AN INVESTIGATION OF TWO TYPES OF QUESTION PROMPTS IN A LANGUAGE PROFICIENCY INTERVIEW TEST AND THEIR EFFECTS ON ELICITED DISCOURSE

Christian Colby

Department of Second Language Education
McGill University, Montreal

October, 2001

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfilment of the requirements of the degree of Master of Arts.

Acknowledgements

Many people gave me their support throughout this project, and I am extremely grateful for it. I would also like to sincerely thank Dr. Carolyn Turner, my thesis supervisor, for her support and guidance, and for her enthusiasm, which inspired me in the 'apprenticeship' of undergoing this research project. I would also like to acknowledge Dr. Turner's help in superb editing, from which I learned so much. Dr. Roy Lyster was generously helpful, by providing insightful guidance regarding how I might draw on discourse analysis in my research concerns. In the recent past, Dr. Jack Upshur of Concordia University, helped me by teaching me basic concepts in language testing which were later further enhanced by Dr. Turner. Fabrice Rouah, of McGill University's Statistical Consulting Service was of exceptional help, in generously guiding and suggesting more proficient methods of analysis of the research data. I wish also to thank Dr. Richard Young for his kind support and for materials sent, and Dr. Ruth Berman for helping me discover her transcription coding manual. I would also like to thank Dr. Gillian Wigglesworth for very kindly suggesting approaches I might be interested in, and also for recommending to me her work on development of a transcription protocol for use with discourse in oral proficiency tests. I thank Dr. Tim McNamara for recent words of encouragement and for having initially inspired the queries leading to the present research, some years ago. I would also like to express my warm thanks to my colleagues at the Montreal office of the Second Language Evaluation Section of the Public Service Commission of Canada who helped and encouraged me, and who were excellent participants in the data collection workshops of the present study, and for Donald Lévesque's translation services. I am deeply grateful for the exceptional support offered me by the managers of the Public Service Commission of the Government of Canada, in both the Montreal and Ottawa offices, who believed in the potential of this research and made it possible for me to undertake it. In this regard, I would especially like to thank Mr. Robert Bisaillon, Mr. Eric Legrand, and Ms. Suzanne Lalonde. Finally, I thank my husband Peter for his support, encouragement and patience.

Abstract

The present research investigates the use of different question prompts and the discourse they generate in the SLE:OI, an ACTFL-variant second language oral proficiency interview test. One hundred and fifty-two question prompts used to elicit the test task of 'supporting an opinion,' were transcribed from 27 SLE:OI tests administered between July and November, 2000. From this, 30 categories of question prompts were identified by 6 SLE:OI raters acting as judges. Independently, the researcher and the judges determined task difficulty/complexity to be the predominant feature differentiating the categories. Using the 30 categories as a basis, the Question Prompt Complexity Questionnaire was produced and administered to the 6 judges. Analysis of the questionnaire data indicated a clear consensus for 3 categories into 'easy' and 'difficult' groups. Subsequently, candidate responses to 11 question prompts from the easy group, and 10 from the difficult group were transcribed, and discourse analyses were carried out to ascertain response levels of L2 fluency (by type-token ratio; frequency of silent and filled pauses, repetitions, and self-repairs), accuracy (by verb morphology and lexical use), and complexity (by clause subordination). The results demonstrated that those candidates tested with 'easy' and 'difficult' question prompts showed strong, significant differences in two aspects of their response fluency, but no significant differences in the accuracy or complexity of their responses. Based on these findings, several recommendations and implications for rater training were cited.

Résumé

La présente recherche vise à étudier l'usage de différentes questions et le discours qu'elles génèrent dans le cadre de l'ELS : IO (Examen de langue seconde : Interaction orale) qui un test oral de compétence en langue seconde s'apparentant à l'ACTFL. Cent cinquante-deux questions utilisées pour évaluer la fonction languagière « soutenir une opinion » ont été extraites de 27 tests d'ELS : IO administrés entre juillet et novembre 2000, puis transcrites. À partir de là, 6 évaluateurs d'ELS : IO agissant comme juges ont relevé 30 catégories de questions. De façon indépendante, le chercheur et les juges ont établi que les aspects difficulté/complexité de la fonction seraient les éléments les plus importants pour distinguer les différentes catégories de questions. Prenant pour base les 30 catégories, un questionnaire sur la complexité des questions a été administré aux 6 juges. L'analyse des données du questionnaire a permis d'établir un consensus clair au sein du groupe pour 3 catégories afin de créer les groupes « facile » et « difficile ». Par la suite, les réponses des candidats à 11 questions de la catégorie « facile » et 10 questions de la catégorie « difficile » ont été transcrites et on a procédé à une analyse du discours pour vérifier les niveaux de réponse sur le plan de la facilité d'élocution en langue seconde (« type-token ratio » et la fréquence des silences et des pauses remplies, les répétitions, l'auto-correction); de la précision (la morphologie des verbes et la précision lexicologique); et de la complexité (les propositions subordonnées). Les résultats ont démontrés que les candidats évalués à l'aide des questions « facile » et « difficile » présentaient des

différences très significatives notamment au niveau de deux aspects de l'élocution mais non pas au niveau de la précision et de la complexité de leurs réponses. Basées sur ces observations, plusieurs recommandations qui auront des implications sur la formation des évaluateurs ont été formulées.

Table of Contents

Page

Acknowledgments.....iii

Abstract.....iv

Résumé.....v

Table of Contents.....vii

List of Tables.....xii

List of Figures.....xiv

Chapter

1. Introduction.....1

Research rationale.....9

2. Literature Review

Origins of ACTFL performance tests.....11

The Proficiency Movement and communicative competence.....17

Reliability defined.....21

Constraints to test reliability.....21

Validity defined.....24

Construct validity defined.....25

Constraints to construct validity.....25

The ACTFL proficiency scale and second language acquisition studies.....25

The ACTFL Guidelines: construct by intuition.....28

Content validity defined.....29

Constraints to content validity.....29

The validity of ACTFL test task assumptions.....29

The validity of the hierarchal sequencing of test tasks.....30

Task authenticity: Oral proficiency tests as 'conversation'.....31

Test tasks and method effects.....32

Calls for empirical research.....36

3. Purpose and Design of the Study.....41

Hypotheses and research questions.....41

Context of the study.....48

Participants.....49

Test candidates.....49

Test raters.....51

English as a second language teacher.....51

Instruments.....51

The SLE:OI oral proficiency interview test.....51

The Question Prompt Categorization Grid.....53

The Criteria for Determining Task Difficulty document.....54

The Question Prompt Category Complexity Questionnaire.....55

Procedure.....55

Procedure: Phase 1.....56

Initial data collection.....56

Delimitation of question prompts.....59

Protocol of question prompt transcription.....62

Workshop 1 Preparatory Categorization trials:

Piloting the methodology.....64

Workshop 1: Protocol of question prompt

categorization.....66

Workshop 2 preparations: Development of the Criteria

for Determining Task Difficulty document 67

Workshop 2 preparations: Creation of the Question

Prompt Category Complexity Questionnaire.....70

Workshop 2: Protocol of questionnaire administration....74

Analysis of the Question Prompt Category

Complexity Questionnaire responses.....75

Procedure: Phase 2.....77

Identification of question prompts issuing from

2-group consensus.....77

Transcription of candidate responses:

The response idea unit (RIU).....78

Analysis of candidate responses: Discourse analysis protocol.....80

Fluency protocol selection rationale.....80

Fluency: Type-token (TTR) measure.....81

Fluency: Fluency feature frequency (FFF) measure.....82

Accuracy measurement.....82

Complexity measurement.....83

4. Presentation and Discussion of Results: Phase 1: Qualitative analyses....85

Workshop 1 preparatory Trials 1, 2, and 3:

Piloting the methodology.....85

Workshop 1: Production of the Question Prompt

Categorization Grid.....91

Workshop 2: The Question Prompt Category

Complexity Questionnaire.....100

5. Presentation and Discussion of Results: Phase 2: Quantitative analyses101

1) Question Prompt Category Complexity Questionnaire:

Consensus identification.....101

2) Discourse Analysis.....105

Analysis of fluency: Type-token ratio.....106

Analysis of fluency: Fluency feature frequency.....108

Analysis of accuracy.....114

Analysis of complexity.....118

6. Conclusions.....122

Introduction.....122

The issue of parallel test forms..... 122

The research question and the research findings..... 125

Implications and recommendations..... 128

Limitations of the study..... 132

Suggestions for further research..... 135

Concluding remarks..... 136

References..... 139

Appendices..... 151

Appendix A - Informed consent to participate in research..... 153

Appendix B - Consentement à participer à la recherche..... 155

Appendix C - Certificate of Ethical Acceptability..... 157

Appendix D – Criteria for Determining Task Difficulty Document..... 159

Appendix E - Question Prompt Category Complexity Questionnaire..... 161

Appendix F – The Response Idea Unit Transcription Coding Protocol..167

Appendix G – Conventions of the Simplified Analysis of Speech Unit..169

Appendix H – Workshop 1 and 2 results: Question prompts and Headings..... 173

Appendix I – Analysis of responses to Question Prompt Category Complexity Questionnaire..... 177

