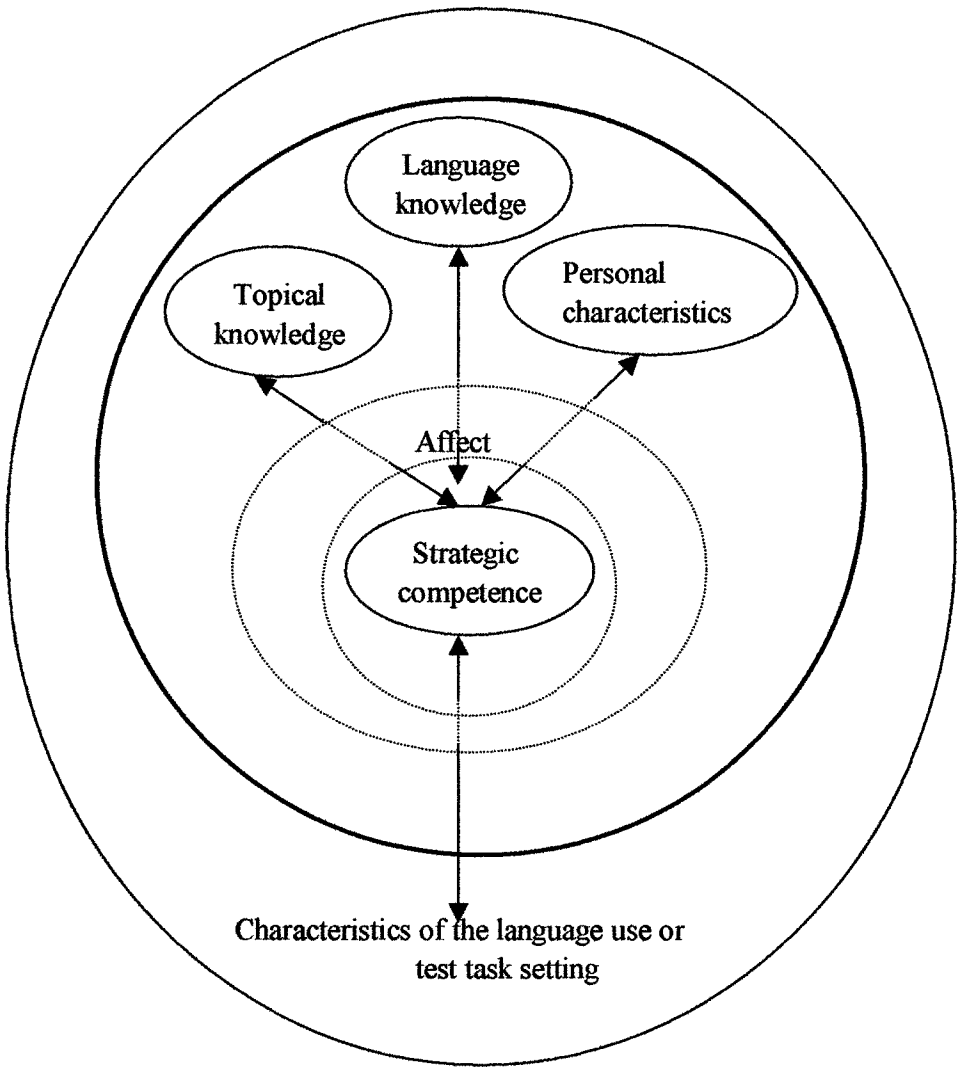

Figure 1. Some components of language use and language test performance (Bachman and Palmer, 1996, p. 63)

Bachman (1990) and Bachman and Palmer (1996) also present a theoretical model of communicative language ability (CLA)[1] model, which they "believe provides a

[1] Interestingly, Bachman and Palmer (1996) seem to deliberately avoid using the term *communicative* though they state that what they call *language ability* is essentially what Bachman (1990) calls *communicative language ability*. In language testing literature, the term *communicative language ability* or *CLA* is usually used to refer to Bachman (1990) and Bachman and Palmer's model.

valuable framework for guiding the definition of constructs for any language testing development situation" (Bachman and Palmer, 1996, p. 67). They define CLA as involving both *language knowledge*[2], and the capacity for implementing, or executing that knowledge in language use in context (*strategic competence*).

The statement of Bachman and Palmer (1996, p. 62) quoted above and Figure 1, where interactions are indicated by the double-headed arrows, show that they believe "interaction" is an important concept in language use. As such, Chapelle (1998, p.44) characterizes their CLA model as an interactionalist construct definition of communicative language ability. Bachman and Palmer emphasize the interaction between the language user, the context and the discourse by including in their CLA model strategic competence, which plays a central mediating role between knowledge structures, language competence, and context of situation through the metacognitive processes of goal setting, assessment, planning and execution.

To Bachman and Palmer (1996), interactions in language use refer to:

1. interactions between different characteristics of individuals. They list four sets of individual characteristics whose effects on language test performance they believe are better understood and considered in the way language tests are designed, developed and used. These individual characteristics include the following:

   (1) personal characteristics, such as age, sex, and native language,
   (2) the topical knowledge that test takers bring to the language testing situation,
   (3) their affective schemata, and
   (4) their language ability (p. 64)

[2] What Bachman and Palmer (1996) call *language knowledge* is referred to as *language competence* by Bachman (1990).

2. interactions between the individual (test taker or language user) and the task (test or TLU). Bachman and Palmer develop a framework of task characteristics[3], which they believe, have significant effects on performance and need to be understood and controlled. These characteristics are broken down into characteristics of the *setting,* the *test rubric,* the *input,* the *expected response,* and the *relationship between input and response* (p.59).

In principle, the CLA model might be extremely rewarding but, in practice, the codifying nature of the underlying competence-oriented model has not interfaced easily with effective predictions to real-world performances (Harley et al., 1990; Skehan, 1998). Skehan (1998) states that the metacognitive processes of goal setting, assessment, planning and execution are important conceptually, but what is vital is to ask what factors influence these processes and how they influence them. He (1998, 2001) argues that to have competences which are unmobilized during performance does not confer much benefit. What is needed also is the capacity to translate these competences. At the most general level, the problem with Bachman's model is that underlying and generalized competences do not easily predict across different performance conditions or across different contexts. Moving from underlying constructs to actual language use has proved problematic in the CLA model. To address this issue, Skehan (1995; 1998), by drawing on insights from

[3] Bachman (1990) uses the term *test method* and *facets* to refer to what Bachman and Palmer (1996) call *task* and *characteristics.*

psycholinguistically-motivated research, proposes the construct of *ability for use*. He (2001) suggests that it is the goal of assessment techniques to devise methods of assessing the construct of *ability for use* as well as the underlying competences. He (1998) states that

> [e]ssentially, 'ability for use ' is seen to mediate between underlying competences and actual performance. The advantages of the construct are that it can incorporate insights from psycholinguistically-motivated research. It also enables a dual-coding perspective to be addressed. That is, rather than draw upon a generalized and stable underlying competence, the second language performer adjusts to performance conditions by trying to allocate attention in appropriate ways. This enables real-time communication to become more feasible and more degree of fluency to be achieved. When communicative pressure is not so heavy, when precision is important, or when task demands emphasize form, a syntactic mode assumes great importance. When pressure is greater and/or when effective communication is paramount, a lexical mode is used more. In this way, the learner, *essentially through a processing competence* [Emphasis in original], is able to handle the fluctuating communicative demands which operate, drawing on parallel coding systems which are available, and whose coexistence enables such flexibility to be used. (pp.168-169)

This is an interesting statement. Skehan develops the CLA model by explicating the way ability for use translates competence in language use. He not only lists the components of competence and ability for use, but also examines the way they interact with each other and their relationship with performance and context.

However, like Bachman (1990) and Bachman and Palmer (1996), Skehan also considers interaction from an individual-focused cognitive perspective. He has cognitive psychology and information-processing as his theoretical orientation, giving prominence to *attention*.

The exclusive cognitive view of interaction has been increasingly criticized in recent years in both SLA and language testing. For example, Atkinson (2002) comments Gass' (1998, p. 84) statement quoted above as " truly a pale reflection of the study of authentic human linguistic interaction"; what is missing is "any concern whatsoever for the dominant forms of interaction in the world, and for those interactions qua interactions" (p. 535).

In the field of language testing, McNamara (1996, P. 85-86) points out that a weakness of current models of communicative competence is that they focus too much on the individual candidate rather than the candidate in interaction and that the idea of performance as involving social interaction has so far featured only weakly in the work of theorists of L2 performance and researchers in language testing.

I will return to this in the next chapter. (See p. 49)

In the following section I will review research on interaction from a sociocultural perspective.

1.2 Sociocultural Perspective on Interaction

While mainstream SLA and language testing researchers have generally adopted an individual-focused cognitive perspective on interaction, a growing number of researchers have emphasized the social aspect of interaction from a sociocultural point of view.

One fundamental notion of sociocultural theory is that social interaction plays a fundamental role in the development of cognition including the development of

language. This notion is captured in the concept of the *zone of proximal development* (ZPD), which is defined by Vygotsky (1987, p. 86) as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers".

This notion provides a theoretical framework for studies on language learning as a mediated process. According to Lantolf (2000b), current research on mediated L2 learning continues to seek to better understand how L2 learning is mediated in the ZPD, especially in the form of peer scaffolding. He claims that "people working jointly are able to co-construct contexts in which expertise emerges as a feature of the group rather than residing in any given individual in the group" (p.84).

Swain and her colleagues, working from the sociocultural theoretical orientation, have carried out a series of insightful studies (Kowal and Swain, 1994; Swain, 1995; Swain and Lapkin, 1998; Swain, 2000; Swain, 2001; Swain and Lapkin, 2001) on the collaborative dialogues among students. Their findings have shown that dialogue among learners, wherein they are able to mediate each other, can be as effective a site for learning to happen as are instructional conversations between teachers and students. Thus, Swain (2001) states "dialogue is not 'enhancing learning' or leading to learning, it is learning" (p.288). The talk students produced provides teachers and researchers with opportunities to observe the underlying L2 learning process. As such, both language educators and SLA researchers are interested in interactions in the form of conversation.

On the other hand, from a sociocultural perspective, language is seen as one of the symbolic tools we use to "mediate and regulate our relationships with others and with ourselves and thus change the nature of these relationships" (Lantolf, 2000b, p. 1). And language use is both socially communicative act and a medium for the internal organization of experience. Therefore from the sociocultural point of view, language is both the result of and the tool for social interaction. It owes both its origin and its continued activation and use to social interaction.

From a sociocultural point of view, everything is co-constructed in social interaction. Thus, Jacoby and Ochs (1995) introduced the term *co-construction*, which they define as "the joint creation of a form, interpretation, stance, action, activity, identity, institution, skill, ideology, emotion, or other culturally meaningful reality" (p. 171).

The co-constructed view of interaction captures the dynamic feature of social interaction. When discussing contextual interaction, Douglas (2000, p. 43) points out that context is "dynamic, constantly changing as a result of negotiation between and among the interactants as they construct it, turn by turn" [4] (quoted from Chalhoub-Deville, 2003, p. 374). Similarly, Jacoby and Ochs (1995, p. 176) state that "interactional event is something that interactants are constantly monitoring, determining, and responding to as interaction unfolds". They suggest that "every

[4] This is a strong social perspective of context and is what Douglas (2000) means by *external context*. However, Douglas favors a focus on the *internal context* in language testing from an individual-focused cognitive perspective. Chalhoub-Deville (2003) comments that "the strong social interactional position advocated by Douglas in places seems to be mitigated or compromised by a need to accommodate conventional testing thinking and practices" (p. 375) and that "Douglas opts to discuss the L2 construct within the framework of the general CLA theory" (p.376).

interactional moment is a unique space for a response to which subsequent interaction will be further responsive" and that "interlocutors are processing and responding to the rich flow of unique interactional moments on-line" (p. 178).

When discussing the implications of co-construction, Jacoby and Ochs (1995) state that:

> [o]ne of the important implications for taking the position that everything is co-constructed through interaction is that it follows that there is a distributed responsibility among interlocutors for the creation of sequential coherence, identities, meaning, and events. This means that language, discourse, and their effects cannot be considered deterministically preordained by alleged "inherent" properties of linguistic structures, by assumed constructs of individual competence and so-called shared knowledge... (p.177)

To emphasize the social dimension of interaction from a sociocultural perspective is in no sense to deny its cognitive dimension, nor is it to simply examine the interaction between the social and the cognitive. Instead, when social and cognitive dimensions are included in one research paradigm, cognition is no longer considered as mere function of prebuilt cognitive structures of the human mind or a private activity that occurs exclusively "in the head". Cognitive activities are socially activated, socially mediated and socially motivated and are defined by social interactions.

As such, Chalhoub-Deville (2003) argues that:

> the ability components the language user brings to the situation or context interact with situational facets to change those facets as well as to be changed by them. The facets aspects of the context the language user attends to dynamically

> influence the ability features activated and vice versa. This perspective, in essence, maintains that ability and context features are intricately connected and it is difficult or impossible to disentangle them. (p. 372)

Therefore, language is both cognitive and social; "it is always mutually, simultaneously, and co-constitutively in the head and in the world" (Atkinson, 2002, p.538). In language studies researchers may have their social or cognitive orientation, but in actuality, the social and the cognitive can never be usefully separated.

1.3 Social interactional context

As mentioned above, a sociocultural approach to language development and language use emphasizes social interaction. However, social interaction itself is not the purpose of language use. People interact by means of language to achieve culturally appropriate goals in a given situation. It follows that social interaction is culturally and situationally specific and to interpret social interaction entails interpretation of cultural and situational contexts under which it takes place.

Research on context of culture and context of situation can be traced back to Malinowsky (1923; 1935) through Firth (1935; 1957) to Halliday's (1978; 1994) systemic functional linguistics (SFL).

In SFL, context of culture is referred to as *genre*. According to Martin (1984, p. 25) "a genre is a staged, goal-oriented, purposeful activity in which speakers engage as members of our culture". Genres are realized or encoded through language. This process of realizing genres in language is mediated through the realization of context

of situation, which is referred to as *register. Register* is described in terms of three variables: *field, mode* and *tenor.* These can be briefly glossed as:

-- field: what the language is being used to talk about;

-- mode: the role language is playing in the interaction; and

-- tenor: the role relationships between the interactants.

This relationship of *genre, register* and *language* is illustrated in Figure 2 by Eggins (1994).

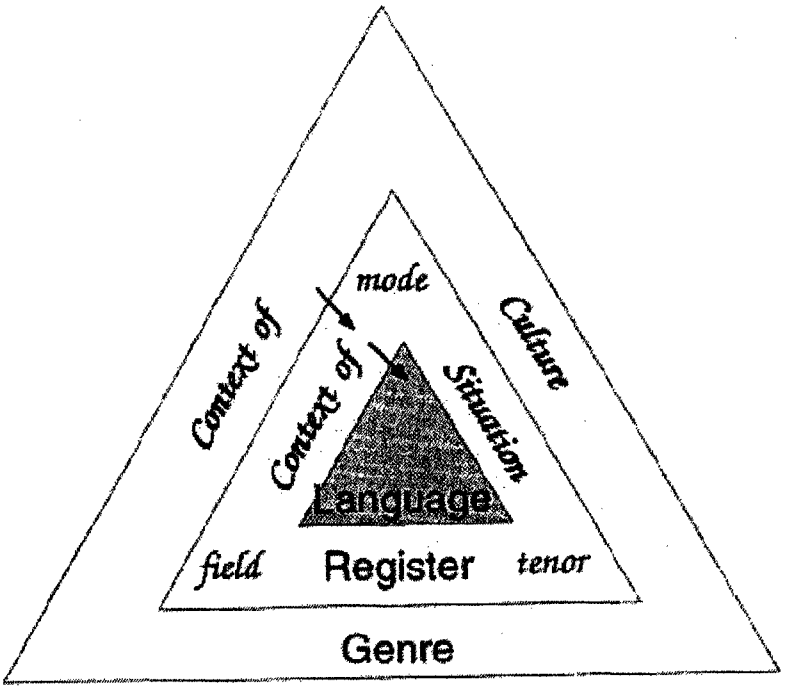

Figure 2 Genre and register in relation to language (Eggins, 1994, p. 34)

Figure 2 shows visually that all language is language in use. That is, language only makes sense under certain cultural and situational context.

It follows that if we take the sociocultural view, language ability is socially and culturally mediated knowledge and socially and culturally mediated cognitive processing capability. It originates in and is both activated by and motivated through social interaction and is situated in social interaction. From this perspective, then, language ability can never be separated from the language use context. As such, Chalhoub-Deville (2003) prefers an "ability -- in language user – in context" representation of the L2 construct over an "ability- in language user" representation, arguing that "ability and context features are intricately connected and it is difficult or impossible to disentangle them" (p.372).

It follows then that in language testing, if we are to assess the test taker's ability to use the target language, we have to take the social interactional context into account since language ability is always situated within the language use context. I argue in this thesis that language ability can only be assessed as it is related to the social interactional context, and that test developers must take this into consideration in designing language test tasks. This would include, for example, considerations of: 1) the social and cultural purpose and meaning of the task; 2) what the task is about; 3) the role of language in the task; and 4) the role and relationship between the interactants in the task. In understanding validity of a language test, these characteristics of the social interactional context will provide researchers with

valuable sources of evidence for test validation. I will discuss this point further in the following chapters.

Chapter 2

Incorporating a Sociocultural Perspective in Second Language Performance Assessment

In Chapter One I reviewed the individual-focused cognitive and the sociocultural perspectives on "interaction" and defined the social interactional context from a sociocultural perspective. In this chapter, I will discuss the implications for and applications in L2 performance assessment of an individual-focused cognitive approach and a sociocultural approach to "interaction". I will argue that it is both necessary and useful to incorporate a sociocultural perspective on "interaction" in L2 performance assessment. I will also argue that we need to take into account the influence on test performance of the social interactional context in task design and validation enquiry of task-based L2 performance assessment. This is an under researched area in language testing. At the end of this chapter, the purpose and the research question of my study will be introduced.

2.1 Second Language Performance Assessment

2.1.1 What Is Performance Assessment?

Fitzpatrick and Morris (1971) define a performance test by comparing it with the usual paper-and-pencil test, stating that a performance test is "one in which some criterion situation is simulated to a much greater degree" (p. 238. cited in McNamara

1997b). Norris *et al.* (1998) suggest that "virtually all language tests have some degree of performance included", so "it might be more appropriate to think of tests as more performance oriented or less performance oriented along a continuum from least direct and least real-world or authentic to most direct and most real-world or authentic." (p. 9) These definitions indicate that *directness* and *authenticity* are two distinguishing features of performance assessment.

In language testing, performance assessment is a term of long-standing discussion. In the 1950s, the Foreign Service Institute Oral Proficiency Interview (OPI) was introduced in the United States to satisfy "the practical requirements of personnel recruitment within the American government for officials who were to be placed in posts abroad requiring an active command of the local language" (McNamara, 1996, p. 30). "The communicative era in the 1970s generated a wave of criticism of the traditional non-communicative tests" (Shohamy, 1995, p. 188) and enthusiasm for more authentic, direct, communicative and performance-based tests. The 1980s witnessed development in sociolinguistics and its influences on language testing (e.g. Shohamy, 1983; Weir, 1988). It became clear that language performance is affected by a wide variety of factors and performance assessment became associated more with specific contexts and tasks (Wesche, 1992).

2.1.2 Task-based Language Performance Assessment

In recent years, increasing interest has been expressed in task-based L2 performance assessment by SLA researchers (e.g. Ellis, 2001), L2 instructors (e.g.

Samuda, 2001) and L2 testers (e.g. Long and Norris, 2000). An indication of this is the special issue of *Language Testing* (2002, 19/4.), which features a collection of empirical, practical and conceptual articles that address task-based language assessment.

Long and Norris (2000, p.60) define task-based language assessment as follows:

> Task-based language assessment takes the task itself as the fundamental unit of analysis motivating item selection, test instrument construction, and the rating of task performance. Task-based assessment does not simply utilize the *real-world* (original italics) task as a means for eliciting particular components of the language system which are then measured or evaluated; on the contrary, the construct of interest in task-based assessment is performance of the task itself. (quoted from Brown *et al.* 2000, p. 9)

Norris *et al.* (1998, p. 8) suggest three characteristics of task-based performance assessment as its working definition: (a) examinees must perform tasks, (b) the tasks should be as authentic as possible, and (c) success or failure in the outcome of the tasks must usually be rated by qualified judges.

On the one hand, task-based language performance assessment "focuses on the elicitation of performances of relevant tasks under conditions that approximate the real world as much as possible as well as on the evaluation of task performances according to real-world criteria" (Brown *et al.*, 2002, p. 11). On the other hand,it takes human cognitive factors and their interaction with context into account when interpreting task difficulty, which is a central issue in task-based language assessment. As such, task-based language performance assessment is more authentic and more direct on the one hand, and on the other hand, it is construct-driven (Messick, 1994),

well supported by theories and research and thus provides a more robust basis for generalization of performances, as stated by Brown *et al.* (2002, p. 11-12):

> [H]uman cognitive abilities interact with the contextual demands placed on performances by a range of variable task characteristics...For assessment purposes, then, if this relationship between cognitive abilities and task characteristics could be understood and modeled, a framework for sequencing tasks according to likely performance difficulty would be made possible. This framework would thus provide a basis for making generalizations from performance on one task to likely performances on tasks with related difficulty sources.

In general, three primary needs have motivated the interest in task-based L2 performance assessment:

(1) *Educational needs.* In the last two decades, L2 instruction has become more communicative with great emphasis placed on students' ability to use the L2 in real-life situations. Crookes and Gass (1993) observe that task-based instruction is one increasingly popular approach to communicative language learning. Pedagogical development has also suggested that tasks can be used as the basis for syllabus organization as well as the unit for classroom activities (Skehan, 2001). There is the need to align assessment with instruction in the form of shared characteristics such as learner-centredness, contextualization and authenticity. (e.g. Chalhoub-Deville, 2001). On the other hand, the use of tasks can provide formative information about the

progress that has been made and diagnostic information about the learner's strengths and weaknesses in terms of the ability for language use[5].

(2) *Social needs.* Since task-based language performance assessment prioritizes the simulation of real-world tasks and associated situational and interactional characteristics, and the inferences we want to make in it are about testees' abilities to accomplish particular tasks or task types in which target language communication is essential, this type of assessment can better inform program- or profession-related actions, such as certification, achievement, or qualification decisions, entrance or exit decisions, or mastery decisions, as to whether or not, or to what extent, students are able to accomplish the task according to real-world criterion elements and expectations. Task-based language performance assessment could provide test users information about the ability of the test taker to accomplish specific target communication tasks, ranging from the survival-related to the job-specific or academic, beyond simply displaying his/her linguistic knowledge (Norris *et al.*, 2002).

(3) *Theoretical needs.* As noted by Bachman and Cohen (1998, p. 22), many SLA researchers consider authentic language use to be the primary source of data for the investigation of language acquisition and hence place great value on "authentic" tasks for SLA research. Task-based performance assessment will better inform inferences made about language acquisition. Task-based language performance assessment can

[5] Here I don't agree with Bachman (2002), who claims that task-based approach to language performance assessment is "inappropriate for many educational purposes, such as diagnosis and assessing the achievement of learning objectives, which are typically stated in terms of areas of language ability to be learned" (p.461).

inform interpretations about the variable influences exerted by task features on examinees' cognitive processes and resulting performances, especially in terms of language production (e.g. Skehan, 1998; 2001; Wigglesworth, 2001).

2.1.3 Validity Issues in Second Language Performance Assessment

The potential of L2 performance assessment to satisfy the needs of different test users is attributable to a number of its attributes: communicative, functional, contextualized, authentic, direct and so on. It promises authentic and direct appraisals of students' abilities to respond to real-life language tasks and as a result, the test score can best inform interpretations and decisions test users make related to what the test taker is able to do with the L2 in future real-world situations. Therefore, it is highly face valid.

However, it is not acceptable to take the view that task-based performance assessments are in and of themselves valid due to their authentic or direct natures, as suggested by Huerta-Macías (1995). First, face validity is neither a necessary nor a sufficient condition for the validity of a test. Secondly, with regard to authenticity, there is the question of authentic to what criterion and for what purpose. In a broad sense, traditional forms of language assessment are not necessarily inauthentic, but can be said to be authentic representations of classroom work. On the other hand, if authenticity, taken as a distinctive feature of performance assessment, refers to the replication of or approximation to the conditions of real-world tasks beyond the testing context with as much fidelity as possible, idealized authentic performance

assessments are hard, if not impossible, to be realized. Jacoby and McNamara (1999, p. 223) point out "the inevitable simplification and dilution of the real-world task when simulated in performance test conditions may result in substantial loss of validity". Thirdly, with respect to directness, because of the influence on the performance of a number of factors and the interactions of these factors in performance assessment, the score which is used to assess the performance of the candidate must be seen as only partly a direct index of his/her actual performance.

Therefore, in task-based second language performance assessment, as in any other forms of testing, adequate justification must be provided for any interpretation of a given test score and for the consequences of the decisions that are made on the basis of the test score. Brown et al. (2002, p.5) argue that the issue of validity must be dealt with for performance assessments just as it is for any other forms of testing "– in an open, honest, clear, demonstrable, and convincing way", especially when the test score is used to make high-stakes decisions about students.

Messick (1989, p. 13) describes *validity* as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores". He uses the term *construct validation* as an overarching term to refer to all aspects of this enterprise. With regard to the validation of performance assessment, Messick (1994) argues "performance assessments must be evaluated by the same validity criteria, both evidential and consequential, as are other assessments" (p.13). He further suggests that:

> [a]uthenticity and directness map, respectively, into two familiar tenets of construct validity, that is, minimal construct underrepresentation and minimal construct-irrelevant variance. Consequently, the issue, as always in test validation, becomes the nature of the evidence accrued to counter the two major threats to construct validity, namely, construct underrepresentation (which jeopardizes authenticity) and construct-irrelevant variance (which jeopardizes directness)". (p.14)

and that:

> [w]here possible, a construct-driven rather than a task-driven approach to performance assessments should be adopted because the meaning of the construct guides the selection or construction of relevant tasks as well as the rational development of scoring criteria and rubrics. Focusing on construct also makes salient the issues of construct underrepresentation and construct-irrelevant variance, which are the two main threats to validity. (p. 22)

Following Messick (1989), Haertel (1992) proposes the relevance of construct validity to both task-centred and construct-centred approaches:

> Construct validity embraces all the evidential basis of test interpretation, including content- and criterion-related lines of validity evidence. It is relevant to performance measurement even if the intended test interpretations do not appear to involve psychological constructs. Test interpretations must be qualified to the extent that the test fails to sample some parts of the performance domain it is supposed to represent (construct underrepresentation) or depends on knowledge or skills from outside of that domain (construct-irrelevant variance) (p.987)

In language performance tests (in SLA research as well, where tests are viewed as SLA elicitation devices) we are not interested in the test taker's performance in the testing context, or the *test,* in McNamara's (1997b) term, for its own sake. Instead, we are more interested in his/her subsequent language behavior outside of the testing setting, or the *criterion,* in McNamara's term. Because this behavior cannot be observed directly, we can only make inferences about it from a limited sample of

performance of the test taker that is available from the test setting. To make these inferences, we need a language performance theory, which helps us to articulate the theoretical rationale for the inferences we made from the learner's or test taker's performance on the test about his/her subsequent target language behaviour. McNamara (1996) points out that a performance theory permits "the necessary clarity, specificity and explicitness in stating the grounds for inferences about candidates' abilities made on the basis of test performance, thereby also facilitating the empirical investigation of the validity of these inferences" (p.49). This can be illustrated by Figure 3.

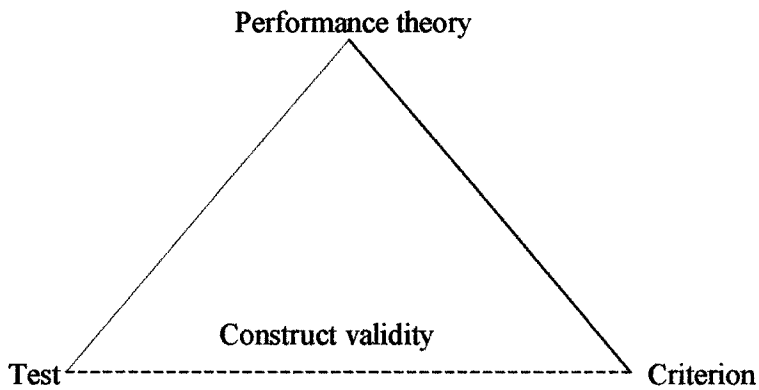

Figure 3 The role of performance theory in performance assessment

However, due to the complexity of language use, to develop a theory on language performance presents theoretical challenges. As such, McNamara (1996) sees the work of modeling performance in language performance assessment as opening Pandora's box ---- a view I will discuss in relation to the findings of my study at the end of this thesis. In the next section, however, I will discuss the implications for performance assessment of the individual-focused cognitive and the sociocultural perspectives on "interaction" discussed in Chapter One.

2.2 Interaction in Second Language Performance Assessment

2.2.1 Implications for and Applications in Performance Assessment of an Individual-focused Cognitive Approach to Interaction

As mentioned above, one distinguishing feature of performance assessment is *authenticity.* In considering authenticity, Bachman (1990) preferred what he refers to as "the interactive ability (IA) approach", which, he asserts, emphasizes "the interaction between the language user, the context and the discourse" (p.302) to a real life approach. In discussing the relationship between performance, abilities and context, Bachman goes on to state that:

> [t]est performance is interpreted as an indication of the extent to which the test taker possesses various communicative language abilities, and there is a clear distinction in this [IA] approach between the abilities to be measured, on the one hand, and the performance we observe and the context in which observations take place, on the other (pp. 302-303).

Bachman (1990) stresses the implications of the IA approach for validation enquiry of language performance assessment, arguing that:

> [i]t [the IA approach] must incorporate a model of abilities and test method facets for both test development and interpretation of test results. It requires a complex process of theory falsification -- construct validation – to provide an evidential basis for validity. But the pay-off is well worth the effort, since this approach provides the only means, in my opinion, of making inferences about the abilities we think we are measuring. (p.331)

In the IA approach to performance assessment, tasks are utilized as methods for eliciting language performance. The performance itself, however, is of less interest than what it reveals of underlying ability. In Messick's (1994) words, performance is the *vehicle* rather than *target* of assessment. Bachman (1990, p. 15) argues for research and development of language tests guided by theoretical frameworks of communicative ability and test method facets. His CLA model discussed in Chapter 1, which has been influential in the language testing field throughout the 1990s and is still considered the one that represents the state-of-the-art of construct definition in language testing (Alderson and Banerjee, 2002), is just such an attempt.

Bachman (1990) and Bachman and Palmer (1996) reiterate throughout their two influential books that the primary interest in using language tests is to make inferences about one or more components of an individual's communicative language ability. The CLA model seems to provide a desirable, feasible and practical framework for the development of language testing to achieve this goal. It has been applied to test development, selection and analysis (See Bachman and Palmer, 1996).

Underlying this advantage is the implicit assumption that such a construct as CLA, whose components and structure can be identified and modeled, does exist in the language user's or testee's mind and this construct can be tested independently of test tasks. If performance in a range of situations is thought to depend on CLA, "the scope to predict performance will accordingly be greater, and broader-based samplings of the abilities should generate more robust predictions, so prediction of

actual performance and even generalizations across contexts can be achieved more effectively" (Skehan, 1998, p. 163).

However, as noted by a number of researchers (e.g. Kunnan, 1998; Douglas, 2000; Chalhoub-Deville, 2003), up to the present, consensus is absent regarding the nature of the components underlying the L2 construct and the manner in which they interact.

To Bachman (1990), language use contexts, which are dealt with as test method facets, are important to the extent that they help draw out those intended abilities through the operation of strategic competence, a very general cognitive capacity. This view is echoed by Douglas (2000), who favors a focus on the *internal context*, which he identifies "as a cognitive construct created by language users for the interpretation and production of language" (pp. 45--46), stating that "what really counts in thinking about context is the internal interpretation of it by the participants themselves" (p.89).

Since, according to Bachman (1990), CLA is of primary interest in language testing, the effects of task characteristics on test performance are considered as sources of measurement error in interpreting test scores, and should be controlled as systematic factors or should be minimized as random factors. To do this, Bachman (1990, Chapter 6) introduces a number of quantitative techniques to examine test reliability.

However, Bachman (1990) points out that the difficulty of distinguishing the ability that we want to measure from the way in which we measure it is the fundamental dilemma of language testing. In order to permit generalizations on the basis of transferable abilities, Bachman maintains the separation of the language use

situation and the abilities underlying performance, focusing on the abilities the individual possesses. But the restriction of language abilities seems to be a vain attempt and leads to unsatisfactory or even unfortunate assessment outcomes (McNamara, 1997a).

To illustrate how a variety of different factors impact upon the performance and the score assigned to the performance of the candidate, Skehan (1998), following Kenyon (1992) and McNamara (1996), proposes the model shown in Figure 4.

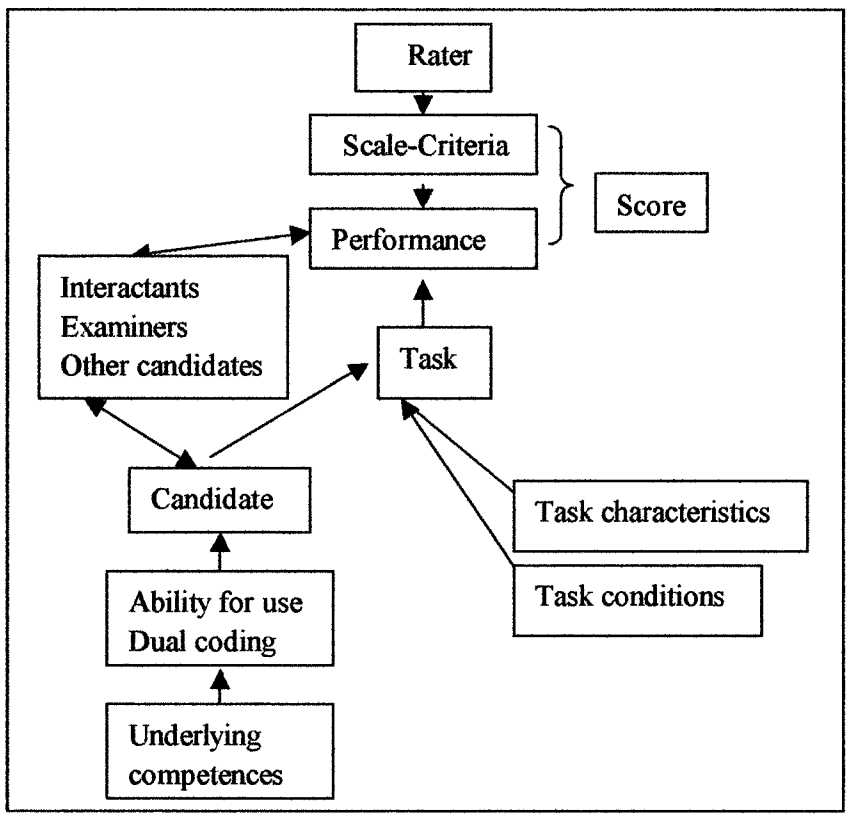

Figure 4 A model of oral test performance (Skehan, 1998, p.172)

Skehan's (1998) model provides a framework to consider a variety of potential

factors that are relevant to consider in interpreting scores of task-based language performance assessment. Since in this model what is of primary interest is the construct of *ability for use*, the influences on performance and on test scores of rater characteristics and the qualities of the rating scales and the interaction of these two factors and the influences of interactants are all considered as construct-irrelevant variances that are to be minimized or controlled.

On the other hand, since Skehan argues for the integration of task characteristics and cognitive abilities in defining the construct of *ability for use* from a cognitive processing perspective, influences of tasks on performance are not regarded in his model as a technical problem at the level of format effect as would be in other types of tests. Rather, they become "*the* central problem in testing" (Skehan, 1998, p.291. Emphasis in original).