Appendix J – SLE:OI Test Information..... 183

List of Tables

Page

Table

1. Criteria for Identification and Selection of Question Prompts.....63

2. Trial 1: Categorization of 152 question prompts after 10 minutes.....86

3. Trial 2: Categorization of 152 question prompts after 3 hours.....87

4. Trial 3: Categorization of 10 question prompts after 5 minutes.....88

5. Workshop 1: Question Prompt Categorization Grid: Final categorization of 152 question prompts in three attempts of 10 minutes.....93

6. Type-token ratio per response idea unit of the *easy* and *difficult* groups.....106

7. Type-token ratio group means and standard deviations.....107

8. Fluency frequency features per response idea unit: *Easy* group.....108

9. Fluency frequency features per response idea unit: *Difficult* group.....109

10. Chi square contingency table for silent pause effect.....110

11. Total pause time in seconds: Group means and standard deviations.....113

12. Total pause time per response idea unit of the *easy* and *difficult* groups.....114

13. Accuracy of verb morphology in occurrences per response idea unit in the *easy* group.....115

14. Accuracy of verb morphology in occurrences per response idea unit in the *difficult* group.....115

15. Chi square contingency table for subject-verb agreement effect.....116

16. Accuracy of a lexical form in occurrences per response idea unit in the *easy* and *difficult* groups.....118

17. Syntactic analysis of clause structures in the *easy* group.....120

18. Syntactic analysis of clause structures in the *difficult* group.....120

List of Figures

Page

Figure

1. Possible outcomes of task difficulty on performance: in C and B/C borderline candidates.....46
2. Possible outcomes of task difficulty on performance: in B/C borderline candidates.....46
3. Analysis of question prompt complexity questionnaire responses.....103

Chapter 1

Introduction

Test fairness is a pivotal concern for many, particularly in Quebec, Canada where French and English are widely spoken and evaluated. Thus a judicious second language test could accurately assess whether a Quebec student should graduate from secondary school, if or how a foreign student should study there, whether international teaching assistants should teach in provincial institutions, if doctors and other professionals should practice in Quebec, and in both private industry and in the public domain, if a candidate for a bilingual position should be employed.

Fairness in second language (L2) testing is also a primary concern of many throughout officially bilingual Canada, and in the Canadian Federal Public Service where thirty percent of existing positions have been designated as bilingual. These positions require varying standards of French or English L2 competence. The President of the Public Service Commission affirmed this role in a recent speech to the Public Accounts Committee by stating that the "The PSC is an **independent Parliamentary agency** [sic] which ensures that staffing and recruitment for the Public Service are conducted according to the principle of merit" (Serson, 2001, p.1). Thus, potential employees can only be legally engaged once their second language abilities have been determined to adequately reflect those required by the target position. In view of this, Canada's Federal Public Service Staffing Directorate has an official mandate based on the concept of

fairness in hiring practices extending to the use of accurate second language testing instruments.

Some twenty years ago, the Canadian Government sought to ameliorate its second language oral testing instruments. The Testing Directorate of the Canadian Public Service opted to use a second language test battery which includes a proficiency interview test to assess the oral abilities of both employees and potential employees. Proficiency interview tests (hereafter called proficiency tests) draw on a conversational format in structured interviews intended to assess oral second language performance.[1] In 1984 the Second Language Evaluation: Oral Interaction (SLE:OI) oral proficiency test was adapted and launched as the Canadian Government's test of second language oral ability.

The SLE:OI is of the lineage of oral proficiency tests of the American Council on the Teaching of Foreign Languages (ACTFL) (see Chapter 2). In fact, the SLE:OI was modelled on the Oral Proficiency Interview (OPI) test, which was developed by an ACTFL agency, the American Interagency Language Roundtable.[2] Thus, in the language testing (LT) literature, the OPI is identified as an ACTFL test. Those tests closely associated with ACTFL tests or derived from them, such as the SEL:OI, are termed ACTFL-variant tests. ACTFL proficiency tests are generally administered by two or three raters and interviewers. The

---

[1] Proficiency tests are also known as performance tests in the language testing literature.

[2] Thus, the SLE:OI and its French version, the 'Evaluation de langue seconde: interaction orale' (ELS:IO), are comparable to the OPI. The OPI, in turn resembles the American Foreign Service Institute's oral proficiency test, known as the FSI – OPI.

SLE:OI, as an ACTFL-variant test, differs from ACTFL tests in that is administered by only one, highly trained rater-interviewer, who accomplishes the two tasks of interviewing and simultaneously rating the interaction.[3]

In an attempt to give each candidate an equal opportunity to succeed, or 'bias for best' (Swain, 1985), the SLE:OI test employs various procedures. Raters of the SLE:OI are carefully chosen and trained in an extensive 5-week training programme which, in the tradition of ACTFL-variant tests, is unusually long and thorough. In contrast, OPI testers require as little as six days of training to qualify as certified OPI testers (American Council on the Teaching of Foreign Languages [ACTFL], 2001). The SLE:OI test requires raters to not only interview and rate candidates in test administrations, but also in normal circumstances to be able to render a score and write a lengthy, detailed defence of the score upon conclusion of the interview.[4] SLE:OI raters re-train in standardizing sessions once a month. The expected result of all of these measures is high test validity overall and fairness; certainly the use of a less than judicious test or testing procedures could result in inexact assessments of examinee second language ability, potentially jeopardizing the present and future employment opportunities of many.

In addressing the testing community of the Canadian Public Service Commission, McNamara (1995b) commented on the overall merit of the SLE:OI oral proficiency test, noting that 'from a practical point of view, the procedures

---

[3] The term rater will hereafter be used to refer to rater-interviewers, as the term pertains to SLE:OI raters, in the present research.

[4] For this reason, training to become an SLE:OI rater is rumoured to be the most gruelling and demanding in the Canadian Federal Public Service.

that you have set in place are exemplary" (videocassette recording of lecture).

Nonetheless, serious concerns have also been raised about the validity and reliability of ACTFL and ACTFL-variant second language proficiency tests by McNamara, Bachman and others (Bachman, Davidson & Milanovic, 1996; Bachman, Lynch & Mason, 1995; Bachman & Savignon, 1986; Jacoby & McNamara, 1999; Lantolf & Frawley, 1985, 1988, 1992; McNamara, 1995a, 1995b, 1996, 1997; McNamara & Adams, 1991; McNamara & Lumley, 1995, 1997 (see Chapter 2 for a discussion of this issue). Moreover, other researchers (e.g., Matthews, 1990; van Lier, 1989) have also questioned elements of proficiency test theory and practice adding their own perspective as practitioners, having worked as oral proficiency test raters themselves. For example, Matthews raised concerns regarding proficiency test task sequencing. Van Lier questioned the assumption that language in oral proficiency interviews approaches that of natural conversation (see Chapter 2).

Having myself worked for ten years in the Canadian Federal Public Service as an oral interaction rater, and in my capacity as a graduate student, I have been in the position of both using an ACTFL-variant proficiency test and of studying ACTFL and ACTFL-variant tests. My job consisted of administering twenty-five oral proficiency tests each week to civil servants being considered for bilingual positions, and to potential civil servants. This situation allowed me to see the inherent value of prudence in particular testing practices. Yet like Matthews (1990) and van Lier (1989), the experience has led me to question certain theoretical and practical aspects of oral proficiency test administration. For

example, although I noticed that SLE:OI test trainers placed a great deal of emphasis on standardization of test content generated by raters, nonetheless a great deal of variation occurred in practice due to the inherent conversational format of the test structure. Consequently, an interest arose as to the effect of variation in test prompts on elicited responses. Moreover, I wondered how this variation might influence the kind of discourse produced in ACTFL and ACTFL-variant tests.

The strength of oral proficiency tests is that they can closely approximate authentic language use, having been designed to simulate natural conversation. Therefore their validity is enhanced by the proximal authenticity of their content. By their nature they consist of mostly spontaneous and changeable language content on the part of interviewers and examinees. The resulting discourse of oral proficiency tests are often said to be unpredictable.

However, this very unpredictability remains a source of potential unreliability since raters regularly employ varied question forms, which are essentially alternate test forms. These alternate forms, occurring spontaneously in the conversational format of the tests and therefore not previously measured for equivalence, threaten to an indefinite extent the reliability of the testing instrument.

Consider if it were possible for interview tests to employ the exact same examiner language across tests. Then variation in the language of the rubric or question prompt would be nonexistent, and this particular threat to the tests' reliability would be nonexistent. However unpredictability and question prompt

variance occur as an inchoate element of the conversational format of oral proficiency tests, after all, they are intended to measure L2 proficiency in unplanned speech. In addition, test administrators welcome the element of unpredictability since it discourages candidates from knowing test questions in advance in order to rehearse or memorize speech samples prior to test administration. Nevertheless, the factors of test reliability (in terms of using the same measure across test administrations), and construct validity (in terms of ensuring that the construct intended to be measured is in fact the one measured), become an important concern given that there is considerable variation in test forms when different question prompts are used across test administrations (see Chapter 2).

Test method refers to how a test is done, and this is defined by the test task or tasks. These in turn are specified to the candidate in part of the test rubric, the instructions. Bachman and Palmer (1996) cite three components of test instructions: the language of instructions; the channel or mode used such as aural, visual or both; and the specification of procedures or tasks. Changes to any of these components result in changes to the test method, known as method effects.

In a discussion of the influence of method effects, Bachman, Davidson, and Milanovic (1996) caution that "It is now well understood that aspects of test methods can have an important effect on performance on language tests and it would thus seem imperative to incorporate information about the characteristics of test methods explicitly into the design of language tests" (p. 126).

Consequently, the question arises as to whether method effects influencing performance would have enough impact to alter candidate scores.

Other test administrators have also raised the issue of method effects in oral proficiency testing. The developers of the Cambridge Assessment of Spoken English (CASE) have broached the problem of the potential unreliability of varying question prompts in the CASE oral proficiency interview test by prescribing both the wording and the order of question prompts in its procedural agenda for its examiner-interviewers, the CASE Interlocutor Frame (Lazaraton, 1996). In this manner the CASE test has integrated controls of its question prompts for both the method and the order effects.

The issue of whether alternate prompt forms may be considered to be equivalent in oral proficiency tests is further complicated by the widespread practice of employing questions based on topics suggested by test candidates themselves. The rationale behind this practice is to allow each test to be tailored to the individual candidate in question. Consequently, the interaction or test content depends to some extent on input from the candidate. ACTFL (2001) has informed candidates of the OPI test of the practice, wherein they note that "the topics discussed during the interview are based on the interests and experiences of the speaker." (p.1). Unlike the CASE administrators, the ACTFL overseers have not controlled for method or order effects as can be seen in their additional information that "There is no script or prescribed set of questions" (ACTFL, online, retrieved April 19, 2001). The merit of adapting the test to candidates in

this way, possibly enhances test validity in the eyes of the candidate. Nonetheless, the impact of this on test reliability and equivalence of forms may be questioned.[5]

In a discussion of the ACTFL-OPI and ACTFL-variant oral proficiency tests, Lazaraton (1996) has lent further weight to the argument that consistency in question prompts is important, as noted in the following comments:

> In fact, the achievement of consistent rating is highly dependent on the achievement of consistent examiner conduct during the procedure, since we cannot ensure that all candidates are given the same number and kinds of opportunities to display their abilities unless oral examiners conduct themselves in similar, prescribed ways. (p. 19)

Nevertheless the SLE:OI oral proficiency test, like many ACTFL variants, uses a range of individually adapted question prompts in the task of '*supporting an opinion.*' The question prompts and the language elicited are presented as equivalent forms. Recently researchers have cast doubt on the validity of this assumption, notably Bachman and Palmer (1996) who observe that in changing the topic in a task, the result is a new task or method, which they consider to be effectively, a new test. Furthermore, Wigglesworth (1997a) studied the effect of task type on candidate discourse in an oral proficiency test, and found differences in candidate discourse as a function of task. She concluded that "the findings of this paper, whilst remaining speculative, do point to the importance of routinely

---

[5] Face validity is the formerly used term for the appearance of validity, particularly from the perspective of test candidates; it is now in disfavour among testing specialists. See Bachman and Palmer (1996, pages 29, 42); and Bachman (1990, pages 285-9) for a discussion of face validity and its limitations.

subjecting test data to rigorous discourse analysis, and to integrating discourse analysis into the process of test validation." (p. 47)[6]

Research rationale

As noted earlier, I have worked for ten years in the Canadian Federal Public Service as an oral interaction rater, administering twenty-five SLE:OI, ACTFL-variant proficiency tests per week to civil servants and potential civil servants being considered for bilingual positions.

Thus, as a testing practicioner I have become concerned with the question of method effects and how they pertain to test fairness for all candidates. I have come to question the conjecture that candidates tested with different prompts or alternate test forms, may in fact be tested with equivalent forms. This is due to the fact that oral proficiency tests regularly employ very different questions and topics and therefore methods, wherein the alternate forms have not been empirically proven to be equivalent.

The present research examines the use of different question prompts (and therefore different methods and alternate test forms) in an oral proficiency test, and the elicited responses to them. Moreover it seeks to investigate the nature of elicited discourse in oral proficiency tests where different question prompts have been employed, by means of discourse analysis of candidate responses. The present study focuses on the discourse produced in the task of *supporting an*

---

[6] Discourse has been defined (Richards, Platt & Platt, 1992) as "a general term for examples of language use, i.e. language which has been produced as the result of an act of communication" and discourse analysis has been defined as "the study of how sentences in spoken and written language form larger meaningful units such as paragraphs, conversations, interviews, etc." (p. 111).

opinion in an oral proficiency test in an effort firstly, to investigate whether method effects may have affected candidate performance, and secondly, to investigate the nature of these effects.

Chapter 2 is comprised of an historic overview of the development of ACTFL oral proficiency tests in the North American context, with the rise concurrently of the Proficiency Movement and of conflicting views of communicative competence. A discussion of constraints to test reliability and validity is also included in the chapter. Chapter 3 describes the rationale and design of the present study. Analyses and discussion of the qualitative and quantitative results are found in Chapters 4 and 5 respectively. Chapter 6 addresses the conclusions of the study, citing limitations, implications, contributions, and recommendations.

Chapter 2

Literature Review

Origins of ACTFL performance tests

An understanding of the rationale behind the design of ACTFL-based oral proficiency interview tests might best be addressed by first considering the context in which they were created.[7]

In the United States it is widely regarded that the demand for a valid interview test of second language oral communication first arose as a result of operational needs in the American military forces during the Second World War (Clapham & Corson, 1997; Lowe, 1983; Spolsky, 1995). Studies by Kaulfers in1944, and Angiolillo in 1947 (as cited in Spolsky, 1995) described U.S. Army programmes which attempted to devise effective language tests. Kaulfers held that L2 tests should provide 'evidence of the examinee's readiness to perform in a life-situation,' and to be scored using 'a kind of ladder' indicating 'performance norms' (as cited in Spolsky). While Kaulfers' proposals were set aside during the war, his novel work nevertheless served to influence subsequent thinking about language testing.

Following the war the American Foreign Service Institute (FSI) developed

---

[7] I confine my discussion to the American arena for two reasons; firstly, the current research deals exclusively with an ACTFL variant oral proficiency test, therefore other oral proficiency testing traditions are not under consideration here, and secondly due to a fact that Fulcher has articulated in noting that, "most work on the testing of speaking has been done in the U.S., and most developments in other countries are based on the American OPI and on rating scales whose ancestor was the FSI" (cited in Clapham & Corson, 1997, p. 77-8).

an interest in improving the speaking skills of diplomats and employees working abroad. This plan included the employment of a valid speaking test. Accordingly, the American State Department took the first step and in 1952 consigned the Civil Service Commission to document an inventory of the foreign language abilities of its personnel. The result was the conception of a 6-band scale of language proficiency descriptors. Lowe (1983) has cited problems with the initial band descriptors owing to the fact that their employee self-assessments included vaguely defined constructs such as being fluent or bilingual. In addition to these difficulties, problems of potential bias in the new test were documented in a study by Sollenberger (1978); it was found in early military test administrations, that scores were affected by the rank and age of the officers tested (cited in Clapham & Corson, 1997).

Consequent to perceiving the descriptor band problems, the State Department responded by commissioning a needs analysis of tasks accomplished by the above-mentioned employees. This evolved into the creation of the Government Definitions, one-line occupational designations which were later expanded in order to be used as a basis for test guidelines (see Lowe, 1983 for a discussion of this).

In 1956 the first American oral proficiency interview test was trialled with State Department employees, becoming mandatory for all Foreign Service Officers a year later. In 1958 the FSI had officially developed its analytic test rating scale, which ranged in scores from 0 to 5, with 'plus' ratings (0+, 1+ and so

on), totalling eleven rating descriptors. At that time user expectations about the test and the rating scale were quite high:

> Confidence in the new FSI testing system was extremely high because of the reported accuracy of measurement, even though it was acknowledged that a test score was only a predictor of effective communication, and not a direct measure of the ability to speak.
>
> (Sollenberger, cited in Clapham & Corson, 1997, p. 76).