Skehan (1998) emphasizes the influence of tasks on performance, suggesting that "tasks will influence difficulty and also have selective influences on different aspects of performance. By concentrating upon task design features, it will be possible to base generalizations on task characteristics that are shared, or not, across the different contexts" (p. 171). As such, he concludes:

- *tasks and processing need to be understood if results are to be interpreted;*
- *tasks and processing provide guidance for the sampling that is necessary to enable generalization: we need to know what sorts of performances people are capable of* (p. 173).

Therefore, Skehan (1998) seems to choose to solve the dilemma of language testing by emphasizing that human cognitive abilities interact with the contextual demands placed on performances by a range of variable task characteristics

However, due to his cognitive theoretical orientation, the focus of Skehan's (1998) research on task is to abstract systematic influences of different variables associated with tasks on cognitive abilities and language performance in terms of task difficulty. As clearly shown in Figure 4, the interactant factor is not integrated in task. By drawing on psycholinguistic research on influences of task features on performance, Skehan (1998) proposes three sets of features that he hypothesizes affect performance on tasks: code complexity; cognitive complexity; and communicative stress.

Skehan's (1998) task-based approach to L2 performance assessment has been employed by a group of researchers at the University of Hawaii at Manoa (Norris *et al.,* 1998; Brown *et al.,* 2002; Norris *et al.,* 2002), who were engaged in a research project to develop prototype task-based L2 performance tests under the practical testing conditions at universities in the USA. They propose that task characteristics and human ability requirements play an interactive role and should be considered as a set when classifying tasks according to difficulty. Abilities that are required for successful task completion are often inherently tied to non-ability characteristics of the task. Descriptors based on these two strands, taken together, should effectively approximate an item specification, each strand contributing crucial information to the performance parameters inherent in a given assessment.

Based on Skehan's (1998) work, the Hawaii group (Norris *et al.,* 1998; Brown *et*

al., 2002) devised a system for coding task characteristics of the particular set of tasks they had identified for their study. Table 1 shows their task characteristic matrix.

Table 1: Assessment of language performance revised task components and characteristics matrix (Brown *et al.*, 2002)

Components	Characteristics	
Code demand	Range +/-	# input/output sources +/-
Cognitive operations	Input/output organized +/-	Input/output available +/-
Communicative adaptation	Mode +/-	Response level +/-

They developed two different types of rating scales: task-dependent and task-independent. The task-dependent scale had been developed to assess the extent to which an individual examinee could accomplish a given target task by real-world criteria for the specific task. This type of scale was developed by a team of informants knowledgeable about both the target tasks and the language issues faced by the target population of the study: international students at American universities. The task-independent scale had been developed, using the processing sources for difficulty estimations, to estimate a student's general level of ability to deal with tasks that involved a range of processing abilities. It includes three scales, one each representing the three cognitive factors that were ostensibly engaged in varying degrees and combinations across the range of the performance tasks developed for the research project.

This approach has many merits. It makes it possible to develop a manageable, relatively easy to understand, and systematic way of examining performances for a range of similar tasks in terms of the variables involved in those tasks. Based on the research into how human cognitive abilities interact with the contextual demands placed on performance by a range of variables of task characteristics, this approach also makes possible a framework for sequencing tasks according to likely performance difficulty. This framework would further provide a basis for making generalizations from performance on one task to likely performances on tasks with related difficulty sources.

The findings of the Hawaii group's research show that a student's success on particular test tasks, as assessed by the task-dependent rating scale, is closely related to his/her general abilities to deal with the processing demands posed by all the full-length task-based performance test, as estimated by the task-independent rating scale. Their findings also suggest that the overall scores, based on average task-dependent ratings or on holistic task-independent ratings, may effectively distinguish among students according to their general abilities to perform the range of tasks found on the test of the research project.

The Hawaii group's work is "the most fully conceptualized, operationalized and researched exemplification of" the task-based approach to language performance assessment (Bachman, 2002, p. 454). Their research has proved the possibility and feasibility of Skehan's approach in criterion-referenced testing context for the purpose of gradation of task difficulty in order to generalize performance on the test to future

real-world task performances.

However, the findings of the Hawaii group's research do not support the use of their task difficulty framework as an inferential basis for generalizing from examinee's performances on specific tasks to likely abilities with related tasks, though some degree of relationship was evidenced between the cognitive factors and task-dependent performance evaluations.

To address this, Norris *et al.* (2002) state that:

> a central focus for research on all types of language performance assessment – including assessment which informs the construct-centred as well as the task-centred ends of the inferential continuum – should be to better understand how tasks are actually accomplished (*in both cognitive and "real-world" terms*) [added emphasis] and what makes a given task more or less "difficult" for different examinees. Without a better understanding of performance along these lines, the validity of inferences about language learners' abilities (of whatever sort) will remain in question. (pp.415-416)

These statements are thought provoking and challenging. Research from a cognitive perspective into task characteristics and task conditions and their interactions with human cognitive abilities in terms of task difficulty (which is far from being complete) will provide necessary but not sufficient information for the design of task-based L2 performance assessment and for the investigation of its validity. What is also needed is research into how tasks are actually accomplished in "real-world" terms, how performance is co-constructed through interaction between or among interactants, and the examinee's perceptions of task difficulty.

If the way tasks are actually accomplished is to be understood in *both cognitive*

and "real-world" terms, a sociocultural approach is most relevant. This approach will give prominence to the social interactional feature of performance and will see a language performance assessment as an essentially social activity.

2.2.2 Implications for and Applications in Performance Assessment of a Sociocultural Approach to Interaction

From a sociocultural point of view, a performance test will be seen as a social activity. This implies that both the elicitation of performance (which involves task development and test taking), and the interpretation of performance (which involves development of rating scales and scoring) will be seen as inherently social acts realized through interactions.

Jacoby and McNamara (1999) examined the validity of the Occupational English Test (OET) in Australia by comparing its assessment criteria with criteria a group of research physicists appeared to be orienting to on a project in America studying an indigenous assessment activity among physicists collaboratively practicing and commenting on one another's presentations in their rehearsals of upcoming conference presentations. Jacoby and McNamara (1999) explain that:

> [f]or the physicists, indigenous assessment is a here-and-now interactional problem-solving activity which aims to help presenters turn what colleagues and mentors judge to be less-than-competent performances into more competent ones suitable for an actual future public performance before an audience of professional peers. (p. 235)

Jacoby and McNamara (1999) found that the OET criteria and the indigenous assessment criteria were different in kind with regard to the relevance of the criteria to the communication task. Their conclusion was that the OET assesses special-purpose communicative performance on an exclusively linguistic plane with language-focused assessment criteria and that the exclusion of the real-world criteria may account for the discrepancy between the test performance and reported real-world performance and thus may jeopardize the validity of the test.

Jacoby and McNamara's (1999) conclusion suggests that in L2 performance assessment, where inferences are to be made about the candidate's performance in the real world on the basis of his/her performance on the test, theory-informed evaluation criteria might also "underrepresent" the construct, which might jeopardize the validity of the inferences drawn from the test. This issue was echoed by Turner (2000) reviewed below.

Upshur and Turner (1999) observe that differences exist between SLA and language testing approaches to the study of L2 performance in that language testing research examines systematic effects on performance test scores "without generally reference to discourse" (p. 82) while SLA approaches investigate effects of tasks on discourse generally without concern for test scores. Upshur and Turner reported some incidental findings from a test-development project in Montreal, Quebec. By using empirically derived rating scales as features of discourse, they were able to show that tasks influence discourse and discourse affects test scores in performance tests. Their findings suggest that interactions between task, examinee, discourse and rater bear on

the likelihood of a candidate getting a particular score. Upshur and Turner propose a performance test taking and scoring model shown in Figure 5, where they highlight the relationship between task and rater to indicate their finding "that judges seem to adapt their strategies to task demands"[6] (p. 106). They also suggest that other relations[7] should be included to develop a more sophisticated model of performance testing, which appears to be an important challenge for both language testing and SLA.

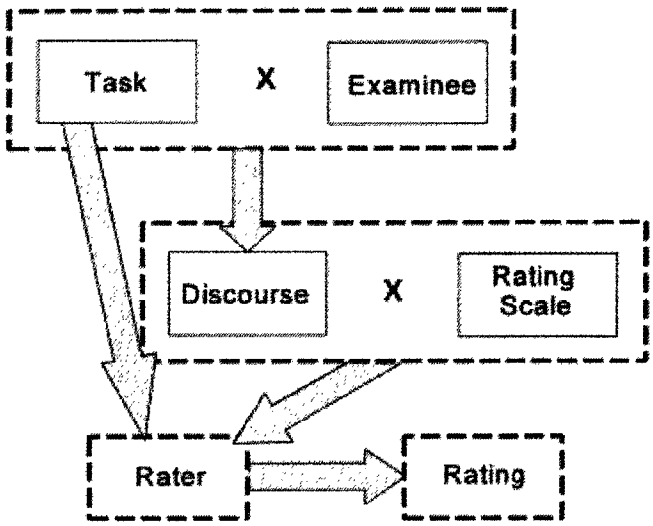

Figure 5 Performance test taking and scoring

(Upshur and Turner, 1999, p.106)

In another study[8], Turner (2000) reported how in the process of constructing

[6] Two tasks were used in Upshur and Turner's (1999) study: the Story Retell (SR) task and the Audio-Pal (AP) task. In the SR task, the students watched a short video, drew a picture to help them remember the story. Then they were required to retell the story using their pictures on a tape recorder. In the AP task, the students were required to compose a letter in the form of a tape recording to an exchange student. Findings of this study showed that "raters were unexpectedly severe or lenient in scoring Story Retell" (p. 103). Upshur and Turner explained that this could be accounted for by different rating strategies the raters used.

[7] Upshur and Turner (1999) also examined relations between raters (who were test developers as well in this study), tasks and the empirically derived rating scales.

[8] This study was conducted "within a Ministry of Education of Quebec project whose objective was to develop empirically based rating scales for secondary-level ESL provincial exams" (Turner, 2000, p.555).

empirical rating scales, actions of participants (*scale development team*) and/or their use of the data *sample* influenced the criteria constructed for the rating scale and the final ratings. Through a qualitative analysis, Turner showed effects on ratings of test methods in terms of empirically derived scales. When discussing implications of her study, Turner states that:

> [o]ne final implication of this study is the contribution the findings can make to the ongoing discussion concerning the basis for rating scales in performance testing. The notion echoed in the literature is that rating scales should be informed by theory and language models.... It appears, however, that empirically derived scales can also inform theory. I see this development as a two-way process, theory informing rating scales and vice versa (grounded theory) (p. 577).

While the studies of Jacoby and McNamara (1999), Upshur and Turner (1999) and Turner (2000) highlight the influence of social interactions on the construction of rating scales and ratings, recent work of Swain and Lapkin (Swain and Lapkin, 1998; Swain, 2000, Swain and Lapkin, 2001; Swain, 2001) has focused on influence on language performance of the more immediate social interaction: collaborative dialogues among students. Their research on collaborative dialogues among middle-school French immersion students within the sociocultural theoretical orientation has provided new ways to examine dialogue from a sociocultural perspective as an important validation source. Swain (2001) suggests that "because cognitive and strategic processes are made visible in dialogue, then studying dialogue will provide us with evidence of how participants in group interaction approach and process the task demands. If understanding these strategies and processes is important

to an understanding of the construct being measured, then the dialogue among participants will be an important source of validation evidence" (p.281).

Swain's (2001) suggestion highlights the usefulness of examining students' performance in relation to the social interactional context in both SLA and language testing. Also, in Chapter One, I argued that in language testing, if we are to assess the test taker's ability to use the target language we have to take into account the social interactional context since language ability is always situated within the language use context. In the next section, I will review three interrelated directions of study concerning test validity in relation to the social interactional context: the effect on performance of a number of variables associated with the interlocutor in face-to-face communication; research on the effects of face-to-face and tape-mediated tests[9] of speaking; and, research on oral proficiency interview (OPI) vs. small group tests.

2.3 Research Related to the Influence on Test Performance of the Social Interactional Context

Researchers have investigated the effect on test performance of a number of variables associated with the interlocutor in face-to-face oral language test, for example, age (O'Sullivan, 1995; O'Sullivan and Porter, 1995; Buckingham, 1997); interaction style (Porter and Shen, 1991); language level (Iwashita, 1997); personality (Porter, 1991a; Berry, 1997); gender (Locke, 1984; Porter 1991a; 1991b; Porter and

[9] Some researchers (e.g. Stansfield, et al. 1990; Shohamy, 1994) distinguish *direct* and *semi-direct*. I choose to use *face-to-face* and *tape-mediated* because I feel that the terms *direct* and *semi-direct* are confusing.

Shen, 1991; O'Sullivan and Porter, 1996; Berry, 1997; Buckingham, 1997); status (Porter and Shen, 1991); power (Van Lier, 1989) and acquaintanceship (Porter 1991a; O'Sullivan, 2002), etc.

Generally, a major concern of these studies has been to investigate systematic effects on the test score and ways to control them. The interactant (the examiner or other candidate) is considered as an additional source of measurement error, which jeopardizes test validity in terms of construct-irrelevant variance.

Most of the research on the effects of face-to-face speaking language tests and tape-mediated tests has centered on the concurrent validity of the tape-mediated test by examining the extent to which it is a valid and acceptable alternative to a face-to-face test (e.g. Shohamy and Stansfield,1991; Stansfield, Kenyon, Paiva., Doyle, Ulsh, and Cowles, 1990; Stansfield,1991).

However, as Shohamy (1982) points out, high correlations between face-to-face and tape-mediated tests provide necessary, but not sufficient, evidence for the appropriateness of test substitution. More convincing evidence for validation should be obtained by examining the specific language samples that two tests elicit and determining whether the tests are, in fact, the same.

To provide more evidence for the validity of the tape-mediated test, Shohamy (1994) collected data from different perspectives using both quantitative and qualitative methods. She examined the functions and topics that were elicited in the test tasks and the specific language samples obtained from the elicitation tasks, focusing on linguistic, strategic and discursive features of the language samples. Her

conclusion was that the elicitation method----face-to-face or tape-mediated interviews----affects the language produced by the testee in terms of speech functions and topics and communicative strategies. She also concluded that "context [in which the language is elicited] alone seems to be more powerful than the elicitation tasks themselves" and that "the context of the test, either 'face-to-face' or 'tape-mediated', can affect or even dictate the type of language that is produced" (p. 118).

Since the tasks of the two tests Shohamy (1994) examined were not identical, some of the discrepancies she reported may be traced to the task type and specific topic rather than the test modalities themselves. However, given the inseparability of language ability and the social interactional context discussed above, I believe that Shohamy's (1994) conclusion is meaningful. In terms of the social interactional context from an SFL perspective discussed in Chapter 1, face-to-face interaction and tape-mediated interaction involve different context of culture and different context of situation. This issue presents a critical challenge to the validity of the tape-mediated test.

Van Lier (1989) has challenged the underlying assumption of the validity of the OPI, that it measures speaking ability in the context of a conversation, by simply asking "Are OPIs examples of conversational language use?". To find the answer to this question, he looked at the interview from the inside to understand what OPIs are and what the participants in them do. His conclusion was that OPIs can and may be designed to elicit conversational language use but frequently is not and that conversation is the best vehicle for the evaluation of oral proficiency. Since van Lier,

a number of researchers (e.g. Perrett, 1990; Lazaraton, 1992; Johnson and Tyler, 1998) have addressed the issue of validation in oral language tests, especially the validity of language proficiency interview (LPI) or OPI, using more qualitative approaches, particularly through discourse analysis. Their findings suggest that "the one-to-one oral interview generates a special genre of language different from normal conversational speech" (Fulcher, 1996, p. 26) and thus it "cannot be considered a valid example of a typical, real-life conversation" (Johnson and Tyler, 1998, p. 28). (See Fulcher, 1996 for a review, and also Young and He, 1998 for recent studies on this topic)

Largely because of dissatisfaction with the OPI or LPI as a solo means of assessing oral proficiency, researchers and test developers have shown interest in search for other types of task, e.g. small group test[10]. There have been reports of successful use of group testing with school students in Israel (Reves, 1980; Shohamy, Reves and Bejarano, 1986; Reves, 1991) and Zambia (Hilsdon, 1991) and with university students in Hong Kong (Morrison and Lee, 1985). Berkoff (1985, p.95) argues that using groups of students overcomes the problems of "artificial conversation" between a "distant examiner" and a "nervous examinee". However, Fulcher (1996, 24) points out that the reports of successful use, the claim of a reduction in "test-type" language and reduced anxiety in the literature are not well supported with empirical evidence. A major reason for the lack of empirical studies

[10] Swain (2001) summarizes a number of reasons for small group test.

on group testing is that the group discussion task has not been usually retained in the test battery because of the logistical problem.

2.4 The Research Gap

From a sociocultural perspective, if the purpose of a language test is to make inferences about the candidate' abilities to use language on the basis of test performance, it is both necessary and useful to examine the dynamic and mutual influences of performance and the social interactional context since language ability and context features are intricately connected. A sociocultural approach will see the social interactional context as a part of the construct to be assessed, whose absence may be a threat to test validity in terms of construct underrepresentation.

Largely influenced by the individual-focused cognitive tradition in language testing, previous research relevant to the influence on test performace has generally seen the social interactional context as something fixed and stable and as sources of measurement error, which are to be voided or controlled.

As such, McNamara (1996, P. 85-86) points out that a weakness of current models of communicative competence is that they focus too much on the individual candidate rather than the candidate in interaction and that the idea of performance as involving social interaction has so far featured only weakly in the work of theorists of L2 performance and researchers in language testing.

2.5 The Study

The current study was designed to explore the usefulness and feasibility of incorporating a sociocultural approach to test development and validation enquiry in a small group oral language test in the EAP context by examining the influence on test performance of the social interactional context. To achieve this research purpose, the following research question was addressed:

How does the social interactional context influence test taker performance?

To explore the influence of the social interactional context on test performance, a sociocultural view was taken to examine the variance in the test performance accounted for by the social interactional context, to investigate the influence of the social interactional context on task difficulty, and to explore ways that analysis of small group oral performance from a sociocultural perspective can inform EAP task-based performance assessment in terms of validation enquiry.