Sollenberger's claim concerning the test's predictive validity is an accurate assessment of how ACTFL and ACTFL-variant language tests function. (This is in contrast to the less-informed view of some who mistakenly assert that the tests are a direct measure of proficiency constructs.) Accordingly, Carroll (1961 [1972: 319]) has expanded on this by noting that the validity of a proficiency test entails "not solely...a good sample of the English language but...whether it predicts success in the learning tasks and social situations to which the examinees will be exposed" (cited in McNamara, 1996, p.29). However, more recently Bailey (1998) has cited the strengths of performance tests as they are "direct, authentic, and highly contextualized" (p.215).[8]

She added that:

> This is because their very design depends on using stimulus materials and posing tasks to the learners that are based directly on the learners' intended (or hypothesized) use of the target language. (p.215)

---

[8] As noted earlier, the term 'performance test' is analogous to 'proficiency test' (see Introduction). Bailey (1998) prefers the former.

Bailey's (1998) positive assessment echoes those of test users of the early 1950's, as reported by Sollenberger, above. At that time proficiency testing provided newer, more authentic methods of determining second language speaking ability. (The issue of how authentic ACTFL oral proficiency tests actually are, however, has been disputed by several researchers [for example Lantolf & Frawley, 1988; Lazaraton, 1992; and Lewkowicz, 2000].)

Over time, from the 1950s onward, the FSI proficiency test became the testing instrument of the Central Intelligence Agency (CIA), the Federal Bureau of Investigation (FBI), and the Defense Language Institute (DLI). And in the late1960's and 1970s the American Peace Corps entered into an agreement with the U.S. Department of Education's Educational Testing Service (ETS) for the development of language training and testing materials including an FSI-type test.[9]

In 1972 the Peace Corps, with the CIA and the DLI came together in 1972 to create the Interagency Language Roundtable (ILR). The ILR included a Testing Subcommittee with the mandate to coordinate the research and development of language tests in U.S. Government language schools (Jones & Spolsky, 1975). In 1975 it was reported that yearly U.S. Government L2 testing amounted to over seven thousand people tested in approximately sixty languages (Jones &

---

[9] ETS itself was founded in 1947 to accommodate the testing needs of the American Council on Education, the Carnegie Foundation for the Advancement of Teaching, and the College Entrance Examination Board. Today it is purported to be the largest private educational testing and measurement organization in the world, developing and administering over 11 million tests annually (ETS, online, retrieved July 30, 2001).

Spolsky).

With the advent of increased travel of foreign students to study in American universities and in recognition of the growing need for uniform second language testing in the American public school system, then President Carter requested a report of the state of foreign and international studies (Fulcher, cited in Clapham & Corson, 1997). Thus one of the recommendations of the ensuing 'Strength through Wisdom: A Critique of U.S. Capability' report of 1979, was the creation of a standard American L2 testing system.

Consequently, the pedagogical division of the American Modern Language Association, ACTFL, and ETS worked to establish a uniform rating scale for both government and academic use, based on the ILR rating scale. The endeavour was called the Common Metric Project.[10] Principal among the Project findings was the observation that the existing ILR scale lacked sufficient lower levels to justify its use in secondary schools and in universities; in other words, it did not discriminate finely enough at the lower levels (Lowe, 1983). Consequently, the ILR rating scale was adjusted to include 3 new subranges at Levels 0 and 1; with 2 levels at Level 2; while the highest level, Superior, was expanded to contain the previous levels 3 through 5. Provisional ACTFL Guidelines including the scale were published in 1982, but it was later, in 1986

---

[10] Lowe (1983) uses this name for the project, citing a 1981 ETS document, *A common metric for language proficiency*. Peckham (online, retrieved November 9, 2000), however refers to 'The Common Yardstick' project. In all probability both authors are referring to the same thing.

that the official and final version of the ACTFL rating scale was published as the ACTFL Proficiency Guidelines.[11]

The ACTFL Speaking Guidelines were further updated in a revision project in 1999 in which certain changes to the rating scale reflected a desire to refine and include greater definition of L2 speaking ability, particularly in the higher levels. Furthermore, ACTFL researchers have defined the rationale behind the revision in the following:

> The purposes of this revision of the Proficiency Guidelines – Speaking are to make the document more accessible to those who have not received recent training in ACTFL oral proficiency testing, to clarify the issues that have divided testers and teachers, and to provide a corrective to what the committee perceived to have been possible misinterpretations of the descriptions provided in earlier versions of the Guidelines.
>
> (Breiner-Sanders, Lowe, Miles & Swender, 2000, p.14)

It is evident that the ACTFL methodology of proficiency testing arose as a result of practical considerations and operational needs. It is less evident that the needs analyses on which test tasks are based was a precise and accurate reflection of target language use (*TLU*), as defined by Bachman and Palmer (1996). This apparent lack would seem to result from the procedural imprecision of basing the *TLU* needs analysis on questionnaire responses from inexpert respondents, as was done with the American Government Definitions on which ACTFL test tasks

---

[11] Thus the new rating scale, jointly adopted by ACTFL and ETS, is occasionally referred to as the ACTFL/ETS rating scale.

were later based.

This is in contrast to the more recently developed Australian performance test, the Occupational English Test (OET). McNamara (1996) recounts that the 1987 – 1989 OET test development resulted from meticulous investigations of various *TLU* domains in specific field domains of the medical profession. The OET task designations arose from consultation at various stages with (a) professional educators for each profession involved, (b) overseas educators for the specific professions, and (c) specialized English as a second language (ESL) teachers experienced in teaching students of particular areas of the health profession (McNamara, 1996). Moreover, following that the data from the above informants was carefully processed and tabulated to reflect *TLU* for various professional domains.

The historic evidence of ACTFL performance testing points to another problematic issue in the test design, that of lack of basis in linguistic theory. The following section describes how second language linguistic ability was viewed in the 1950s.

The Proficiency Movement and communicative competence

With the emergence of oral proficiency testing from the early 1950s onward, a new faction of thought arose, called 'The Proficiency Movement.' The movement proponents saw the need to replace traditional, discrete-item approaches to testing speaking ability with the more naturalistic one of ACTFL. Clark (1988) noted the new movement 'placed a premium on the accurate and reliable measurement of functional language skills, especially listening

comprehension and speaking, within a real-life language use context (p. 187). At the time Lowe (1988) defined proficiency in the following way:

> Proficiency equals achievement (functions, content, accuracy) plus functional evidence of internalized strategies for creativity expressed in a single global rating of general language ability over a wide range of functions and topics at any given level. (cited in McNamara, 1996, p.77)

However, at that time several researchers were influenced by communicative competence models reflective of the distinction between language perfomance and competence, as put forth by Chomsky:

> We thus make a fundamental distinction between *competence* (the speaker-hearer's knowledge of his language) and *performance* (the actual use of language in concrete situations [italics in original].
>
> (cited in McNamara, 1996, 55).

Hymes' (1967, 1972) theory of communicative competence also distinguished between language knowledge and ability for use (cited in McNamara, 1966, p.54-55). Thus again, the division was made between the constructs of *knowledge of language* (realized yet possibly undemonstrated), and *potential to actually use language* (which could be realized or demonstrated). Therefore, this model also differentiated the underlying construct of language knowledge or competence; and that of performance, as it might be demonstrated in language tests such as those of proficiency testing.

Canale and Swain (1980, 1981) proposed a framework for communicative competence that involved *grammatical competence*, (knowledge of rules of

syntax, morphology, lexical use, sentence-grammar semantics and phonology), *sociolinguistic competence*, (cohesion and coherence and appropriateness of language in context), and *strategic competence*, (communication strategies, verbal and nonverbal, used to compensate for communication breakdowns). Later Canale (1983a, 1983b) would revise his interpretation of the theory to include *discourse competence*, which he defined as "mastery of how to combine grammatical forms and meanings to achieve a unified spoken or written text in different genres" (cited in McNamara, 1996, 64).

More recently, Bachman and Palmer's (1996) model of language ability have included *language knowledge* and *strategic competence* where each has been finely defined, or deconstructed. For example, Bachman and Palmer's *language knowledge* includes the following components:

- Organizational knowledge (how utterances or sentences and texts are organized)
- Grammatical knowledge (how individual utterances or sentences are organized)
- Textual knowledge (how utterances or sentences are organized to form texts)
- Pragmatic knowledge (how utterances or sentences and texts are related to the communicative goals of language users and to the features of the language use setting)
- Functional knowledge (how utterances or sentences and texts are related to the communicative goals of language users)

- Sociolinguistic knowledge (how utterances or sentences and texts are related to the features of the language use setting)

In addition, Bachman and Palmer's *strategic competence* includes:

- Goal setting (deciding what one is going to do)
- Assessment (taking stock of what is needed, what one has to work with, and how well one has done)
- Planning (deciding how to use what one has)

Clearly, all of the above has served to elucidate the many facets of language competence as they appear in language learning. In other words, it is no longer sufficient 'to know' a language, one must demonstrate that knowledge through language use. In second language testing, Bachman and Palmer (1996) have defined components of language use and test performance as including topical knowledge, language knowledge, personal characteristics all of which, affect strategic competence. Moreover, Bachman and Palmer also find that strategic competence is affected by setting, and characteristics of the language use or test task.