Chapter 3 Methodology

As mentioned at the end of the last chapter, the purpose of this research was to explore the usefulness and feasibility of incorporating a sociocultural approach to test development and validation enquiry in a small group oral language test in the EAP context by examining the influence on test performance of the social interactional context. To achieve this purpose and to address the research question at the end of Chapter 2, the current study was conducted in an EAP program at Carleton University, Ottawa. Two tasks from the Oral Language Test (OLT) of the Canadian Academic English Language (CAEL) Assessment were used and parallel task versions were developed by changing test method facets. Both quantitative and qualitative methods were employed to analyze the data.

3.1 Background

The CAEL Assessment is a standardized test of English in use for academic purposes. It is designed to describe the level of English language of test takers planning to study in English-medium colleges and universities in Canada. Since its first use in 1989, test takers in many parts of the world have taken the CAEL Assessment as part of the process of admission to universities and colleges in Canada, Europe and the United States.

The CAEL Assessment is administered in three stages. Stage One consists of registration, self-assessment and a personal essay. Stage Two consists of a

tape-mediated test of oral language proficiency in an academic context. Stage Three consists of written responses to academic readings, an academic lecture and an essay prompt which asks test takers to agree or disagree with a statement. All of the evidence collected about a test taker's ability to use English for academic purposes at these stages are reviewed by a placement team. An overall result is assigned on criteria which reflect the standards of proficiency accepted at Canadian universities who use the CAEL Assessment.

According to Fox (1999; 2000), the CAEL Assessment OLT is a task-based tape-mediated oral language test of spoken English in use for academic purposes. It consists of five tasks which represent ways in which students talk about their academic work within colleges and universities. These tasks include:

Task 1 (2 minutes): making short presentations;

Task 2 (5 minutes): relaying information;

Task 3 (5 minutes): explaining choices;

Task 4 (5 minutes): summarizing main points

Task 5 (8 minutes): listening and responding to group discussion.

The scoring for the OLT is undertaken by trained raters who listen to the test takers' recorded responses to the five tasks. Points are assigned to each task on an analytic scoring rubric on specific features of each task. The overall oral performance is then assessed holistically. The analytic points and the holistic point are put together

to make the raw scores, which are converted to criterion-related band scores that range from 10 to 90. The score criteria for bands 3-9 are listed in Appendix A. (For more detailed information about the CAEL Assessment, see Fox, 1999; 2000.)

"The CAEL Assessment differs from other standardized tests in that its specific purpose is to identify test-takers who have the ability to use EAP in university classrooms" (Fox, 2004, p. 438). What is interesting is that at Carleton University, the CAEL Assessment is also used as a placement test in EAP support programs. Results of the test place the students into one of three main levels:

(1) Intensive ESL course level: for students whose English level is not high enough to begin credit courses;

(1) Credit ESLA course level: for students who are admitted to a degree program but whose language skills require some additional support;

(2) English as a second language requirement has been satisfied. No ESL is required.

There is some sociocultural belief underlying the development and use of the CAEL Assessment. First, it emphasizes ability to use the English language in particular context ---- university academic classroom. Second, it sees university students as practicing members of certain academic communities and their academic development as "a specialized form of socialization through which culturally specific knowledge, language, discourse, cognition, skills, and practice are transmitted and developed" (Jacoby and McNamara, 1999, p. 224). Therefore, it is to the advantage of

the development of the students' EAP ability that they are admitted to an academic program even though their language is below the required level for full-time study. Thirdly, since the process of socialization is achieved through language, EAP ability is essential for academic success in English-mediated universities in that it allows the student to participate in specific academic practices. Therefore, to begin a program the candidate has to be able to attain an EAP threshold level (Fox, 2004). The CAEL Assessment serves as a "gatekeeper" that allows or denies access to a program in universities. It assesses the candidate's discipline-general capabilities (Fox, 2001) of participating in specific academic practices by means of language at the initial point of his/her contact with the new academic community (Fox, 2004). Finally, if a candidate whose language is below the required level for full-time study is allowed to enter an academic program as a result of the CAEL Assessment, it is advantageous to the student in terms of his/her language development and his/her academic success that some kind of EAP support is provided. There seems to be some relationship between English level, EAP support and discipline-specific work. This relationship is best articulated by Fox (2004):

> [T]he lower the level of English, the greater the amount of EAP support that is required and the smaller the number of discipline-specific courses that are allowed. Conversely, the higher the level of English, the greater the number of discipline-specific courses that are allowed, and the lower the amount of EAP support that is required (p. 440).

At Carleton University, the threshold level for the admission of an academic program is Band 40 of the CAEL Assessment. The candidate is not required to take an

EAP course if he/she is able to attain CAEL Band 70 or above. Candidates whose band scores are between 40 and 70 are placed in three different EAP support programs, where different EAP courses are offered and the students are allowed to take a limited number of discipline-specific courses. This is illustrated in Figure 6. The students have to receive a B- or better in the current EAP course to enter the next higher EAP program.

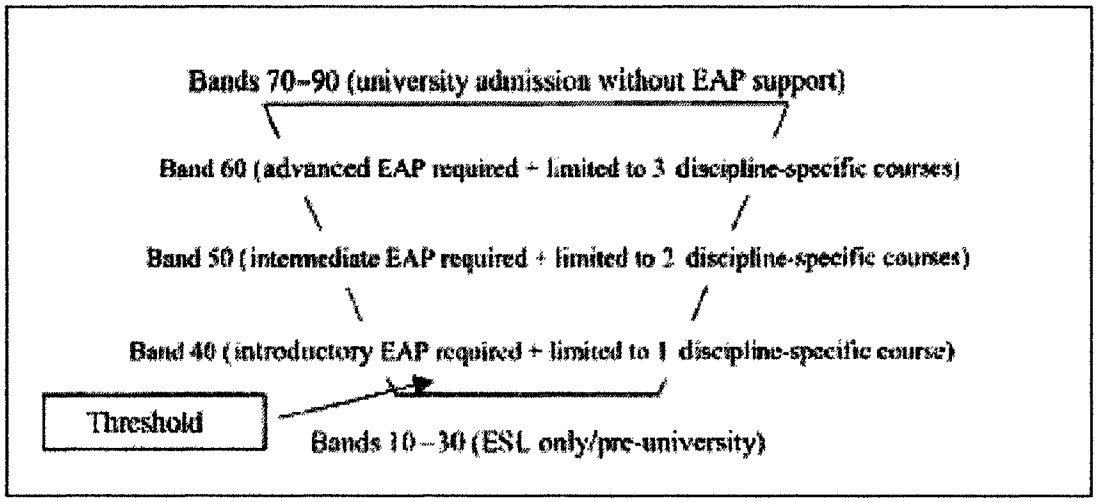

Figure 6. The relationship between test performance (band score) and program placement in one university program (Fox, 2004, p. 440)

3.2 Participants[11]

Students:

23 students from an introductory EAP program (see Figure 6) at Carleton University participated in this study. These students had taken the CAEL Assessment

[11] Before the study was conducted, an ethic proposal was submitted by the researcher to Carleton University Research Ethics Committee. The proposal was reviewed and approved by the committee, so the current study was appropriate for human participants.

before they registered in this program. They achieved an Overall Band Score of 40 (approximately 190 on computer-based TOEFL or 520 on the paper-based)[12] with Band Scores of 30 or above in all of the Reading, Listening, Writing and Speaking Band Scores. (The descriptor of Band 40 of the CAEL Assessment is provided in Appendix A.)

At Carleton University, the introductory EAP course is a one-term 9-hours/week 1.0 credit course. The students on the introductory EAP program are also permitted to register in one 0.5 credit course.

Of the 23 students, 18 were from Mainland China, 1 from Taiwan, 1 from Kuwait, 1 from Vietnam and 2 from Japan. They had studied English as a foreign language in their native countries for 5-8 years and had studied in Canada for 1-3 years. Their ages ranged from 20 to 28 years. 12 of the participants were male students and the other 11 were female.

EAP teacher

The EAP teacher in the study had taught the 23 students for one term. She had completed her MA courses and was working on her thesis at the time of the study. She was both an experienced EAP teacher and an experienced CAEL Assessment rater.

[12] This approximate TOEFL equivalence for a CAEL Assessment test result is provided in the *Test Score and Users' Guide* of the CAEL Assessment to meet the needs of score users. It should be noted that it is a guide to interpretation only. Because of these two tests operationalize very different constructs of the English language, a precise comparison of them is both impossible and inappropriate (Fox, 2002).

Raters

The EAP teacher rated the students' performance on the small group discussion in the classroom. Recordings of the students performing the tasks in the computer lab were rated by the other two raters. One of these raters with a doctoral degree had taught in the EAP program and was one of the originally developers of the test and an experienced rater. She had been engaged in research related to the CAEL Assessment for a number of years. The other rater was the researcher, an MA student who had taught English as a foreign language at a Chinese university for over ten years. He was also an experienced and trained rater.

Pearson correlation co-efficients were computed between the two sets of marks assigned by the two raters to the students performing the tasks in the computer lab. The correlations ranged from .84 to .94. Given the consistently high correlations across the tests, the researcher was confident of the interrater reliability of this study.

3.3 Instrumentation

Tasks

Task 5 of the CAEL assessment OLT (listening and responding to group discussion) was used in the current study. In this task on the CAEL Assessment, test-takers are required to explain a choice for participation in a group project. The test taker is given a handout, where some topics and some prompt details relevant to these topics are listed. In the CAEL assessment OLT, the professor's instructions and

the talk of other members are pre-recorded on the computer. First, the test taker listens to pre-recorded professor's instructions for a group oral presentation. After the instructions, the test taker has one minute to familiarize him/herself with the topics. Then, the test taker listens to pre-recorded responses of other members of a group who explain their preferences choosing from the list of topics on the handout for participation in the presentation. After listening to the other group members, the test taker is given one minute for planning and then is asked to explain his/her own presentation choice from the topics on the handout which have not been talked about by the other group members.

The two tasks used in the current study had different topics. Task A was about *youth and employment.* The other task, Task B was on *violence in society.* (See Appendix B for the prompt instruction and Appendix C for the handout for Task B[13]). Each of these two tasks was used in parallel versions by changing test method facets. Two test methods were used in the study: the individual test and the group test. (See Procedures below).

Post-test Questionnaire

With the principle that test-takers' reactions to the test offer useful information about validity of the test (Brown, 1993; Fulcher, 1996), a post-test questionnaire (see Appendix D) was designed to capture the students reactions to and opinions on the tasks and test formats used in this study. The questions on the questionnaire were

[13] The handout for Task A is not included in this paper because of task security.

intended for eliciting information about the students' reactions to the two tasks and to the two test formats and information about their preferences for the test.

The questionnaire consisted of five parts:

Part A collected the students' personal information in terms of name, sex and nationality and identified the tasks they undertook.

Questions in Part B concerned the students' reactions to and opinions on the two tasks in terms of validity (questions 1-2; questions 5-6; questions11-12), self-evaluation of task performance (questions 3-4); adequacy of timing (questions 7-8), task familiarity (questions 9-10) and task difficulty (questions 13-14).

The twelve questions in Part C were designed to elicit information on the students' perception of the two test methods: the individual test and the group test, in terms of test validity (questions 15-18), test-related anxiety (questions 19-20), preference of test method (questions 21-22), test difficulty (questions 23-24) and test fairness (questions 25-26)

Question 27 in Part D of the questionnaire asked the students about their preference of the topic and the test method.

In the last part of the questionnaire, the students were encouraged to make any comments on the tests they had taken.

3.4 Procedures

The study was conducted as a part of the EAP support course final exam at the end of the term when the students had completed this course. Each of the two tasks

was used under two different social interactional contexts: the individual context (hereafter IC) and the group context (hereafter GC). The procedures of the IC were identical with those of the CAEL Assessment OLT described above.

The recordings of Task A and Task B were similar in terms of length of time: instructions 2 minutes 17 seconds for each task; prompt conversation 2 minutes 40 seconds for Task A and 2 minutes 43 seconds for Task B; planning time 2 minutes for each task; altogether approximately 7 minutes for each task.

In the GC task, the participants were given the handouts in the classroom and were asked to discuss the topics in small groups for ten minutes. A tape-recorder was placed on each of the desks where each group of the students sat. The students were told that the tape-recording was for research purposes only and that scores would not be assigned to their performance on the group discussion in the classroom. The co-constructed group discussion was organized by the researcher and the prompt instruction for each of the tasks (which was similar to that listed in Appendix B) was delivered to the students by the EAP teacher. Immediately after the small group discussion in the classroom, the students were led to the computer laboratory to explain their choices of the topics for the oral presentation. It took approximately 1 minute for the students to walk from the classroom to the computer lab. The test takers had no planning time in the computer lab.

The participants were divided into two groups. There were 10 students in Group 1 and 13 students in Group 2. Each group was required to undertake two tasks. The split-plot design (Huynh & Feldt, 1976) of this study is shown in Table 2.

Table 2. Split-Plot Design of the Study

	Individual Context	Group Context
Task A	Group 1	Group 2
Task B	Group 2	Group 1

In order to avoid any order effect, first, Group 1 took the Task A – IC test in the computer lab while Group 2 discussed the topics of Task A in small groups (3 groups of 3 and 1 group of 4) in the classroom. Then, Group 1 had a small group discussion (2 groups of 3 and 1 group of 4) on topics of Task B in the classroom while the students in Group B made presentations individually on the computer about the topics of Task A they had chosen and then took the Task B – IC test. After Group 2 finished the Task B – IC test in the computer lab and Group 1 finished small group discussion, the students in Group 1 were required to make their individual presentations on the computer. All small group discussions in the classroom were tape-recorded. After each of the two groups had taken the two tests in the computer lab, the students were invited to complete a questionnaire in the classroom (see Appendix D). These procedures are illustrated in Figure 7:

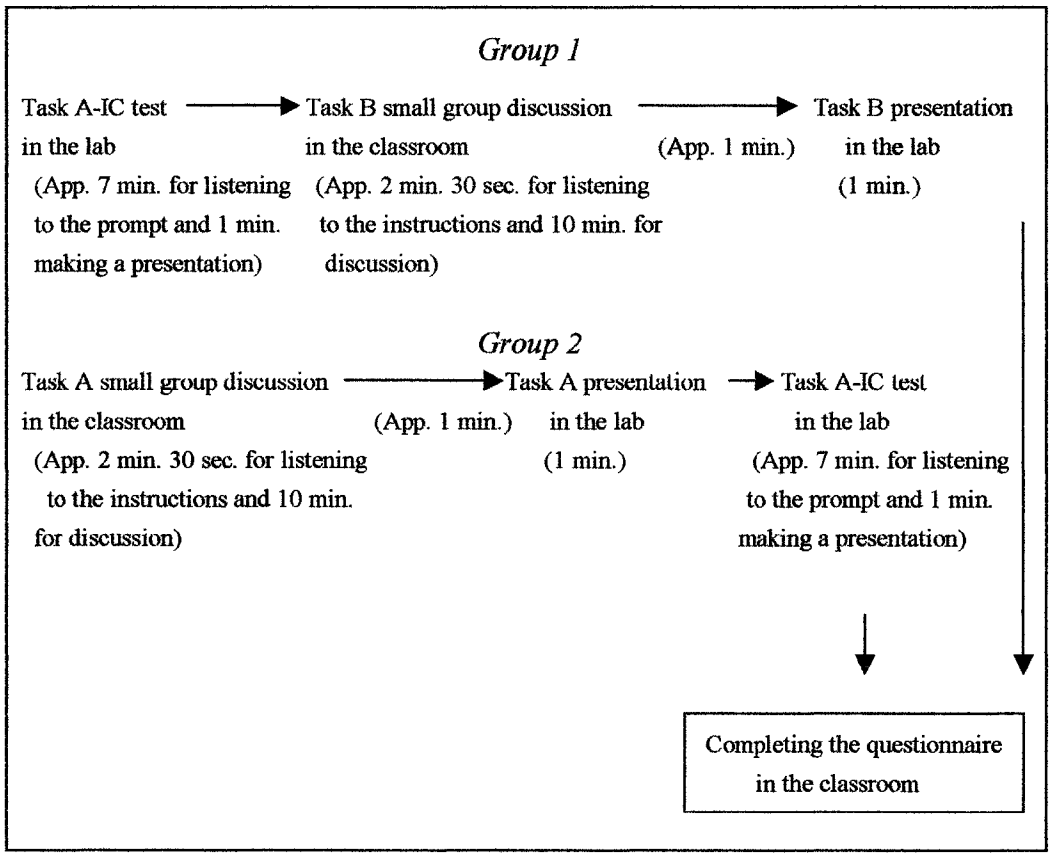

Figure 7 Flow chart of test procedures

Therefore, there were four tasks used in this study:

- Task A-IC: *youth and employment* in individual context
- Task B-IC: *violence in society* in individual context
- Task A-GC: *youth and employment* in group context
- Task B-GC: *violence in society* in group context

The two IC tasks were different from the two GC tasks in that the latter involved an additional stage of small group discussion in the classroom. Due to the inclusion of this stage, we could say that the GC tasks involved two activities: the small group

discussion (group work) and the individual presentation (individual work) while the IC tasks involved only the individual work. The two activities involved different social interactional context in terms of genre and register variables discussed in Chapter One. (See pp. 19-22) These differences are listed in Table 3:

Table 3. The social interactional Contexts of the individual work and the group work

Social interactional Context	***Individual Work***	***Group Work***
Genre	Oral presentation on the computer to remote partners	Small group discussion to choose a topic to talk about
Field	The topic of youth and employment or of violence in society	The topic of youth and employment or of violence in society
Mode	Monologue	Conversation
Tenor	Presenter to remote group member	Peer to peer

The performances of the students on the small group discussion in the classroom were assessed by the EAP teacher and the performances of all the students on the tests in the lab were rated by the two raters using the holistic Band Score Criteria listed in Appendix A. The procedures and the reliability of the CAEL Assessment scoring criteria have been reported in literature, e.g. Fox, Pychyl & Zumbo (1993). However, to allow for the rating of borderline performances, intermediate levels were included between each band score by using "+" and "-". In addition to rating the students' performance, the raters were encouraged to make comments on the candidates'

performances. The recordings of the small group discussion and of the performance of the students on the computer were transcribed for analysis.

3.5 Analysis

The procedures described above produced the following data:

a. The EAP teacher's ratings of the students' performances on the small group discussion in the classroom;

b. The two raters' ratings on the students' performances in the computer lab;

c. The completed questionnaires;

d. Transcriptions of the small group discussions in the classroom and the recordings from the computer lab.

In order to address the research questions, both quantitative and qualitative methods were employed to analyze the collected data in this study. In the quantitative data analysis, an alpha level of .05 ($p<.05$) was set. SPSS Version 12.0 for windows was used for the analysis. It should be noted that because of the small sample size of this study, any "significant" result can only be interpreted as suggestive.

The scores assigned by the two raters to all the performances of the students on the four tests ranged from 30 to 60-. These band scores were converted into nine-point Likert scales for calculation purpose, as shown in Table 4.

Table 4. Nine-point Likert scales converted from the scores by the two raters

Band score	30	30+	40-	40	40+	50-	50	50+	60-
Likert scale	1	2	3	4	5	6	7	8	9

After calculations of Pearson correlation co-efficient for interrater reliability, the two sets of scores by the two raters were averaged. The averaged scores were used for all of the following calculations:

Pearson correlation co-efficient was also calculated to examine the correlation of the averaged scores of the GC tasks by the two raters with the scores assigned by the EAP teacher to the students' performance on the small group discussion in the classroom.

In order to examine variance in test performance related to social interactional context in this task-based small group oral language test, the students' performances on the same task in the two formats IC and GC were compared in terms of test scores. Because the sample size was rather small and the numbers of students of each group were different, t-tests were used, with the test scores being the dependent variable and the format being the independent variable, to establish any differences in the scores assigned to the students' performances on the same task under IC and GC.

However, because the scores were achieved by two different groups of students in the two formats, the scores were also dependent on the students' abilities. In order to compare the abilities of the two groups of participants, a t-test was conducted to compare the means of the scores on the two tasks assigned to the two groups.

Because the two groups of students performed two different tasks under IC and GC, I also needed to examine whether there was any significant difference between Task A and Task B. In order to compare these two tasks, the means of the scores assigned to the participants' performances across these two tasks were compared using independent samples t-test. (T-tests are more robust to sample size.) The students' responses to questions in Part B of the questionnaire were also analyzed to see if there was any difference between Task A and Task B in terms of the students' perceptions.

As mentioned above, because the sample size of this study was small, any "significant" results from the quantitative analysis can only be interpreted as suggestive. To better understand the influence of the two test methods on the test performance, the students' responses to questions in Part C of the questionnaire were analyzed to obtain information on the students' perception of IC and GC.

To explore how analysis of small group oral performance from a sociocultural perspective can inform EAP task-based performance assessment in terms of validation enquiry both quantitative and qualitative methods were employed. First, correlations were calculated between the ratings of the EAP teacher assigned to the students performances on the small group discussion and the raters' ratings to the students' performances on the GC test. However, since, as suggested in Chapter 2, similar quantitative scores may represent qualitatively different performances, and different scores may not necessarily reflect differing performances, transcriptions of the recordings of the students engaged in small group discussions in the classroom

and of the students performing the tasks in the computer lab were qualitatively analyzed to see if they produced the same kind of language samples. The focus of qualitative analysis was on the influence of the social interactional context on test performance from a sociocultural point of view. The students' responses to questions on the questionnaire and the comments they made were also analyzed as sources of validation evidence.

To better understand task difficulty from a sociocultural point of view, transcriptions of the recordings of the students engaged in small group discussions in the classroom and of the students performing the tasks in the computer lab were qualitatively analyzed to see how the social interactional context influenced task difficulty. The students' responses to questions on the questionnaire and the comments they made were also analyzed to reveal more information about task difficulty from the students' perspective.

In analyzing the transcriptions of the recordings of the classroom small group discussions, Swain's (2001) approach to dialogue analysis was taken. In analyzing the transcriptions of the recordings of the lab presentations, the transcriptions were read and categorized for discourse features through a process of analytical induction (Hicks, 1994). I identified recurrent themes in the discourse features. A process of analytical induction was also employed in analyzing the students' responses to the questions on the questionnaire.

In analyzing the transcriptions of the recordings of the lab presentations, first, samples from the same test method (IC and GC) were compared. Recurrent themes

emerged and these were identified as features of the discourse produced from the IC task or the GC task. Then, discourse features of the two test methods were compared to identify any differences between the IC tasks and the GC tasks in terms of discourse features they produced.

In the following chapter, I will present results of these analyses and discuss these results in relation to the research question.

Chapter 4 Results and Discussions

In this study, data was collected from the following sources: the students' performances on the two tasks under IC (the individual context) and GC (the group context), the completed questionnaire and the small group conversation. In this section, I will bring together the results of the data analysis from these sources to discuss the research question. That is: how does the social interactional context influence test takers' performance?

4.1 Results

Comparison performances in the Individual Context (IC) and the Group Context (GC)

In this study two tasks were used with two test methods: the IC and the GC methods. Thus, four tasks were generated. The students' performances on these four tasks were rated by two raters. Table 5 shows descriptive statistics for the averaged scores across the four tasks.

The comparison of the scores assigned to the students' performances on the same task in the IC and GC formats shows that the scores awarded to the students on Task A in the GC format are significantly ($p<.05$) higher than those on the same task in the IC format ($p<.03$, $t = 2.353$, $df = 21$), but the students' performances on Task B are not significantly different in these two contexts.

Table 5. Descriptive statistics for the averaged scores across tasks

Tasks	$\bar{X}$	*Median*	*Min.*	*Max.*	*SD*	*Skew*
Task A - IC (Group 1, N = 10)	4.45	4.75	1.50	6.50	2.02	-.19
Task B - GC (Group 1, N = 10)	5.75	6.00	2.50	8.50	1.85	-.42
Task B - IC (Group 2, N = 13)	6.15	6.5	2.00	8.00	1.52	-1.76
Task A - GC (Group 2, N = 13)	6.31	7.00	2.50	9.00	1.76	-.80

☐ IC tasks ▒ GC tasks

Pearson correlation co-efficient calculations showed significantly high correlation of the averaged scores of the GC tasks by the two raters with the scores assigned by the EAP teacher to the students' performances on the small group discussions in the classroom (for Task A-GC, $r = .83$; for Task B-GC, $r = .89$).

Analysis of the students' responses to some relevant questions on the questionnaire revealed some information concerning the students' perceptions of these two test methods. The students' responses to questions 21-22, which asked about their preferences of the test methods, showed general agreement among the students that Format 2 (GC test) was preferable to Format 1 (the IC test). This preference was evidenced by the participants' responses to question 27, which was about the students' preference of the tasks and the test methods, as shown in Table 6.

The students' preferences for GC were related to their perceptions of the test difficulty. In responding to questions 23-24, which concerned test difficulty in relation to test methods, nineteen out of the twenty-three students disagreed that

Format 2 made the test more difficult while fourteen of the students believed that Format 1 made the test more difficult. This was also consistent with the students' responses to questions 19-20, which concerned test anxiety related to test methods. Eleven of the participants agreed that they felt nervous in taking the IC test while five agreed that they felt nervous in taking the GC test.

Table 6. Students' Preferences of the Test

Test	First choice	Second choice	Third choice	Fourth choice
Task A in IC	1 (4.8)	2 (9.5)	8 (38.1)	10 (47.6)
Task A in GC	9 (42.8)	9 (42.9)	2 (9.5)	1 (4.8)
Task B in IC	1 (4.8)	2 (9.5)	9 (42.9)	9 (42.8)
Task B in GC	10 (47.6)	8 (38.1)	2 (9.5)	1 (4.8)

Notes: 1.Two participants didn't make their choices, commenting that there was no difference between these four tests;
2. Numbers show the frequency;
3. Numbers in brackets show the percentage.

With regard to questions 17-18 (whether the test methods provided the students with enough opportunity to show their English speaking ability) seven students from Group 1 believed that they had enough opportunity to show their ability to speak English in doing the IC test while eight students from Group 2 believed so. For GC test, nine students from Group 1 and eleven from Group 2 felt it provided them with enough opportunity to show their ability to speak English.

In responding to questions 15-16, all the students agreed that they understood what they were supposed to do in the IC test while two of them chose "neutral" for the GC test.

With respect to the fairness of the test, the participants seemed to be more in favor of the individual tape-mediated test. Seventeen students believed it was a fair form of test while twelve regarded the co-constructed small group test as a fair test form.

However, it should be noted that since the students' ability and the task difficulty are two other major factors that had influences on the students' performance and the test score, we must take these two variables into account when examining the influences of IC and GC on the students' performance and the test score.

Comparison of Task A and Task B

The t-test result showed no significant difference between the two tasks in terms of the test scores on them awarded to the students. This statistic was consistent with the students' ratings of the task difficulty (questions 13-14) on the questionnaire. No significant difference was found between Task A and Task B in terms of the participants' general evaluation of task difficulty: in response to question 13, 11 out of the 23 students rated difficulty of Task A below 3, and 11 students rated difficulty of Task B below 3 as well in response to question 14 on the questionnaire. Responses to questions 9-10 indicated that there was also great agreement among the students as to their familiarity with the two tasks.

Responses to questions 1-2, and 11-12 showed that the majority of them believed that both Task A and Task B were valid in assessing their ability to speak English and in producing the EAP type of English. In responding to questions 7-8, most of the students felt that they had enough time to do both the tasks.

Comparison of abilities of Group 1 and Group 2

In respect to the ability of the two groups of students, statistical analysis showed a significant difference ($p<.05$) between them ($p<.04$, t = -2.124, df = 44). This difference was also evidenced by the participants' reactions to the task difficulty. Although the participants' general evaluation of task difficulty didn't suggest significant difference between Task A and Task B, closer examination of the students' responses to questions 13-14 in terms of frequencies of low ratings (below 3) and high ratings (3 or above) of task difficulty in Group 1 and Group 2 suggested that generally more students in Group 1 felt the two tasks were difficult than in Group 2. 7 out of the 10 students of Group 1 rated the difficulty of Task A 3 or above while 5 out of the 13 students of Group 2 did so. For the difficulty of Task B, 5 students of Group 1 and 7 students of Group 2 rated 3 or above.

Table 7 shows the averaged ratings of task difficulty by each group of students:

Table 7. Students' Ratings of Task Difficulty

Task	Group	Averaged rating of task difficulty
Task A-IC	1 (N = 10)	3.1
Task B-IC	2 (N = 13)	2.7
Task A-GC	2 (N = 13)	2.3
Task B-GC	1 (N = 10)	2.5

It also should be noted that the participants' responses to questions 5-6 on the questionnaire revealed that more students in Group 1 felt the tasks, regardless of the social interactional context considered here, provided them with enough opportunity to show their ability to speak English.

Consistency was found between the students' perceptions of task difficulty (questions 13-14) and their self-evaluations of their performances (questions 3-4). That is, all the students that rated task difficulty as below 3 believed that they did well on the task and vice versa. The students' responses to questions 13-14 and questions 3-4 were also consistent with the scores awarded to their performances on the test by the raters. The students who felt the task was more difficult and didn't believe they did very well on the task generally obtained lower marks and vice versa.

With regard to questions 5-6, 8 students from Group 1 believed that they had enough opportunity to show their ability to speak English in doing Task A while nine students from Group 2 believed so. For Task B, seven students from Group 1 and

seven from Group 2 felt it provided them with enough opportunity to show their ability to speak English.

Comparison of the language samples from the IC and the GC tests

Analysis of the transcriptions of the recordings of the participants performing the two tasks under IC and GC suggested some general differences in the participants' performances on the tasks in these two contexts in the following aspects:

A. Range of Vocabulary

The range of the students' vocabulary seemed to be more limited on the IC task. They used more general words and made more repetitions. For example:

Example 1

And that is most common and serious problem in the family. And this is really, really bad for the children. Sometimes in the family, the parents will beat their children or their parents will beat each other. It's really, really bad in front of the children. (Task B-IC)

Example 2

It is true that students spend their spare time to work. So they feel stress. It is true. (Task B-IC)

In contrast, on the GC task, the students used a relatively wider range of vocabulary and relatively more complex sentence structures. Unnecessary repetitions were significantly reduced. For example:

Example 3

I work in a convenience store when I was in university. And I gained some pocket money to buy anything I like. And I also gained some good experiences such as how to solve problems in any situations and how to face people in public, and, but I also got some bad influence. I am exhausted after work. I don't have time to do my homework, prepare the test and work on my presentations and work at the same time. So at the end I gave up my job because I feared the risk to fail my study. (Task B-GC)

Example 4

It is true that if there is no suitable environment for the students to learn in the school because the violence is contained such things like fighting, robbery and something like that and if that happened, the school is not like a school. It's kind of like the society outside. (Task A-GC)

B. Organization of the discourse

In the IC task, the discourse seemed to be more locally managed and to be produced on-line. That means though the students knew which topic to talk about and which aspect(s) to focus on, they didn't seem to have a global plan about how to organize their discourse. The utterance was produced more spontaneously. For example:

Example 5

Well, I'd like to choose the topic of combining study and work. [pause] Why I choose this topic is because I... I...I as a student, I have a part-time job outside so I think I have something to talk about this. First, I think the stress is the big problem...(Task A-IC)

This student then talked about the 'stress' until the end. He didn't talk about a second point concerning 'combining study and work' though there was some time left when he finished talking. It seemed that the student was prepared to discuss the topic of 'combining study and work' and to focus on one aspect: 'stress', but the pause between the first clause and the second clause indicated that he noticed there was a gap between what he had said and what he wanted to say next. In order to keep the flow of his speech, he decided spontaneously to talk about the reason for his choice. I say "spontaneously" because the student didn't seem to have planned beforehand to talk about the reason for his choice, as revealed by the repetitions of "I". The process of deciding to talk about the reason seemed to be mediated by the first clause. Since the student had planned to focus on only one aspect of the topic, the word "first" was also produced spontaneously. The production of this word was to a large extent the result of the mediation of the previous utterance "I have something to talk about this".

Example 6

OK. Today, I'd like to talk about the topic of violence on the street. This topic, the violence on the street is [pause] main problem in the society [pause] because it not only effect on the ordinary citizen but also cost for police power to protect the ordinary people. First, the effect on the ordinary citizen...(Task B-IC)

As in the previous example, this student only talked about one aspect of the topic he had chosen, "the effect of street violence on the ordinary citizen", though it seemed at the beginning that he would talk about another aspect "cost for police" as well. After introducing his choice of the topic in the first clause, the student seemed to be puzzled about how to go ahead. He then made redundant repetitions of the topic. These repetitions, on the one hand, contributed to the flow of his speech, and, on the other hand, saved attentional resources for the speaker so that he could pay more attention to what to talk about next. The repetitions of the topic also revealed the self-mediation process.

In contrast, on the GC task, the discourse seemed to be generally better planned and organized. For example:

Example 7

The topic I choose is the attitude towards youth and employment. I will talk about the positive and negative influence of doing a part-time work. At the beginning, I would like to talk about the positive side. [After discussing the positive aspects of having a part-time job] The negative side I will talk about the work influence the study...(Task A-GC)

This student took a very formal approach in constructing the discourse. It was more like a written essay, well organized from the beginning, rather than conversational style.

Example 8

Ok. I'm going to talk about violence in the school. There has been issues that has happened couple of months ago and it's in the US. There is a high school student use a gun to shoot for the whole classroom. [After describing the US accident] What's been a problem today is the violence, and especially in the schools. [The student went on to make some comments on the negative effects of violence on campus and on weapon control in high schools.] (Task B-GC)

The structure of this discourse (introduction of topic – explaining by examples – comments) also indicates a formal style. Between the first clause and the second clause, there seemed to be a shift of information but there was no processing time for this shift. And the mediating function of the first clause was not as obvious as in Example 5 and Example 6. The second clause didn't seem to be produced spontaneously.

C. Exemplifications

On the IC test, the students used more personal experiences as examples to talk about the topic. For example:

Example 9

First, I think the stress is the big problem. I have because when I work I have to think to finish my homework to prepare for the test and when I study I have to think about my work so I think this is problem and the part-time job also has the problem with the grade just because when I work, I reduce my study time for preparing something so maybe I will get low mark for test and because I am so tired of the work, I can't have enough sleep or some good sleep. [Task A-IC]

In contrast, on the GC task, students not only used personal experiences, but also experiences and accounts of other people to talk about the topic they chose. For example:

Example 10

From working, we can gain some money to buy anything we like and we can learn some experience from working like how to face people, and how to solve problems in some situation... Some people work and study at the same time. They have to compete with adults for jobs because some bosses they will probably prefer to hire experienced people to work because they have more experience but usually some bosses will hire young people because their salary is much cheaper than adults. [Task A-GC]

This student used what her partner had talked about in the small group discussion to explain "the positive side" and "the negative side" of youth involved in employment.

Some students recounted the US accident (see Example 4 above) as an example of violence in the schools in the GC, but none of them did so in the IC task.

Qualitative analysis of the small group discussion in the classroom

Qualitative analysis of the small group discussion in the classroom revealed how the students collaboratively solved problems in performing the tasks. It suggested that the problem-solving process was co-constructed by all participants. The individual's performance was largely mediated by the social interactional context.

The following example is an excerpt from the transcription of the recording of a group of three students engaged in doing the "youth and employment" task in the classroom.