Consequent to the recent and more refined understanding of the constructs involved in second language testing, apprehensions have been raised regarding the construct of communicative competence (Bachman & Savignon, 1986; Lantolf & Frawley, 1985; Shohamy, 1988, 1990). Yet the ACTFL Proficiency Guidelines assert that they are "not based on a particular linguistic theory or pedagogical method, since the guidelines are proficiency-based, as opposed to achievement-based, and are intended to be used for global assessment" (ACTFL

online, retrieved June 20, 2000). Lantolf and Frawley (1988) have been among the strongest objectors to this comment. Regarding ACTFL's disinterest in linguistic theory, Lantolf and Frawley argue "against a definitional approach to oral proficiency and in favor of a principled approach based on sound theoretical considerations" (p. 181).

Reliability defined

Test reliability has been defined by Genesee and Upshur (1996) as "the consistency of test scores for the same individuals .... A test that yields the same score for a given individual on two separate occasions would be considered reliable" (p. 244). Inter-rater reliability refers to the consistency of assessments made by different raters, or in McNamara's (2000) words, "the extent to which pairs of raters agree" (p. 134).

Constraints to test reliability

The indiscriminate element of the ACTFL approach to test content suggests the need firstly, for greater research into performance test content in order to determine test reliability with more accurate measures than those based on assumptions; and secondly, of ensuring the validity of the constructs on which performance tests are based. Empirical evidence can indicate how closely performance testing approaches its goals reliably and validly assessing second language speech. In fact, only in the presence of empirical evidence can test reliability and validity be asserted with any precision.

As a result of these precise concerns, several researchers (Bachman & Palmer, 1996; Lazaraton, 1996; Madsen & Jones, 1981; Salaberry, 2000; Spolsky,

1995; Stansfield & Kenyon, 1992; Young, 1995a) have addressed issues of reliability in oral performance testing.

However in historical terms, as early as 1890 Edgeworth (as cited in Spolsky, 1995) analyzed inter-rater reliability in assessments of Latin prose essays. Spolsky indicated that by the 1930s, studies of various second language writing tests had also found inter-rater reliability to be problematic in subjective assessments, reporting that "the most patent causes of unreliability were luck in being asked the right question and 'adventitious variation' [sic] in the state of the candidate at the time of the examination" (p. 65).

A hundred years after Edgeworth, research into oral performance test reliability was also primarily concerned with inter-rater reliability. However, in the last 20 years oral performance testing research into test reliability (Lazaraton, 1996; Madsen & Jones, 1981; Salaberry, 2000; Stansfield & Kenyon, 1992; Young, 1995a) has increasingly regarded rater *behaviour* during test administrations as an essential element influencing oral performance test reliability. The behaviour concerned relates to the types of questions raters ask in oral performance tests. For example, Bachman and Palmer (1996) have suggested that "if an extensive set of instructions is used on one form of a test and an abbreviated set on another, test takers' performance on the two forms may be unstable" (p. 139).

In order to address the issue of test reliability, Stansfield and Kenyon (1992) compared two ACTFL-variant proficiency tests, the Simulated Oral Proficiency Interview (SOPI) test involving audiotaped instructions to examinees,

with the Oral Proficency Interview test, in which examinees interact with interviewers. They found the SOPI to be more reliable than the OPI test, given the fact that the instructions were exactly the same over test administrations in the former, as opposed to the latter in which considerable variation in question format occurred.

Young (1995a) cited a similar occurrence when he compared the University of Cambridge Local Examinations Syndicate (UCLES) First Certificate in English (FCE) oral interview test with the OPI. Young noted that "less scripted interview formats such as the ACTFL OPI aim for valid, fluid, interactive assessment but leave open the issue of comparability of interviewer style – an important aspect of reliability (see Shohamy, 1988)" (p. 29).

Lazaraton's (1996) concern about interviewer style and examiner conduct in the CASE led her to advise that follow-up after examiner training should include the use of examinee-interviewer control checklists. The checklists would record examiner adherence to specified question formats as well as the frequency of their 'speech behaviours.' These include repeating answers, rephrasing questions, slowing rate/increasing pitch, and intervening to encourage talk, among others.

Salaberry (2000) has cited reliability concerns as having led several researchers to question the institutional use of the ACTFL Guidelines. In addition, Madsen and Jones (1981) have suggested that many second language teachers have avoided oral proficiency testing citing inconsistent testing methods at their

disposal, and their concerns that the use of these would result in inadequate reliability across test administrations.

As one of several checks for test reliability, Bachman and Palmer (1996) asked "to what extent do characteristics of the test rubric vary in an unmotivated way from one part of the test to another, or on different forms of the test?" (p. 139). It follows from Bachman and Palmer's query that in terms of question content in oral performance testing, it would be preferable if the effects of using different questions were well understood. Thus, any consequent constraints to test reliability could then be minimized.

Validity defined

Validity is concerned with inferences about how appropriately a measure estimates what it is intended to measure. Messick (1989) has defined validity as "an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on scores or other modes of assessment [italics in original]" (p. 13). The concept of test validity may be understood in the question: are we measuring what we think we are measuring?

In terms of language testing, Bachman and Savignon (1986); Lantolf and Frawley (1985); and Shohamy (1988, 1990) have raised questions regarding a priori assumptions involving construct, content, and predictive validity in performance tests, and including the validity of test designs.

Construct validity defined

Bachman and Palmer (1996) have characterized construct validity as "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (p. 21). Messick (1989) clarified the concept further in noting that "the measure is taken to be one of an extensible set of indicators of the construct" (p. 17).

Constraints to construct validity

The ACTFL proficiency scale and second language acquisition studies

The proficiency scale of the ACTFL Guidelines assume the ordered existence of a number of constructs related to second language performance. The Guidelines have been viewed both from the perspective of LT and second language acquisition (SLA) theorists. The Guidelines describe a number of functional language abilities which have been postulated as developing in steady progression in learners as they acquire superior levels of proficiency, gradually moving from weaker to stronger. Thus it is presumed that there exists a continuum of discrete bands of language proficiency which L2 learners exhibit in neat progression through the descriptor bands. However, researchers in both the fields of SLA and in LT have noted the inconsistency of a progressional model in view of observed L2 use (Gatbonton, 1978; Shohamy, 1988; Young, 1995b; Fulcher, 1996a).

Tarone (1998) has defined this feature of L2 language, in the following way:

> *Interlanguage (IL) variation* is the tendency for a second language learner's utterances, produced in the attempt to convey meaning, to vary systematically in grammatical and phonological accuracy as specific situational features change... The term *variation* ought to be reserved to refer to shifts *within* the performance of any given individual and not to differences *across* individuals [italics in original]. (p. 73)

Young (1995b) also noted that 'a majority of longitudinal studies of interlanguage development have shown instead that the interlanguage system goes through a period of restructuring and reorganization. One result of such restructuring is that intermediate stages may be further from the target than either beginning or advanced stages – a pattern commonly called U-shaped behaviour. He also observed that describing learner language is a considerable problem due to interlanguage variation (1989).

In a similar vein, Ellis (1997) also affirmed that 'acquisition follows a U-shaped development; that is, initially learners may display a high level of accuracy only to apparently regress later before finally once again performing in accordance with target-language norms.' Based on the arguments cited, Young (1995b) has observed that ACTFL-based oral proficiency tests do not appear to accurately reflect actual L2 learner ability.

As noted above, the ACTFL rating scales assume a discrete, linear progression in SLA, where learners experience interlanguage progress while developing in step with other language ability components, in a prescribed fashion. Young (1995b) noted that oral proficiency in actuality is a

multidimensional rather than a unitary and linear construct, and he calls for rating scales to reflect this. Young also observed that in view of this, the developers of the CASE have opted to use 11 different scales to accommodate various proficiency components in L2 speech samples (1995).

Elsewhere Young noted that learner interlanguage is 'subject to rules of its own, and is to some extent at least independent of either first or target language. This is what Corder (1967), Nemser (1971), and Selinker (1969, 1972) have variously called "idiosyncratic dialect," "approximative system," or "interlanguage." (1991).

Smith (1989) studied situational context and interlanguage, and documented instances of variation in performance accuracy in general content and field-specific versions of the Spoken Proficiency English Assessment Kit (SPEAK) oral test (cited in Tarone, 1998).[12]

Lantolf and Frawley (1985, 1988, 1992) have voiced harsh criticism of the ACTFL rating scale descriptors, citing the element of the arbitrary in assigning proficiency levels. Thus, these authors observe that while height is metric, colour is scalar in nature, and therefore "the question for OP [oral proficiency] testing is whether or not relevant linguistic behaviour is metric or scalar (1988, n.3, p. 192).