Example 11

S1: [Reading from the handout] Attitudes towards youth and employment. Positive, experience and money. Yeah. Eh...Youth? So we have to focus youth, not old people and... (1)
S2: Adult (2)
S1: [Reading from the handout] Interfere with study. So, they has study problem. That... So the youth involved in the, involved in the employment may have problems with study. So we also can talk about this, this pers... as...aspect. [Reading from the handout] Compete with adults for jobs. So, that should be the negative...negative effect, affect. (3)
S2: Negative effect? (4)
S1: Yeah. (5)
S2: Interfere. (6)
S1: Interfere. So, interrupt. Not interrupt. (7)
S2: Interfere. (8)
S1: Effect. I say it is negative effect. (9)
S2: (Reading from the handout) Compete with adults for jobs. (10)
S1: (Reading from the handout) Combining study and work. That makes problem. That's problem comes out. (Reading from the handout) Stress and problems with grade. I think this negative negative [pause] effect. This...(Reading from the handout) Self confidence and money help...This positive. It's positive effect. (Reading from the handout) Types of work available for youth. (11)
S2: But we don't need to talk about this. (12)
S1: I think we should make a choice to... (13)
S3: Yeah, make a choice (14)
S1: To pick up one topic that we want. (15)
S2: Right. (16)
S3: I think so. So, why can't we choose this, positive? (17)
S1: Yes, I think...(18)
S3: It is really good experience and we can save money to do something else. (19)
S1: I think we talk about the whole topic. So we need, we need talk the positive and both the negative. (20)
S3: Both? (21)

S1: Yes. This topic has two points here. So, each four, each of four topics have this point. So we both, we both have to talk. (22)
S2: I think the third is easy to talk. (23)
S3: Yeah, there have so many to prepare. (24)
S1: Yeah, such as stress, problem with grade. (25)
S2: So, of the four topics, we choose one. We choose combining study and work. (26)
S1: Let's practice # 3, OK? (27)
S2 and S3: OK. (28)
S1 (to S3): Could you tell me what's the stress they may have who have work in school? (29)
S3: En... (30)
S1: The stress. (31)
S3: They, the student have to work a lot, and so they can't have enough time to sleep... (32)
S1: No, "sleep" is here. Stress. (33)
S3: What does "stress" mean? (34)
S1: *Yali* [The Chinese word for "stress"] (35)

In this activity, the goal of the students performing the task was to choose a topic to talk about. This goal was achieved: they decided to focus on the third topic and to talk about both the negative and the positive aspects of combining study and work. This was done collaboratively by the three students.

S1 first took the responsibility to read aloud the second topic. His reading drew his attention to the external context: what was written on the handout. He held up the word "youth" for attention and reflection and then he externalized this process (turn 1). This externalization not only revealed S1's self-mediation process but also served as an implicit invitation to his partners for their involvement. It also framed the next turn. In turn 2, S2 picked up this invitation by contributing the word "adult" to show his involvement. Similarly, the third turn was largely a self-mediation process on the part of S1 and his externalization of the process also formed the context for his

partners. It revealed that S1 seemed not to be able to use the word "effect" fluently. S2 seemed to have noticed this and he offered some help in the following turns. In turns 4—9, S1 and S2 "negotiated" over the words "effect" and "interfere". It was obvious that both of them benefited from this negotiation. Turn 9 suggested that the interaction between S1 and S2 increased the former's control of the word "effect" and the phrase "negative effect". Turn 11 revealed the interaction between the language user (S1) and the external context (the handout) and the task (to talk about the topic). The activated phrase "negative effect" seemed to play a significant role in this interaction. Turn 12 was S2's reaction to S1's last utterance. It was also S2's contribution to the co-constructed performance of the task.

S3 seemed to keep silent until turn 14, by which he made his contribution to the performance of the task. However his utterance revealed his reaction to what had been going on. He made more contribution in turns 17 and 19, drawing his partners' attention to the third topic by emphasizing and detailing it. Presumably because of this, S1 turned to S3 in turn 29 asking him to talk about "the stress of combining study and work". S1 and S2 collaboratively talked about the topic of combining study and work in turns 29--35. What seemed to be interesting was S1's use of the Chinese word in the last turn to facilitate S3's understanding of the English word "stress". This revealed S1's intention to keep the interaction going and, to keep on the topic and to elicit more information from S3 about the aspect of "stress" of this topic.

The next example is an excerpt from the transcription of the recording of a group of three students engaged in doing the "violence in society" task in the classroom.

Example 12

[After talking about the first topic listed on the handout they found that the second aspect, "effect on male/female relationship" of the topic "violence in movies" was hard to talk about. So they decided to give up the first topic.]

S1: Maybe... Let's move on second topic "violence in families". (1)

S2: Families is, I think.... Just like you guys said before, the effect on children, if the... (2)

S1 and S3: Yeah? (3)

S2: I think if the father and mother fight every day in the house and the children see that. This is the problem when they grow up. (4)

S3: They consider that all the families just like that. (5)

S2: Yeah, if children... if parents... (6)

S1: Beat their children? (7)

S2: Yeah, beat their children all the day. (8)

S3: Abuse. (9)

(All the three students laughed)

[Then they had some discussion on the second aspect of the topic "violence in families" and found it was hard to talk about so decided to give up the second topic.]

S3: So, well, the first one we don't know how to talk about effect on male/female relationship. The second one we don't know how to talk about problem with the elderly people. (10)

S2: So now, let's move on to violence in schools. [pause] I think weapon control in high schools is to be easier for us because we... (11)

S1: Yes, it's easy to talk about. (12)

S2: We have the example in the US high school. (13)

S3: Actually we don't need to talk about that example. Only one minute. (14)

S2: That's enough. (15)

S1: Just some details. (16)

S3: So, effect on learning, what should we talk about? (17)

S1: Effect on children's learning? (18)

S3: Maybe if we are the classmates and I know you have a gun there, and I cannot focus on my study, feel scared or if someone... I don't know. (19)

S2: I just think weapon control in the high school. This is the problem. (20)

S3: Effect on learning *can* [stressed] be something like that. Or if, you know... (21)

S2: Yeah, yeah, you remember just like if other students don't like you, just for example, if they don't like you and they beat you at the time and make you unconfidence. This is the effect on learning. (22)

[9-second pause]

S2: If you don't like me and you avoid me, I'll feel unconfidence and I'll feel lonely. This is the effect on the learning. (23)
S1: That maybe a kind of violence. (24)
S2: This should not be the *violence* [stressed]. Violence should be the kill and... (25)
S1: Kill another people and... (26)
S2: Something like that. OK. Let's say some violence on the street. (27)
S3: No, no. How can we say weapon control? (28)
S1 and S2: Weapon control... (29)
S2: is the simple example just like what I say of the US high school. (Recounting the US accident again) (30)
S3: And it also effect on learning. (31)
S1 and S2: Yeah. (32)

[Then they began to relate weapon control in the school with the effect of violence on learning before moving on to the next topic.]

As in Example 11, the goal of the students performing the task was to choose a topic to talk about. This goal was achieved collaboratively by the three students.

The conversation was produced on a turn-by-turn basis. The turn distribution showed that there was a shared, joint and distributed responsibility among the three interlocutors for the creation of this discourse. S1's utterance in turn 1 drew attention to the topic of violence in families. S2's utterance in turn 2 revealed that his thinking process of this topic was mediated by what the other two students had discussed before. In turn 3, S1 and S3 displayed their connectedness with S2 by showing interestedness in the content of S2's utterance. In the following turns (4--9), the three students collaboratively worked on the topic of the effect on children of violence in

families. It seemed that S2 was the initiator of the idea and S1 and S3 contributed by expanding his idea and by offering some linguistic devices. S3's utterance in turn 10 suggested that what they had talked about had been internalized in her mind. Her summary of the first two topics suggested that she felt the first two topics were not easy to talk about and she would like to work on some other topics. S2 and S1 responded this suggestion in turns 11—13. In turns 13—16 the students referred to the US accident, which had been recounted by S2 as an example of the effect on children of violence in movies. This showed that this example had become shared knowledge among the three students. S3's utterance in turn 19 revealed a self-mediated process in her mind. While she was externalizing this process, she encountered some difficulty and gave up. However, S3's utterance facilitated S2's thinking, which he externalized in turn 22. The 9-second pause between turn 22 and turn 23 indicated that S1 and S3 seemed not to understand what S2 had said. And in turn 23, S2 took the responsibility to make further explanations by paraphrasing his last utterance. S2's suggestion in turn 27 indicated that he felt they had talked enough about the topic "violence in schools". However, his suggestion was refused by S3, who insisted on keeping on the topic of "weapon control". Therefore, S2 took his responsibility again to make further explanations, this time, by re-recounting the US high school example.

4.2 Discussion

In this section, I will discuss the research question in relation to the results presented above and where necessary, I will include some additional results which have not been presented.

Two tasks were used in the current study: Task A, *youth and employment* and Task B, *violence in society.* Each of these two tasks was used in parallel versions by changing test method facets. Two test methods were used in the study: the IC test and the GC test. Thus, four tasks were developed in this study:

- Task A-IC: *youth and employment* in individual context
- Task B-IC: *violence in society* in individual context
- Task A-GC: *youth and employment* in group context
- Task B-GC: *violence in society* in group context

The two IC tasks and the two GC tasks were different in terms of the social interactional context. The IC tasks involved only one social interactional activity: the individual work of presenting one's choice of a particular topic on the computer, while the GC tasks involved a group social interactional activity in addition to the individual work. The individual work and the group work differed in terms of the social interactional context. My assumption was that the group work activity would represent a more natural social interactional context than the individual work. From a sociocultural perspective, small group discussion where peers are engaged in face-to-face oral interaction for the purpose of choosing a particular topic for presentation is a common and meaningful activity, or a genre, in the university setting. For the IC tasks, the absence of (a) "real" interactant (s) denaturalized to some extent

the interactional situation of the tasks, though recordings were provided in these tasks.

My assumption was evidenced by the results of the qualitative analysis of the students' performances on the small group discussion in the classroom. The turn-to-turn analysis of the group work activity revealed how the students' ability to speak English interacted with the social interactional context. Context was not something predetermined by the task itself; it was dynamic and constantly changing as a result of interaction between and among the interactants as they constructed it, turn by turn. Each utterance was framed by what had been said and contributed to establishing a space for subsequent utterances to be produced. Utterances produced by interactants were not merely manifestations of their knowledge or ability but also manifestations of the mediated cognitive and strategic process. This process arose in the interaction between and among the interactants. Therefore ability and context features were intricately connected and it was difficult or impossible to disentangle them.

To better understand the influence of the social interaction on test performance, I examined two other factors influencing performance: the two tasks used and the ability of the two groups of students. It should be noted, however, that there might be other factors other than the social interactional context that have influences on test performance in this study. For example, as shown in Figure 7 (p. 62), the two test methods are different not only in terms of the social interactional context, but also in terms of planning time. Also, the number of choices in the two test methods is also

different. In the GC tasks, the students could choose from the four topics while in the IC tasks, they could only choose from the two remaining topics. The different order of the two groups of students taking the IC task and the GC task might also produce difference between test performances of the two groups. It is expected that analysis of the students' responses to the questions on the questionnaire could help validate the discussions below.

If it is right that there is no significant difference between the two tasks and there is a significant difference between the two groups of students in terms of their abilities to speak English, it would follow that the differences between the means of scores across the four tests are attributable to the gap between the two groups of participants in terms of their ability to speak English and to the influence of the two test formats. Given that Group 2's performances are generally better than Group 1's performances across IC and GC, it would be reasonable to conclude that there is ability to speak English that is transferable across the IC situation and the GC situation.

Given the high correlations of the two raters' ratings assigned to the students performance on the GC tasks with the EAP teacher's ratings assigned to the students' performance on the small group discussion in the classroom, it would also be reasonable to conclude that there is ability that is transferable across the GC situation in the computer lab and the live small group discussion in the classroom. Considering the advantages of the tape-mediated test, this conclusion may be useful information.

However, given that the small sample sizes of this study made it impossible to generalize the results of either the t-tests or the correlations, these conclusions can only be considered as suggestive.

Statistics in Table 5 show the influences of the test methods on the performances of the two groups of students. For both groups, the score on the GC test is better than that on the IC test in terms of means, median, and minimum and maximum scores. Though, in terms of mean and median, the score of Task B – IC (by Group 2) is higher than that of Task B – GC (by Group 1), the latter is higher than the former in terms of minimum and maximum scores. This conclusion is also supported by the students' preference for this type of test over the IC test, as revealed by questionnaire analysis. What seems certain is that the GC test is a means of "biasing for best" (Swain, 1983; 2001) in terms of anxiety reduction and thus will help test users make better decisions, on the basis of the test score, related to what the test taker is able to do with the L2 in future real-world situations.

Given that there is no significant difference between Task A and Task B and Group 1 and Group 2 are clearly different in ability, statistics in Table 5 also suggests that the IC task separates the two groups more. However, statistics of Group 1 and Group2 show that the IC task does not separate the students within either group more than the GC task does. To address this issue, and considering the fact that the small sample sizes of this study made the t-test results only suggestive, we examined individual students' scores in each group across the tasks and revisited the data of their responses to the questions on the questionnaire and the comments they made and

of the transcriptions of the recordings. The examination revealed more interesting information.

Of particular interest are three individuals, two from Group 1 and the other one from Group 2. The two students from Group 1 obtained the lowest two scores on Task A- IC (1.5 and 2.5). However, their scores were increased by 3 points on Task B-GC to 4.5 and 5.5 respectively. Both of them expressed their preferences for the GC format over the IG format consistently in responding relevant questions on the questionnaire. One of them made the following comment on the questionnaire:

In the group work, the talk of the group members gave me some ideas about what to talk about in the test.

Both of these two students were active in making their contributions to the small group discussion.

The student from Group 2 obtained 6 on Task B-IC; surprisingly, however, he obtained 4 on Task-GC. What is more interesting is that the scores of the other two students who had been in the same small group with this student in the classroom discussion also decreased (one from 6 to 5 and the other from 7 to 6). One of the raters commented the performances of these two students as "good phrasing but slow pace".

Revisiting the data of the recording of this small group in the classroom discussion revealed that most of their discussion involved talking about the language itself, e.g. meanings of some words on the handout. The relatively low marks of the

three students in this small group on Task A-GC were consistent with their relatively low scores assigned by the EAP teacher in the classroom.

The student who obtained 7 on Task B-IC made the following comments on the questionnaire:

The thing that I worry most about in spoken test is I can't say what I want to say because I don't know some of the words. It's good that in the group work I can ask my partners how to say something. In the group work, I tried to remember some words, but I forgot them in front of the computer.

The student who achieved 6 on Task B-IC commented that

I don't see any differences between the two tests I take. The group work does not make difference. A test is a test.

These findings might present some challenges to the validity of the tape-mediated test. For example, "superficial fluency" in some students' speaking on a tape-mediated test might mask their deficiency in speaking ability of participating in small group discussion, an issue also discussed by Fox (2004). How to identify "superficial fluency" in speaking on a tape-mediated test and how to highlight it through rating criteria specification might be an interesting topic for future research. Evidence obtained from comparing the candidate's performance on the small group discussion in the classroom and his/her performance on the GC task in the computer lab might usefully inform this research.

As mentioned above, findings of this study have shown that if the purpose of task-based L2 performance assessment is to assess test takers' ability to do certain types of tasks, the social interactional context needs to be taken into account. Locally

examining the test taker's ability engendered by the features of specific social interactional contexts and systematically exploring the test taker's inconsistent performances across context will provide the tester with a clearer picture about the test taker's ability and qualitative analysis of the dynamic and mutual influences of performance and interactional context will provide useful evidence for validation enquiry of task-based L2 performance assessment.

Comparison of the performances of the students on the IC format test and the GC format test and the small group discussion in the classroom revealed that the students paid attention to different aspects of their performances. The classroom discussion seemed to be a here-and-now interactional problem-solving activity where the students paid more attention to what to talk about and the effectiveness of communication, neglecting linguistic errors to keep the flow of communication. On the IC and GC tests in the computer lab, the participants paid more attention to how to talk. On the IC test, they seemed to allocate their attention more to monitoring their language. As a result, they made more self-corrections when they were aware of errors in their utterances and made more repetitions when they noticed a gap between what they wanted to say and what they could say or when they could not work out a solution to the gap. They didn't often "take the risk" of experimenting with language forms which they were not in good control of. On the GC format test in the lab, on the other hand, the students produced more planned language. These points may also be connected to the issue of "superficial fluency" mentioned above. Thinking along these lines, I would suggest that discourse produced in the small group discussion

may form a useful basis for constructing empirically based rating scales (Upshur and Turner, 1999; Turner, 2000) for small group oral language performance test in the EAP context.

It is also noticeable that though, as mentioned above, the significantly high correlation between the two raters' averaged scores for the students' performance on the GC tasks and the EAP teacher's scores for their performance on the small group discussion shows there is ability that is transferable across the GC situation in the computer lab and the live small group discussion in the classroom, qualitative analysis of the language samples in these two situations shows differences existed among the individual participants in terms of the aspects of the classroom discussion they transferred to the GC test in the lab and the extent to which they transferred them. Closer examination of the variances in the individual candidate's performance across these two situations may better inform our interpretation of the candidate's ability to speak English.

With regard to the influence on task difficulty of the social interactional context, statistics in Table 6 show that overall students found the two IC tasks more difficult than they did the two IC tasks. The students' responses to the relevant questions on the questionnaire and their comments on the two test formats also indicate that the small group discussion helped reduce the difficulty of the task on the part of the test takers in terms of more planning time and test enjoyment. Generally, the spontaneous support provided by the interactants positively affected the performance of the students and helped reduce test anxiety. Here are some comments made by the

students on the questionnaire:

- *Format 2 is really interesting to me.*
- *I like Format 2 because we have more time to prepare for the test.*
- *I like Format 2 and I don't feel nervous in it because it is what we do in class almost every day.*
- *I like Format 2 because if I have a question, I can ask my friend.*

However, some students showed their concern about the somewhat negative influence of their interactants. For example:

- *I don't think Format 2 is fair because some partners are helpful some others are not.*
- *If there is someone who is very talkative in our group, the other people will not have many opportunities to prepare for the test.*

To see if there was any connection between the students' perception of task difficulty and the influence of the interactional context, I further examined the students' ratings of task difficulty attempting to identify students who rated the IC task as more difficult than the GC task and students who felt the difference between the IC task and GC task in terms of difficulty to a relatively greater extent. Two cases were identified. Student A from Group 1 chose 4 in rating Task A-IC difficulty and chose 1 in rating Task B-GC difficulty. Student B in Group 2 chose 3 in rating Task A-GC difficulty while she chose 2 in rating the difficulty of Task B-IC. Then I re-examined the transcriptions of the recordings of these two students participating in the small group discussion. It was revealed that Student A seemed to be relatively more active in contributing to the small group discussion, asking for support from his

partners and lending support to his partners during the group work while Student B was relatively more quiet in her group work, being a listener most of the time to her partners. Interestingly, Student A was the student who made the comment "I like Format 2 because if I have a question, I can ask my friend" on the questionnaire. However, Student B didn't make any additional comments on the questionnaire.