---

[12] These examples of a sentence completion task from Smith's study, are compelling:

1. General topic test response: By saving our money, we will be able to buy a house.
2. Specific topic response: By calibrating your instrument, you should be careful and patient. (cited in Tarone, p. 72-3)

Thus, Lantolf and Frawley suggest that language, as a variable and non-unitary construct, should not be measured as if it were so. In other words, the measuring exercise involved in oral proficiency testing should not be presented as being straightforward and analogous to a calibration of discrete units.

The ACTFL Guidelines: construct by intuition

Several researchers have questioned the fact that the Guidelines not only have been found to inaccurately reflect actual language usage among second language speakers, but that they were devised by arbitrary rather than empirical means (Bachman, 1990; Bachman & Savignon, 1986; Barnwell, 1989; Lantolf & Frawley, 1985, 1988, van Lier, 1989). This is notwithstanding the fact that the descriptors bands were devised from expert judgements (ACTFL, 1986). These judgements arose from expert intuitions at the time, and as such are subject to questioning. Intuition may be a good device in *taking* second language tests; it would seem to follow that *rating* language tests should involve more proven means.

Consider the parallel case of student performances in piano recitals. The performers may decide to perform using intuition to best express the music at hand. However, an expert adjudicator would hopefully not base his or her judgement of the performance solely on intuition, but rather on a thorough knowledge of the construct of musical performance. Similarly, opponents of the Guidelines caution that their use may essentially dictate unrealistic and unfounded candidate performance expectations. For example, Lantolf and Frawley (1985) have posited that ACTFL tests *impose* competencies on the examinees and

measure the extent to which the person deals with the imposition [italics in original]" (p.339). In view of this concern, several researchers have observed actual L2 language use. They have concluded that authentic L2 language use differed from that specified in the ACTFL scale (Bachman & Savignon, 1986; Lantolf & Frawley, 1985; Bachman, 1990; Matthews, 1990; Bachman & Palmer, 1996; Bachman & Cohen, 1998).

Content validity defined

McNamara (2000) has called content validity "the extent to which the test appropriately samples from the **domain** of knowledge and skills relevant to performance in the **criterion** [sic]" (p. 132). It follows that an examination of tennis performance measured by a paper and pencil essay on the history of the sport, would have poor content validity.

Constraints to content validity

The validity of ACTFL test task assumptions

The construct of developmental stages described in the Guidelines are reflected in the content of ACTFL test tasks. Some of the ACTFL developmental levels used in their oral proficiency tests and their corresponding test tasks are:

1. Listing (*Novice*)
2. Giving information (*Intermediate*)
3. Describing, narrating in the present or past time, summarizing, comparing and contrasting, and instructing (*Advanced*)

4. Supporting and defending an opinion, hypothesizing, persuading (*Supr.*) [13]

(Kenyon, 1998, p. 23)

LT researchers cite a lack of empirical evidence to support the assumption that the ACTFL-based oral proficiency test tasks accurately reflect the construct of L2 developmental stages (Bachman & Savignon, 1986; Lantolf & Frawley, 1985; Bachman, 1990; Matthews, 1990; Bachman & Palmer, 1996; Bachman & Cohen, 1998). Therefore, in view of the prescriptive nature of the ACTFL band descriptor determination and given the lack of empirical evidence supporting their ordering, the issue of their validity can at best be described as unresolved. When the construct of L2 developmental stages are manifested in the content of oral proficiency tests, issues of content validity ensue.[14]

The validity of the hierarchal sequencing of test tasks

Some researchers (e.g., Matthews, 1990; Lantolf & Frawley, 1988) have questioned the sequence of tasks in ACTFL-variant oral proficiency tests. In fact, it can be said to remain a matter of supposition that mastery of a narrative task should occur prior to mastery of an argumentation task, for example. The fact that some ACTFL-variant tests do not subscribe to the same order as others further compounds this issue. Indeed, Matthews (1990) warned that "the categories

---

[13] Supporting and/or defending an opinion is frequently referred to in discourse analysis as argumentation.

[14] In an intriguing, yet perhaps extreme view, Barnwell (1996) has observed that 'had the ACTFL/ETS procedure been a drug or a domestic appliance, it would have been withdrawn from the market, because its proponents supplied no proof that it did what it claimed to do. It was all development and no research' (p. 174, cited in Fulcher (1999).

employed by current tests invariably overlap to a greater or lesser extent, and at worst stand in an inclusive or covariant relationship to each other." (p.118).

On the other hand, it is possible that many candidates feel that oral proficiency test tasks *are* trustworthy. Indeed, Kenyon (1998) found evidence to support the ACTFL test task hierarchy since candidates in his study ranked the difficulty of the Simulated Oral Proficiency Interview (SOPI) test tasks in a manner that was similar to that presupposed by its developers.

Task authenticity: Oral proficiency tests as 'conversation'

Research into oral proficiency tests in recent years has also sought to establish if the discourse of oral proficiency tests could qualitatively be equated with natural conversation. Fulcher (1996b) discussed this issue:

> It has often been claimed that certain oral tests are valid on the grounds of the test task selected. (...) Thus, Wilds (1979: 12) argued that the validity of the Foreign Service Institute (FSI) test was 'unquestionable' because the oral interview was based upon a demonstration of speaking ability in a 'natural context' related to living and working abroad. (...) It need not be repeated that the appeal to face validity is neither a necessary nor sufficient condition for the validity of a test (Stevenson, 1985a; 1985b), but the issue of whether or not the task design used in a test is capable of producing a context for 'natural language output' is one which is worthy of investigation. (...) *Much of the work which has been done on 'interview talk' suggests that one-to-one oral interview generates a special genre of language different from normal conversational speech* (Lazaraton, 1992;

MacPhail, 1985; Perrett, 1990; Silverman, 1976; van Lier, 1989) [italics added] (p. 26).

Young (1992) has also noted that (...) " the Guidelines deal only cursorily with interactional discourse." (p. 3). Moreover, textual and discourse analysis have also demonstrated the unlikelihood that much of the language produced in oral proficiency tests does in fact approach that of authentic conversation. (Jennings, Fox, Graves & Shohamy, 1999; Lantolf &Ahmed, 1989).

Test tasks and method effects

Bachman (1990), and Bachman and Palmer (1996) have analyzed the elements of language tests resulting in a new awareness of various characteristics, or facets, that factor into the equation of how final scores are determined. Some of these test facets include the testing environment, the test rubric, the input and the expected response, the latter incorporating the format and nature of language. In addition, Rasch Measurement has shown that many variables contribute to the final rating in oral proficiency tests.[15]

As test facets such as 'test method' and 'test task' and others have become better defined and better understood, a strong interest in discourse analysis in oral proficiency testing has recently arisen in the language testing community. Accordingly, various researchers have investigated the effect of task on discourse produced (Bialystok, 1991; Young, 1995; Wigglesworth, 1997a).

---

[15] Rasch Measurement refers to a statistical tool used for finely estimating probabilities of various factors of rating scales, for example item difficulty (McNamara, 1966). Also known as the Rasch Model, it is widely used in the analysis of educational test data, and was introduced by Georg Rasch in the 1950s (Wright & Masters, 1982).

Textual and discourse analysis have also demonstrated the unlikelihood that much of the language produced in oral proficiency tests does in fact approach that of authentic conversation (Jennings, Fox, Graves & Shohamy 1999; Lantolf & Ahmed, 1989). Young and Milanovic (1992) used discourse analysis of an oral proficiency test and found that "The major influence on discourse as a whole was task." (abstract of document resumé). As noted previously, Wigglesworth (1997a) also studied the effect of task variation in an oral interaction (oral proficiency) tests. According to her "It was found that where there is an information gap, the nature of candidate discourse differs in both quantity and quality from the discourse elicited where no information gap exists." (p. 35).

More positively, Kormos'(1999) discourse analysis research suggests that the task of enacting a role-play, (a task often employed in oral proficiency tests), is a more effective approximation of natural conversational interaction than are other tasks used in them, and thus it is a useful measure of second language ability where naturalistic interaction is intended.

In the realm of second language acquisition as well, concerns have arisen regarding task and discourse produced. Long and Crookes (1992) have stated that "little empirical support is yet available for the various proposed parameters of task classification and difficulty." (cited in Robinson, 1995, p. 128) Robinson (1995) examined discourse elicited in the oral task of 'giving a narration'; his findings support the view that complex tasks elicit less fluent, yet more accurate and complex speech than do simpler tasks.

Turner and Upshur (1995) analyzed the discourse of ESL students in two different oral tasks and found differences in communicative effectiveness and grammatical accuracy which could be attributed to the task. The authors comment, "It is quite reasonable that different communicative tasks would make use of different component abilities." (p. 23) They also call for more research to investigate this question.

Douglas and Selinker (1985) used discourse analysis in studying oral proficiency interview tests and found evidence to support the concept that examinees perform better when topic domains are closely familiar to them.

Robinson defined the need for additional research of task type, in his comment that "determining valid criteria for the relative difficulty, and hence grading, of tasks for second language learners will require research aimed at establishing empirical differences between pairs of activity of the same type, set at different levels of complexity" (p. 128).

Bachman and Palmer (1996) have called for greater understanding and control of test method effects, as they have noted:

> There is also considerable research in language testing that demonstrates the effects of test method on test performance. This research and language teachers' intuitions both lead to the same conclusion: the characteristics of the tasks used are always likely to lead to affect test scores to some degree, so that there is virtually no test that yields only information about the ability we want to measure. The implication of this conclusion for the design, development, and use of language tests is equally clear: since we

> cannot totally eliminate the effects of task characteristics, we must learn to understand them and to control them so as to insure that the tests we use will have the qualities we desire and are appropriate for the uses for which they are intended. (p. 46)

Overall there has as yet been little research, however, on certain aspects of the discourse of oral proficiency tests. Young (1995b) observed this when he stated that "Although studies of language proficiency interviews abound, remarkably few researchers have examined in any detail exactly what participants say in these interviews. (...) This descriptive work is an indispensable foundation for studies of construct validity and the design of more effective instruments for assessing oral proficiency." (p. 7) Fulcher (1995) has also called for more research into the effect of task difficulty on test scores in oral language testing. (cited in Upshur & Turner, 1999, p. 87) And as early as 1988 Bachman, Kunnan, Vanniarajan and Lynch (1988) noted the importance of calibrating precise task difficulty to expected response for the purposes of rating.

Upshur and Turner (1999) analyzed the discourse produced in oral tasks, in creating an empirically-based rating scale. They provide salient reasons to suggest that oral performance rating scales should be task-oriented, for greater reliability of ratings across test tasks. They note:

> The weight of evidence suggests, therefore, that rating scales should be task-specific, not just population-specific. (...) On the basis of our evidence we do not believe that a more general scale-type should be

assumed. A further implication of our findings is that effective rating scales may reflect task demands as well as discourse types. (p. 105)

Bachman (2000) observed that we now have many more resources than in the past to deal with the challenges and difficulties of language testing. Bachman cited Albert Einstein's statement that "not everything that counts can be counted, and not everything that can be counted counts" (p.1). This in a sense exemplifies the current concern in the language testing community regarding the need for calibration of measures such as those used in oral proficiency testing, in order to achieve fair and just equivalent test forms.

Calls for empirical research

Hymes (1967) called for and proposed a model of a descriptive theory of language, warning about the consequences of discounting scientific, empirical approaches to language use. Consequently, Hymes observed the danger in the following:

> Diversity of speech, within the community and within the individual, presents itself as a problem in many sectors of life – in education, in national development, in transcultural communication. When those concerned with such problems seek scientific cooperation, expecting to find a body of systematic knowledge and theory, they must often be disappointed. *Practical concern outpaces scientific competence* [italics added].
>
> (Hymes, 1967, p. 8)

The concerns of the present discussion reflect those of Hymes. That is, that practical need has, in some measure driven test procedure in the ACTFL performance testing tradition, rather than has linguistic theorem.

Much later Fulcher (1995) wrote of the hazards of the naïve assumption that language performance is equated with competence in his warning that "maintaining the distinction between competence and performance does make a great deal of sense in any scientific enquiry... These consequences [of not maintaining the distinction] are the opposite of scientific enquiry" (p.30).

It follows that painstaking care should be shown in devising and maintaining second language tests. In view of these concerns, several researchers have called for empirical study as a basis for both test development and of test validation (Bachman, Davidson & Milanovic, 1996; Bachman, Lynch & Mason, 1995; Bachman & Savignon, 1986; Jacoby & McNamara, 1999; Lantolf & Frawley, 1985, 1988, 1992; Matthews, 1990; McNamara, 1995a, 1995b, 1996, 1997; McNamara & Adams, 1991; McNamara & Lumley, 1995, 1997; van Lier, 1989; Wigglesworth, 1997a, 1997b). Some researchers, for example Wigglesworth (1997a) have called for discourse analysis of test content data, even to the point of making this a routine validation endeavour.

Heretofore the testing approach of ACTFL and ACTFL-variant performance tests has not incorporated scientific, empirical investigations of test data. Rather, the raison d'être of these tests has traditionally been based on intuitions of how the tests and their rating scales should function. However, much contradictory evidence refutes many of the claims of ACTFL.

Numerous SLA and in particular, interlanguage studies, have weakened some assumptions pertaining to the validity of the ACTFL rating descriptor bands having shown that L2 acquisition, in fact occurs under circumstances of considerable variation in a multidimensional and non-unitary format (Ellis, 1997; Gatbonton, 1978; Shohamy, 1988; Young, 1995; Fulcher, 1996a).

Similarly, LT researchers have cited a lack of empirical evidence to support the assumption that the ACTFL and ACTFL-variant proficiency tests accurately reflect the construct of the developmental stages of L2 proficiency (Bachman, 1990; Bachman & Cohen, 1998; Bachman & Palmer, 1996; Bachman & Savignon, 1986; Lantolf & Frawley, 1985; Matthews, 1990).

In addition, some LT researchers have found in test content analysis, that while the interactions of performance tests have proven useful, they do not approach that of authentic conversation, as suggested by ACTFL (Fox, Graves & Shohamy, 1999; Jennings, Lantolf &Ahmed, 1989; Lazaraton, 1992; MacPhail, 1985; Perrett, 1990; Silverman, 1976; van Lier, 1989).

Finally, task method effects have been shown to contribute to test performance and to discourse produced (Bachman, 1990; Bachman & Palmer, 1996; Bialystok, 1991; Douglas & Selinker, 1985; Kormos, 1999; Robinson, 1995; Upshur & Turner, 1999; Young, 1995; Wigglesworth, 1997a). Calls have been made for more research into how method effects influence discourse produced in performance testing (Fulcher, 1995; Young, 1995). Fulcher in particular has observed the need for research addressing the question of the effect of task difficulty in oral performance testing (cited in Upshur & Turner, 1999, p.

87). The present research seeks to add to empirical knowledge within the language testing community regarding task difficulty and discourse produced.

To paraphrase Barnwell (1996) after a fashion, many assumptions dating from the 1950s (for example, that smoking is a relaxing and harmless pastime), would not pass the more rigorous norms of inspection of 2001. Language theorem and thought has evolved from the mid-twentieth to the early twenty-first century. Similarly, second language test development worthy of integrity, is no longer based on expert intuitions, but rather on sound, empirical study. This is particularly salient in the case of second language performance tests of high stakes, such as the Canadian Government SLE:OI test.

Lazaraton (1992) observes that "while objections have been (and continue to be) raised about numerous aspects of the OPI, there seems to be widespread agreement that the oral interview is the most appropriate vehicle for measuring oral proficiency" (p.373). This is undoubtedly the case of the SLE:OI proficiency test due to the high procedural standards it employs, particularly in training, in testing protocol, and in maintenance of standardized rater judgements (see Introduction).

Nevertheless, SLE:OI raters grapple daily with issues related to appropriacy of test content; it is a real and practical concern of SLE:OI and other ACTFL-variant oral proficiency test examiners. The present study attempts to some degree, to address these concerns.

The language testing community has called for more closely determined performance test content reliability and validity, based on empirical evidence.

Therefore, the present study contributes to this need, through the means of discourse analysis of the language generated in the SLE:OI performance test, and by investigation of method effects inherent in using equitable or inequitable question prompts in that test. There remains much work to be done in order to address this issue in the literature of second language testing.

In conclusion, this research will add to the as yet incomplete body of language testing research wherein ideas about what should validly be included in language proficiency interview test content are based not on intuition, but rather on empirical evidence.

Chapter 3

Purpose and Design of Study

Hypotheses and research question

In view of recent calls in the language testing community for empirical evidence pertaining to the actual speech generated in ACTFL-variant oral proficiency tests, I became interested in further inquiry into the kind of discourse generated in some of the functions, or tasks in these tests.[16] Essentially the question arose as to whether different question prompts used for the same task in oral proficiency tests elicited similar responses, since this is the basic premise of the test developers; as noted earlier, variation in question prompts is both standard practice in oral proficiency tests, and in the case of ACTFL-based tests, it is encouraged. In the North American ACTFL tradition, little concern is directed to the standardization of content prompts, as it is in the tradition of the British CASE oral proficiency interview test, for example (see Lazaraton, 1996). Therefore without empirical evidence, the assumption of reliability across different test forms in oral proficiency testing remains uncertain.