These results suggest that there seems to be some relationship between the extent to which the candidate contributes to the co-construction in the social interaction and his/her perception of task difficulty. On the other hand, the extent to which the candidate participates in the interaction is constrained by the social interactional context – the performance of the interlocutor(s). Therefore when considering task difficulty in task-based L2 oral performance assessment, we should take into account not only the stable and global features of the task but also the dynamic influences of the interactant's performance. This also implies in developing small group discussion oral test, pairing may be a critical issue to consider.

Chapter 5 Conclusions and Implications

Sociocultural theory has drawn increasing attention from researchers in the field of L2 education (Lantolf, 2000b). Moving away from the previous exclusive focus on the individual in SLA research, a growing number of L2 researchers have taken a sociocultural perspective exploring the rich sociocultural contexts of language learning with the underlying assumption that language learning is not just an individual psychological process of input, interaction and output, or a process of accumulation of knowledge and skills, but also a process of socialization, a process of participation, of becoming a member of a certain community (Lantolf, 2000a; Atkinson, 2002).

Sociocultural theory gives prominence to social interaction in discussing the development of cognition and the use of strategic competence and cognitive processes. In the field of L2 education, recent research has shown how learning and language acquisition are realized through a collaborative interactional process in which the language of the interaction provides affordance which learners appropriate for their own purposes. A well-used form of interaction in the L2 education context is small group discussion. The function of small group discussion for language development in terms of peer scaffolding or peer mediation from a sociocultural perspective has been well documented in SLA literature (See Lantolf, 2000b for a detailed a review). In the university setting small group discussion is also one common discipline-general academic activity. It is also frequently used in EAP programs

However, in spite of the popularity of small group discussion in the field of L2 education, the group discussion task has received little attention in the language testing community (Bonk and Ockey, 2003).

The current study was designed to explore the necessity, usefulness and feasibility of incorporating a sociocultural approach in a small group oral language test in the EAP context by considering the influence of the social interactional context on test performance from a sociocultural view. Two tasks from the OLT of the CAEL Assessment were used and parallel task versions were developed by changing test method facets (the IC test and the GC test). Both quantitative and qualitative methods were employed. In this chapter, I will draw some conclusions from this study and discuss their theoretical and practical implications for language testing.

5.1 Limitations of the Study

First, however, it is important to acknowledge the limitations of the study.

There are a number of limitations of this study that should be noted. For example, as has been repeatedly mentioned in the last two chapters, because of the small sample size (N = 23), any results from the analysis of the current study should be only interpreted as suggestive. Also, the sample was taken from one class; the students were familiar with each other, and the EAP teacher, who was the rater of the classroom group work in this study, was also from the same class. Further, because of the time limit, no follow-up research was conducted to collect additional data to

support the findings of this research. Therefore, it is important to interpret the findings with these limitations in mind.

5.2 The Sociocultural Context: Re-thinking Pandora's Box

As previously indicated, generally the students' performance on the GC task was better than that on the IC task and the GC task was generally favored by the students over the IC task. The GC task, which involved a small group discussion procedure, could also be viewed as more authentic to EAP use in the university setting in the sense that the group work, which was an outcome of the GC task, provided the students with the possibility of using the English language in a more authentic and natural social interactional context: small group discussion is a meaningful and purposeful sociocultural activity wherein peers are engaged in face-to-face oral communication working on a specific topic. In contrast, the absence of (a) "real" interactant (s) in the individual work activity denaturalized the language use situation.

Considering these advantages of the GC task, it seems that it could be a potentially attractive substitute to the IC task currently used in the OLT of the CAEL Assessment.

Further, because at the final stage the candidate was required to perform the GC task individually on the computer, it was possible to evaluate the individual's performance on the GC task. Moreover, the language sample generated during the small group discussion that was an outcome of the GC task also provided test developers and researchers a useful source for test validation evidence and a

potentially useful basis for constructing empirically based rating scales (Upshur and Turner, 1999; Turner, 2000) for small group oral language performance test in the EAP context.

Therefore, findings of this study may have practical implications for task development, validation enquiry and rating scales construction.

As mentioned above, although Bachman (1990) sees the inseparability of language ability and language use situation as the fundamental dilemma of language testing, he maintains the separation of these two in order to permit generalizations on the basis of transferable abilities. Although Skehan (1995; 1998) attempts to integrate task characteristics and task conditions in defining the construct of *ability for use*, because of his psycholinguistic and cognitive theoretical orientation, his focus is on modeling selective requirements on ability of task characteristics in terms of task difficulty. The underlying assumption is that task characteristics and their influences on ability are stable and they can be described globally.

The individual-focused cognitive approach of modeling, conceptualizing and explaining the composition of the stable L2 ability and the interaction of its components is consistent with the nomothetic tradition of SLA research (cf. Markee, 1994).

From a sociocultural perspective of interaction, however, context is dynamic, constantly changing and co-constructed and must be analyzed locally. From a sociocultural perspective, social interactional contexts will not be seen as problems that need to be avoided or controlled, or as unwanted construct-irrelevant variances in

terms of test validity Research on task difficulty from a cognitive perspective will provide necessary but not sufficient information for the design of task-based L2 performance assessments and for the investigation of their validity. What is also needed is research into how tasks are actually accomplished in "real-world" terms, how performance is co-constructed through social interaction between interactant(s) in particular social interactional context, and the influence on task difficulty of the social interactional context and the test taker's perception of task difficulty. Without such research, the generalizability of task-based performance test scores will remain in question. To do such research entails the employment of qualitative research methods.

Therefore, research on the influence of the social interactional context on test performance from a sociocultural perspective also has methodological implications for language testing.

SLA and language testing researchers have earnestly occupied themselves with models of abilities underlying language performance since Hymes (1972) coined the term *performance* and *communicative competence.* For over half a century, different models have been proposed (Davies, 1989; Taylor, 1988; Bialystok and Sharwood-Smith, 1985; Widdowson, 1979; Canale and Swain, 1980; Canale, 1983; Bachman, 1990; Skehan, 1998); however, consensus is still absent among the researchers. McNamara (1996) sees the work of modeling performance in language performance assessment as opening Pandora's box. Researchers in mainstream SLA and the language testing community have generally taken an individual-focused

cognitive approach to open this box, viewing performance as manifestation of a trait or a bundle of traits (Young, 2000)-- the individual's internal stable core language abilities.

McNamara (1997a) points out that the exclusive focus on the ability of the candidate in the cognitive approach views the candidate in a strangely isolated light, who is exclusively responsible for the performance. He argues "clearly a performance is not a simple projection of what is in the head of the candidate, even if that display is mediated by the candidate's strategies for dealing with the interactional context in which it is to be achieved" (p.453).

Taking a sociocultural perspective, we might not see the Pandora's box as a "black hole". Instead, we might see it as a kaleidoscope, or a dynamic landscape of meaning and relationship, where the individual plays a particular role as a participating member in a particular sociocultural context.

Traditionally, L2 testing has been informed by SLA and language testers have been the consumer of SLA theories. Studies on language performance in relation to the social interactional context may suggest the significant role language testers can play in constructing theory.

Therefore, research on the influence of the social interactional context on test performance from a sociocultural perspective also has theoretical implications for language acquisition.

References

Alderson, J.C. and Banerjee, J. 2002. Language testing and assessment, part 2. *Language Testing* 35, 79-113.

Atkinson, D. 2002. Toward a Sociocognitive approach to second language acquisition. *The Modern language journal* 86 (4) 525-545

Bachman, L. 1990. *Fundamental consideration in language testing.* Oxford: Oxford University Press.

Bachman, L. F. 2002. Some reflections on task-based language performance assessment. *Language Testing.* 19 (4): 453-476.

Bachman, L.F., Lynch B.,K. and Mason, M. 1995. Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing.* 12 (2): 238-257.

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

Bachman, L. F. and Cohen, A. D. 1998. Language testing-SLA interfaces: An update. In Bachman, L. F. and Cohen, A. D. (ed.) *Interfaces between second language acquisition and language testing research.* New york: Cambridge University Press. 1-31.

Banerjee, J. and Luoma, S. 1997: Qualitative approaches to test validation. In Clapham, C. and Corson, D. (eds.) *Encyclopedia of language and education, Volume 7: Language testing and assessment.* Dordrecht: Kluwer Academic Publishers, 275-287.

Berkoff, N.A. 1985. Testing oral proficiency: a new approach. In Lee, Y.P. (ed.) *New Directions in Language Testing.* Oxford: Pergamon Institute of English, 93-100

Berry, V. 1997. Gender and personality as factors of interlocutor variability in oral performance tests. Paper presented at 19th Annual Language Testing Research Colloquium. Florida, USA March 6-9

Bialystok, E. and Sharwood-Smith, M. 1985. Interlanguage is not a state of mind: an evaluation of the construct for second language acquisition. *Applied Linguistics* 6 (2) 101-117.

Bonk, W., J. and Ockey, G., J. 2003. A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing* 20 (1) 89—110.

Brindley, G. 1994. Task-centered assessment in language learning: the promise and the challenge. In Bird, N., Falvey, P., Tsui, A., Allison, D. and McNeill, A. (eds.) *Language and learning: papers presented at the Annual International Language in Education Conference.* (Hong Kong, 1993).

Brown, A. 1995. The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing.* 12: 1—15

Brown, G., Anderson, A., Shilcock, R. and Yule, G. 1984. *Teaching talk: strategies for production and assessment.* Cambridge: Cambridge University Press.

Brown, J., D., Hudson, T, Norris, J and Bonk, W., J. 2002. *An investigation of second language task-based performance assessments* (Technical Report # 24). Honolulu: University of Hawaii, Second Language Teaching & Curriculum Center.

Buckingham, A. 1997. Oral language testing: do the age, status and gender of the interlocutor make a difference? Unpublished MA dissertation, University of Reading.

Bygate, M., Skehan, P., and Swain, M. (eds.) 2001. *Researching pedagogic tasks: second language learning, teaching and testing.* New York: Pearson Education.

Canale, M. 1983a. From communicative competence to communicative language pedagogy. In Richards J, C. and Schmidt R.W. (ed.) *Language and communication.* London: Longman. Pp 2-27.

Canale, M. 1983b. On some dimensions of language proficiency. In Oller, J., W. (eds.) *Issues in language testing research.* Rowley, Mass.: Newbury House.

Canale, M. and M. Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics* 1: 1—47.

Chalhoub-Deville, M. 2001. Task-based assessments: characteristics and validity evidence. In Bygate, M., Skehan, P., and Swain, M. (eds.) *Researching pedagogic tasks: second language learning, teaching and testing.* New York: Pearson Education, 210-228.

Chalhoub-Deville, M. 2003. Second language interaction: current perspectives and future trends. *Language Testing 20 (4)* 369-383.

Chapelle, C. 1998. Construct definition and validity inquiry in SLA research. In Bachman, L. and Cohen, A. (eds.) *Interfaces between second language acquisition and language testing research.* New York: Cambridge University Press. 32-70.

Chomsky, N. 1965. *Aspects of the theory of syntax.* Cambridge, Mass.: MIT Press.

Crookes, G. V. 1986. *Task Classification: A cross-disciplinary review.* Technical Report No. 4. Honolulu, Center for Second Language Research, Social Science Research Institute, University of Hawaii at Manoa.

Crookes, G., and S, Gass. 1993. *Tasks and language learning: integrating theory and practice.* Clevedon, Avon: Multilingual Matters.

Davies, A. 1989. Communicative competence as language use. *Applied Linguistics* 10 (2): 157-170.

Davis, K. 1995. Qualitative theory and methods in applied linguistic research. *TESOL Quarterly.* 29 (3) 427—454.

Douglas, D. 1994. Qunantity and quality in speaking test performance. *Language Testing.* 11 (2): 125-144.

Douglas, D. 1998. Testing methods in context-based second language research. In Bachman, L. and Cohen, A. (eds.) *Interfaces between second language acquisition and language testing research.* New York: Cambridge University Press. 141-155.

Douglas, D. 2000. *Assessing language for specific purposes.* Cambridge: Cambridge University Press.

Duff, P. 1986 Another look at interlanguage talk: taking task to task. In R. Day (ed.) *Talking to learn: conversation in second language acquisition.* Rowley, MA: Newbury House.

Duranti, A. 1988. The ethnography of speaking: toward a linguistics of the praxis. In Newmeyer, T. (ed.) *Linguistics: the Cambridge Survey. Vol. IV: Language: The Socio-cultural context.* Cambridge: Cambridge University Press. 210-228.

Duranti, A. and Goodwin, C. (ed.) 1992. *Rethinking Context: Language as an interactive Phenomenon.* Cambridge: Cambridge University Press.

Educational Testing Service 1982. *Oral proficiency testing manual.* Princeton, NJ: ETS.

Eggins, S. 1994. *An introduction to systemic functional linguistics.* London: Pinter Publishers.

Ellis, R. 1997. SLA and language pedagogy: An educational perspective. *Studies in Second language Acquisition,* 19, 69—92.

Ellis, R. 2001. Non-reciprocal tasks, comprehension and second language acquisition. In Bygate, M., Skehan, P., and Swain, M. (eds.) *Researching pedagogic tasks: second language learning, teaching and testing.* New York: Pearson Education, 49-74.

Elman, J.L., Bates, E.A. Johnson, M.H., Karmiloff-Smith, A., parisi, D. & Plunkett, K. 1996. *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT press.

Firth, J. R. 1935. The technique of semantics. *Transactions of the Philological Society.* 36-72.

Firth, J. R. 1957. *Papers in Linguistics 1934-1951.* London: Oxford University Press.

Fitzpatrick, R. and Morrison, E. J. 1971. Performance and product evaluation. In Thorndike, R. L. (ed.) *Educational measurement* (second edition). American Council on Education, Washington, D. C. 237-270, reprinted in Finch, F. L. (ed.). 1991. *Educational performance assessment.* The Riverside Publishing Company. Chicago, 89-138.

Foster, P. and Skjehan, P. 1996. The influence of planning and task type on second language performance. *Studies in Second Language Acquisition.* 18: 299-323.

Fox, J. 1999; 2000. *The Canadian Academic English Language Assessment: test score and users' guide.* Ottawa: Language Assessment & Testing Research Unit, Carleton University

Fox, J. 2001. It's all about meaning: language test validation in and through the landscape of an evolving construct. Unpublished PhD thesis. McGill University. Montreal, Quebec, Canada.

Fox, J. 2002: Test takers' and test raters' accounts of three L2 writing tests. Paper presented at the Language Testing Research Colloquium. Hong Kong, China, December.

Fox, J. 2004. Test decisions over time; tracking validity. *Language Testing* 21 (4): 437-465

Fox, J., Pychyl, T.A., & Zumbo, B.D. 1993. Psychometric properties of the CAEL Assessment: An overview of development, format, and scoring procedures.

Carleton Papers in Applied Language Studies. Vol. 10, Carleton University, Ottawa, Canada. 1-12.

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

Gass, S. 1998. Apples and oranges: Or, why apples are not orange and don't need to be: A response to Firth and Wagner. *Modern Language Journal.* (82) 83-90.

Haertel, E. 1992. Performance measurement. In Alkin M. C. (ed.) *Encyclopedia of educational research,* 6th edition. pp 984-989.

Halliday, M. A. K. 1978, *Language as social semiotic: the social interpretation of language and meaning.* London: Edward Arnold.

Halliday, M. A. K. 1994. *An introduction to functional grammar* (2nd edition). London: Edward Arnold.

Harley, B., Allen, J.P.B., Cummins, J. and Swain, M. 1990. *The development of second language proficiency.* Cambridge: Cambridge University Press.

Hall, J. K. 1995. "Aw, man, where you goin?": classroom interaction and the development of L2 interactional competence. *Issues in Applied Linguistics,* 6, 37—62.

He, A.W. and Young, R. 1998. Language proficiency interviews: a discourse approach. In Young, R. and He, A. W. (eds.) *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency.* Amsterdam: John Benjamins Publishing Company.

Henning, G. 1996. Accounting for nonsystematic error in performance ratings. *Language testing.* 13 (1): 53—61.

Hicks, A. M. 1994. Qualitative comparative analysis and analytical induction : the case of the emergence of the social security state. *Sociological Methods and Research,* 23, 1, 86-113 .

Hilsdon, J. 1991: The group oral exam: Advantages and limitations. In Alderson, J. C. and North, B. (eds.) *Language testing in the 1990s.* Modern English Publications in association with the British Council. London: Macmillan.

Huerta-Macías, A. 1995. Alternative assessment: responses to commonly asked questions. *TESOL Journal, 5* (1), 8-11

Huynh, H., & Feldt, L. S. (1976). Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split-plot designs. *Journal of Educational Statistics, 1,* 69-82.

Hymes, D. 1972. On communicative competence. In J., Pride and J., Holmes (eds.) *Sociolinguistics.* Harmondsworth: Penguin.

Iwashita, N. 1997. The validity of the paired interview format in oral performance testing. Paper presented at 19th Annual Language Testing Research Colloquium. Florida, USA March 6-9

Jacoby, S. and Ochs, E. 1995. Co-construction: An introduction. *Research on Language and Social Interaction* 28 (3) 171-183.

Jahnson, M. 2001: *The art of nonconversation: a re-examination of the validity of the oral proficiency interview. New Haven, CT: Yale University Press.*

James, G. 1988. Development of an oral proficiency component in a test in English for academic purposes. In Hughes, A. (eds.) *Testing English for university study. ELT Documents* 127. Oxford: Modern English Publications and the British Council.

Jacoby, S. and Ochs, E. 1995. Co-construction: an introduction. *Research on Language and Social Interaction.* 28/3: 171-183.

Jacoby, S. and McNamara, T. 1999. Locating competence. *English for Specific Purposes* 18 (3) 213-241.

Johnson, M. and Tyler, A. 1998. Re-analyzing the OPI: How much does it look like natural conversation? In Young, R. and He, A. W. (ed.). *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency.* Amsterdam: John Benjamins Publishing Company. 27-51.

Kenyon, D. 1992. Introductory remarks as symposium on development and use of rating scales in language testing. 14th Language Testing Research Colloquium, Vancouver, 27 February to 1 March.