In designing a research study of this nature, it was necessary to choose an appropriate test task, or function for investigation. Hatch (1992) defines argumentation as "the process of supporting or weakening another statement whose validity is questionable or contentious" (p. 185). In this case argumentation

---

[16] I will use the terms 'task' and 'function' indiscriminately, as is done in ACTFL-type oral proficiency testing.

is elicited in the SLE:OI (an ACTFL-variant test) task requiring candidates to '*support an opinion.*' This function was selected for investigation in the present research for the following reasons. The content validity of the task has occasionally been found to be problematic since it may not accurately reflect the target language use (Bachman & Palmer's termed '*TLU*') needs of many occupations dependent on oral proficiency tests (1996). Bachman and Palmer have suggested the following approach to appropriately accommodate the *TLU* in language testing:

> In language testing, our primary purpose is to make inferences about test takers' language ability, and in most cases we are not interested in generalizing to just any, or all language use domains. Rather, we want to make inferences that generalize to those specific domains in which the test takers are likely to need to use language. In other words, we want to be able to make inferences about test takers' ability to use language in a target language use domain. (1996, p.44)

It will be recalled that the ACTFL OPI, the most widely-used North American oral proficiency test, was originally derived from the needs of FSI diplomats. It is logical to suppose that a needs analysis of diplomatic work might include work involving the supporting of opinions. On the other hand, it is debatable whether clerical workers and many others who are required to support and defend an opinion for ACTFL-variant oral proficiency tests actually ever need to do so in the course of their occupational duties. In fact, in my own experience in oral proficiency testing, test candidates have on occasion

themselves expressed to me their perception that the opinion function was not representative of their actual on-the-job tasks.

Again, from my viewpoint as a testing practicioner, it appears from my practical experience that the function of *supporting an opinion*, among all ACTFL oral proficiency test functions has the greatest actual or potential variation in structure. This variation may thereby pose the greatest threat to both construct validity and to reliability. Excessive prompt variation threatens construct validity since the construct being measured is less controlled under these circumstances. Thus, with less certainty can it be said which construct is being measured. Related to this, Robinson (1995) found that prompts used to elicit an opinion sample tended to elicit that of narrating instead. Even the dubiously -viewed face validity, (the appearance of validity), may suffer should excessive variation in prompts be noticed and disapproved by test candidates (see Introduction).

Cohen (1994) noted that "reliability asks whether an assessment instrument administered to the same respondents a second time would yield the same results" (p.36). Bachman and Palmer (1996) suggest a procedure for developing an oral interview test with controls for three sources of error relating to reliability which include "inconsistency of questions, lack of equivalence of different sets of questions, and lack of consistency among interviewer/raters" (p.184-185). It is my view that the SLE:OI test could benefit from Bachman and Palmer's approach by addressing the issue of prompt-related sources of potential reliability inconsistency in the task of *supporting an opinion.*

In the SLE:OI, the 'opinion' task was of interest in the present study for the added reason that in the test, it serves as a kind of breakpoint or threshold function. That is, in order for a candidate to successfully perform at the advanced C-level, the function *must* be accomplished. It follows that all of the B/C borderline performance candidates would have had to have successfully completed the task in order to have been awarded the C-level.

There are three ways for raters to arrive at opinion topics in ACTFL-variant oral proficiency tests, such as the SLE:OI. They are: 1) by pure invention, 2) by picking up cues from candidates, 3) from a bank of predetermined topics. The first is problematic since there is no way to accurately assess equivalence across tests. The second is in some measure unreliable since in this case the task may essentially be 'self-selected' by the candidates themselves.[17] The third appears the most effective, given that it affords some degree of reliability of content in alternate test forms, however, in cases where the topics have not been proven to be equivalent there is no guarantee that it is any better or more reliable than the first and second.

Therefore, for the previously stated reasons, the task of *supporting an opinion* may be crucial for borderline candidates in the SLE:OI, an ACTFL-variant test, and the issue of test fairness and equivalence of test forms becomes decisive. Let us consider the hypothetical case of a possible outcome of question prompt quality, on various candidate performances in the task of stating an

---

[17] This is due to the ACTFL test tradition of 'tailoring' the test to the individual candidate (see Introduction).

opinion. For this purpose we may consider 3 ranks of performance: highly proficient, clearly-defined C-level performances; strong B/C borderline (known as B/C level); and low B/C borderline (known as B/C level).[18] It may be hypothesized that the highly-proficient C-level candidates should encounter no difficulties in answering either undemanding or demanding question prompts. Similarly, B/C candidates should conceivably be able to adequately accomplish the task administered with either an undemanding or demanding prompt.

Conversely, we might be concerned that the performance of a low B/C borderline candidate may be compromised by the use of either an undemanding or demanding question prompt. It is possible that those who are asked an undemanding question may accomplish the task, while those asked a demanding question might not be able to do so. In this way the measurement instrument would not be reliable, since the construct under consideration (*supporting an opinion*), would essentially be split into 2 discrete constructs – supporting an undemanding opinion, and supporting a demanding one.

Consequently, weaker B/C borderline candidates would be unfairly penalized should they be asked a demanding, or difficult question. Conversely, stronger B/C borderline candidates may be given undue advantage should they be asked an easy one. Figures 1 and 2 illustrate in graphic form the possible outcomes of variable task difficulty, as discussed in this hypothetical situation.

---

[18] In this nomenclature, the underlined level indicates the final rating score. Thus a B/C candidate would ultimately be awarded a B-level rating, and a B/C would receive a C-level rating for the task of global test performance.

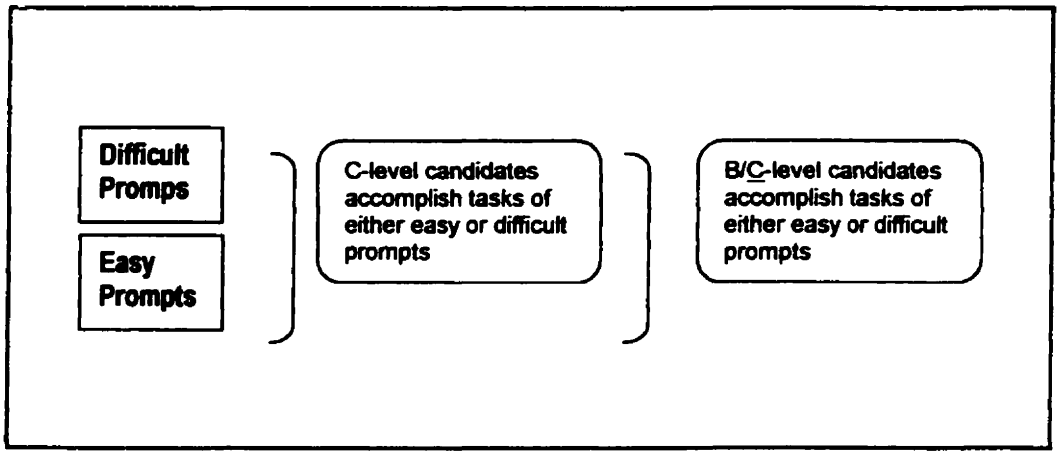

Figure 1. Possible outcomes of task difficulty on performance: in C and B/C borderline candidates.

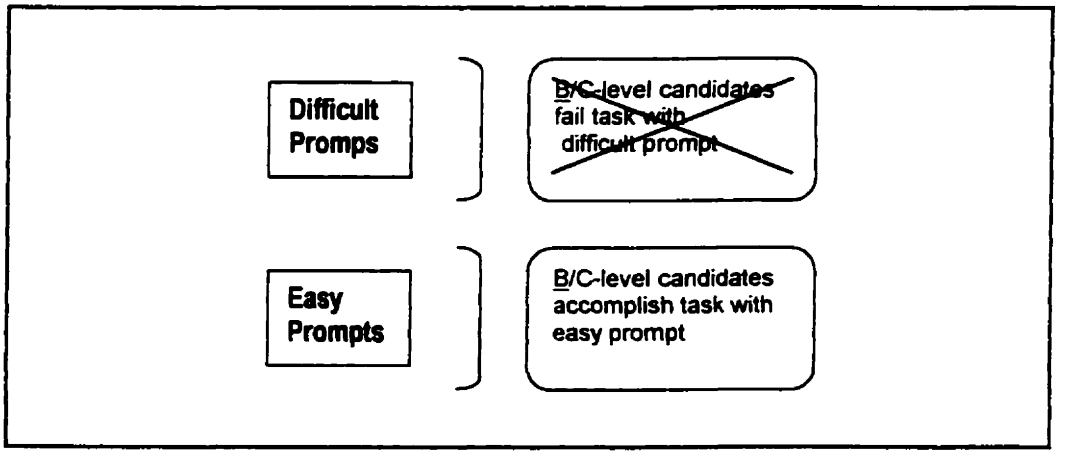

Figure 2 Possible outcomes of task difficulty on performance: in B