Kowal, M. and Swain, M. 1994. Using collaborative language production tasks to promote students' language awareness. *Language Awareness* 3. 73—93.

Kunnan, A. 1998. Approaches to validation in language assessment. In Kunnan, A. (ed.) *Validation in language assessment.* Nahwah, NJ: Lawrence Erlbaum, 1-16.

Lantolf, J. P. 2000a. Introducing sociocultural theory. In Lantolf, J. P. (eds.) *Sociocultural theory and second language learning.* Oxford: Oxford University Press. 1-26.

Lantolf, J. P. 2000b. Second language learning as a mediated process. *Language Teaching* 33, 79—96.

Lantolf, J.P and Appel, G. (Eds.) 1994. *Vygostskian approaches to second language research.* Norwood, NJ: Ablex.

Lazaraton, A. 1992. The structural organization of a language interview: a conversation analytic perspective. *System* 20, 373-386.

Long, M. H. and Norris, J. M. 2000. Task-based language teaching and assessment. In Byram, M. (ed.) *Encyclopedia of language teaching.* London: Routledge. 597-603.

Long, M. 1989. Task, group, and task-group interaction. *University of Hawaii Working Papers in English as a Second Language.* 8 (20) 1—26.

Long, M. L. 1997. Construct validity in SLA research; A response to Firth and Wagner. *Modern Language Journal.* (81), 318—323.

Low, P. and Clifford, R.T. 1980. Developing an indirect measure of overall oral proficiency. In Firth, J.R. (eds.) *Measuring spoken language proficiency.* Washington, DC: Georgetown University Press.

Lumley, T., and T. McNamara. 1995. Rater characteristics and rater bias: implications for training. *Language Testing* 12: 55—71.

Malinowsky, B. 1923. *The problem of meaning in primitive languages*, supplement I to C. K. Ogden, and I. A. Richards (eds.) *The meaning of meaning.* New York: Harcourt, Brace, and World, Inc.

Malinowsky, B. 1935. Coral Gardens and their magic, a study of the methods of tilling the soil and of agricultural rites in the Trobriand Islands, Vol. 2. *The language of Magic and Gardening.* London: Allen and Unwin.

Markee, N. P. 1994 Toward an ethnomethodological respecification of second-Language acquisition studies. In Tarone, E.E., Gass, S.M. and Cohen, A.D. (eds.) *Research methodology in second-language acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc, Publishers. 89-116.

Martin, J.R. (1984) Language, register, and genre, In F. Christie[eds.] *Language studies: Children writing.* Geelong, Victorka: Deakin University Press. pp. 21-30

McNamara, T. F. 1990. Item response theory and the validation of an ESP test health professionals. *Language Testing.* 7(1): 52—75.

McNamara, T. F. 1995. Modeling performance: opening Pandora's box. *Applied Linguistics* 16/2: 159-179

McNamara, T. F. 1996. *Measuring second language performance.* London: Longman.

McNamara, T. F. 1997a. 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics* 18 (4) 446-466.

McNamara, T. F. 1997b. Performance testing. In Clapham, C. and Corson, D. (eds.) *Encyclopedia of language and education, Volume 7: Language testing and assessment.* Dordrecht: Kluwer Academic Publishers, 131-139.

Mehnert, U. 1998. The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition.* 20 (1): 83-108.

Messick, S. 1989. Validity. In R. L. Linn (ed.) *Educational Measurement.* Third edition. New York: Macmillan. 13-103.

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

Morrison, D.M. and Lee, N. 1985. Simulating an academic tutorial: a test validation study. In Lee, Y.P. (eds.) *New directions in language testing.* Oxford: Pergamon Institute of English, 85—92.

Norris, J. M. 2002. Interpretations, intended uses and designs in task-based language assessment. *Language Testing.* 19 (4) 337-346.

Norris, J., Brown, J., Hudson, T. and Yoshioka, J. 1998. *Designing second language performance assessments* (Technical Report # 18). Honolulu: University of Hawaii, Second Language Teaching & Curriculum Center.

Norris, J., Brown, J., Hudson, T. and Bonk, W. 2002. Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing* 19 (4) 395-418.

Norton, B. 2000: Writing assessment: language, meaning, and marking memoranda. In Kunnan, A.J. (ed.) *Fairness and Validation in Language Assessment: Selected*

Papers from the 19th Language Testing Research Colloquium, Orlando, Florida. New York: Cambridge University Press, 20—29.

Ochs, E. 1979. Transcription as theory. In Ochs, E. and Schieffelin, B. (ed.) *Developmental Pragmatics.* New York: Academic Press. 43-72.

O'Sullivan, B. 1995. Oral language testing: does the age of the interlocutor make a difference? Unpublished MA project, University of Reading.

O'Sullivan, B. 2002. Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing* 19 (3) 277—295.

O'Sullivan, B. and Porter, D.1995. The importance of audience age for learner-speakers and learner-writers from different cultural backgrounds. Paper presented at the RELC Conference, Singapore, April,1995.

O'Sullivan, B. and Porter, D.1996. Speech style, gender and oral proficiency interview performance. Paper presented at the RELC Conference, Singapore, April,1996.

Pennycook, A. 1997. Cultural alternatives and autonomy. In P. Benson & P. Voller (Eds.) *Autonomy and independence in language learning.* London: Longman. pp35—53.

Perrett, G. 1990. The language testing interview: a reappraisal. In de Jong, J.H.A.L. and Stevenson, D.K., (ed.) *Individualizing the assessment of language abilities.* Philadelphia, PA: Multilingual Matters, 225-238.

Pica, T., Kanagy, R. and Falodun, J. 1993. Choosing and using communication tasks for second language instruction. In Gass, S. and G. Crookes (eds.) *Tasks and language learning: integrating theory and practive.* (pp. 9--34) Clevedon, Avon: Multilingual Matters.

Porter, D. 1991a: Affective factors in language testing. In Alderson, J. C. and North, B. (eds.) *Language testing in the 1990s.* Modern English Publications in association with the British Council. London: Macmillan, 32—40.

Porter, D. 1991b: Affective factors in the assessment of oral interaction: gender and status. In Arnivan, S. (ed.) *Current development in language testing.* Singapore: SEAMEO Regional Language Centre. Anthology Series 25, 92—102.

Porter, D. and Shen Shu Hung. 1991. Gender, status and style in the interview. *The Dolphin, Volume 21.* Aarhus: Aarhus University Press, 117—28.

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 Alderson, J. C. and North, B. (eds.) *Language testing in the 1990s.* Modern English Publications in association with the British Council. London: Macmillan.178—88.

Robinson, P. 1995. Task complexity and second language narrative discourse. *Language Learning.* 45/1: 99-140.

Samuda, V. Guiding relationships between form and meaning during task performance: the role of the teacher. In Bygate, M., Skehan, P., and Swain, M. (eds.) *Researching pedagogic tasks: second language learning, teaching and testing.* New York: Pearson Education, 119—140.

Seidenberg, M. S. 1996. language acquisition and use: learning and applying probabilistic constraints. *Science,* 275, 1599—1603.

Shohamy, E. 1982. Predicting speaking proficiency from cloze tests: theoretical and practical considerations for test substitutions. *Applied Linguistics.* 3, 161—71.

Shohamy, E. 1983. The stability of oral proficiency assessment in the oral interview procedure. *Language Learning 33* 527-40

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

Shohamy, E. 1995. Performance assessment in language testing. *Annual Review of Applied Linguistics,* 15, 188-211.

Shohamy, E., Reves, T. and Bejarano, Y. 1986. Introducing a new comprehensive test of oral proficiency. *English Language Teaching journal* 40, 212—20.

Shohamy, E. and Stansfield, C. 1991. The Hebrew oral test: an example of international cooperation. *AILA Bulletin 7.*

Skehan, P. 1995. Analysability, accessibility, and ability for use. In Cook, G. and Seidlhofer, B. (eds.) *Principle and practice in applied linguistics.* Oxford: Oxford University Press.

Skehan, P. 1996. A framework for the implementation of task-based instruction. *Applied Linguistics* 17(1) 38--62

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

Skehan, P. 2001. Tasks and language performance assessment. In Bygate, M., Skehan, P., and Swain, M. (eds.) *Researching pedagogic tasks: second language learning, teaching and testing.* New York: Pearson Education, 167—185.

Skehan, P. and Foster, P. 1997. The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research.* 1 (3): 185-211.

Spolsky, B. 1989. Competence, proficiency and beyond. *Applied Linguistics* 10 (2): 138-156.

Stansfield, C.W. 1991. A comparative analysis of simulated and direct oral proficiency interviews. In Anivan, S. (eds.) *Current developments in language testing.* Singapore: SEAMEO-RELC. 199-209

Stansfield, C.W., Kenyon, D.M., Paiva, R., Doyle, F., Ulsh, I. and Cowles, M.A. 1990. The development and validation of the Portuguese speaking test. *Hispania 73,* 641-51.

Swain, M. 1985. Large-scale communicative language testing: a case study. In Y. P. Lee, A. C. Y. Fok, R. Lord, and G. Low (eds.) *New directions in language testing.* Oxford: Pergamon Press: 35-46.

Swain, M. 1995. Three functions of output in second language learning. In Cook, G. and Seidlhofer, B. (eds.) *Principles and practice in applied linguistics: studies in honour of H. G. Widdowson.* Oxford: Oxford University Press. 125-144.

Swain, M. 2000. The output hypothesis and beyond: mediating acquisition through collaborative dialogue. In Lantolf, J. P. (eds.) *Sociocultural theory and second language learning.* Oxford: Oxford University Press. 97—114.

Swain, M. 2001: Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing* 18, 275-302.

Swain, M. and Lapkin, S. 1998. Interaction and second language learning: two adolescent French immersion students working together. *Modern Language Journal* 82, 320—337.

Swain, M. and Lapkin, S. 2001. Focus on form through collaborative dialogue: exploring task effects In Bygate, M., Skehan, P., and Swain, M. (eds.) *Researching pedagogic tasks: second language learning, teaching and testing.* New York: Pearson Education, 99-118.

Taylor, D. S. 1988. The meaning and use of the term 'competence' in linguistics and applied linguistics. *Applied Linguistics* 9 (2): 148-168.

Turner, C.E. 2000. Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *Canadian Modern Language Review, 56(4),* 555- 584.

Upshur, J. A. and Turner, C. 1999. Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing.* 16 (1) 82-111.

Van Lier, L. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23:* 489—508.

Van Lier, L. 2000. From input to affordance: social-interactive learning from an ecological perspective. In Lantolf, J. P. (ed.) *Sociocultural theory and second language learning.* Oxford: Oxford University Press. 245-259.

Vygotsky, L. S. 1978. *Mind in society: the development of higher psychological processes.* Cambridge, MA: Harvard University Press.

Vygotsky, L. S. 1987. Thought and speech. In Rieber, R. W. and Carton, A.S. (eds.) *The collected works of L. S. Vygotsky: Volume 1.* New York: Plenum, 243-285.

Weir, C. J. 1988. The specification, realization and validation of an English language proficiency test. In Houghs, A. (eds.) *Testing English for university study.* Modern English publications/British Council, London pp.45-110.

Wesche, M. 1992. Performance testing for work-related second language assessment. In E. Shohamy and R. Walton (eds.) *Language assessment for feedback: testing and other strategies.* Kendall/Hunt Publishing Company. 103-122.

Widdowson, H. G. 1979. *Explorations in applied linguistics.* Oxford: Oxford University Press.

Widdowson, H. G. 1989. Knowledge of language and ability for use. *Applied Linguistics.* 10/2: 128—37.

Wigglesworth, G. 1997. An investigation of planning time and proficiency level on oral test discourse. *Language Testing.* 14 (1): 101-122.

Wigglesworth, G. 2001. Influences in performance in task-based oral assessments. In Bygate, M., Skehan, P., and Swain, M. (eds.) *Researching pedagogic tasks: second language learning, teaching and testing.* New York: Pearson Education, 186—209.

Young, R. and He, A. W. (eds.) 1998. *Talking and Testing: Discourse Approaches to the Assessment of Oral Proficiency.* Amsterdam: John Benjamins Publishing Company.

Young, R. 2000: Interactional competence: challenges for validity. Paper presented at the Language Testing Research Colloquium. Vancouver, Canada, March.

Appendix A. Band Score Criteria

1. Overall Result Band Score Criteria

Band Score	Score Criteria
40	Marginally Competent User Demonstrates uneven control in using academic English. Fluency, accuracy, and flexibility are impediments to overall competence in the academic setting

2. Speaking Performance Band Score Criteria

Band Score	Score Criteria
30	Limited Speaker Speaks with some difficulty; hesitations or false starts Mispronounces some words Searches for words or provides studied and careful responses
40	Marginally Competent Speaker Speaks with some fluency and flexibility Speed of response (either too fast or too slow) sometimes limits communication
50	Competent but Limited Speaker Speaks with some fluency and flexibility Speaks unevenly—at times there is a natural and easy quality to the response and at other times the response breaks down
60	Competent Speaker Speaks fluently, flexibly and with a degree of ease Compensates strategically for limitations Communicates most required information clearly

Appendix B. Prompt instruction for Task A and Task B

Prompt Instruction

Task 5. Often in a university course, you're required to work in a small group to prepare an oral presentation. In this task, you'll be asked to participate in a small group activity. Please look at Task #5 on your handout. This is the handout that you professor has given you describing a group project. Listen as he or she gives you an instruction about the group project and an oral presentation:

[Professor:] *Ok, now, as you can see we'll be looking at the issue of 'youth and employment'/'violence in society'. You'll be looking at several different aspects of 'youth and employment'/'violence in society' and presenting your information to the class in the form of a group oral presentation. I'll put you in groups of three. When you meet with your groups today, you should examine the topics of 'youth and employment'/'violence in society' listed on your handout sheet. Each of you must then decide which topic of 'youth and employment'/'violence in society' you'd like to prepare.*

Take a few moments now to read over the topics. Now take one minute to familiarize yourself with the topics and the details from the list on the handout.

[One minute pause] Now listen to the other two students in your group talking about their topics and why they have chosen them. After they finish speaking, they'll ask you to choose another topic from the two remaining topics and to explain your choice. You'll be given a short time to decide and then you'll be asked about your decision. Listen as the two students talk about their decisions.

[Recording of two "students" talking about their choices of the topics]

OK. So that leaves two other topics for you. Which one would you like to do and why?

Now, take a short time to decide which topic you would like to talk about. Be sure to choose from the two remaining topics. Don't choose one of the topics that have already been discussed. You have one minute to plan your answer.

[One minute pause] OK. Now tell your group which topic you have chosen and explain your choice. You'll have up to one minute to discuss your topic. Begin speaking after the beep.

Appendix C. Handout for Task B

TASK *5:*

In this task you will be asked to talk about your choice of topic for a group project.

HANDOUT

1300 A – Violence in Society
Instructor: James Woods

GROUP PROJECT WORK SHEET

VIOLENCE IN SOCIETY

Your group will give an oral presentation on each of the following aspects of *violence in society.* Each member of the group will choose **one topic** listed below.

Choose your topic from the list below:

TOPIC	DETAILS
Violence in movies	X effect on children X effect on male/female relationships
Violence in families	X effect on children X problems with elderly people
Violence in the schools	X effect on learning X weapons control in high schools
Violence on the streets	X effect on ordinary citizens X cost for police, etc.

END OF TEST

Appendix D. Questionnaire

We would like to know your reactions to the test. Please answer the questions in as much detail as possible. This will assist us in our research. Your cooperation is appreciated.

A. Please complete these details:

Name_____ Sex_____ Native country_____

Now you have taken two tests: two tasks in different formats. The two tasks you have undertaken are:

Task A: Youth and Employment
Task B: Violence in the Society

The two formats are:

Format 1: Computer-based test (as what you did on the CAEL Assessment)
Format 2: Small group discussion test (You had a group discussion in the classroom before making your presentation on the computer.)

Please specify which task you have undertaken in the two formats:

Computer-based test_____; small group discussion test _____
A. Youth and employment B. Violence in the society

*B. Please note the following questions are asked for the two **TASKS** you have undertaken.*

Task A: Youth and Employment
Task B: Violence in the Society

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

1. I believe that Task A would provide an examiner with an accurate idea of my ability to speak English.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

2. I believe that Task B would provide an examiner with an accurate idea of my ability to speak English.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

3. I believe I did well on task A.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

4. I believe I did well on task B.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

5. I had enough opportunity to show my ability to speak English in doing Task A.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

6. I had enough opportunity to show my ability to speak English in doing Task B.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

7. I had enough time to do task A

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

8. I had enough time to do task B.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

9. I am familiar with the content of Task A.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

10. I am familiar with the content of Task B.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

11. I believe Task A produces the type of language required of students in studying a university course.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

12. I believe Task B produces the type of language required of students in studying a university course.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

13. Please rate Task A for difficulty.

easy ----------▶ difficult

1 2 3 4 5

14. Please rate Task B for difficulty.

easy ----------▶ difficult

1 2 3 4 5

*C. Please note the following questions are asked for the two **FORMATS** of the test you have taken.*

Format 1: Computer-based test (as what you did on the CAEL Assessment)
Format 2: Small group discussion test (You had a group discussion in the classroom before making your presentation on the computer.)

15. I understood what I was supposed to do during the Format 1 test.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

16. I understood what I was supposed to do during the Format 2 test.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

17. I believe that Format 1 would provide me with an opportunity to show my ability to speak English.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

18. I believe that Format 2 would provide me with an opportunity to show my ability to speak English.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

19. I felt nervous while I was doing the Format 1 test.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

20. I felt nervous while I was doing the Format 1 test.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

21. I like Format 1.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

22. I like Format 2.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

23. Format 1 makes the test more difficult.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

24. Format 2 makes the test more difficult.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

25. Format 1 is a fair form of test.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

26. Format 2 is a fair form of test.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

D. *Your preference of the test*

27. If you were going to take an oral test in an examination, which one of the three tests would you prefer to take? Put a '1' next to the task you would prefer most, a '2' next to your second choice, and a '3' next to your third choice, and a '4' the test you would least like to take.

Task A in Format 1

Task A in Format 2

Task B in Format 1

Task B in Format 2

Please add any other comments you wish to make: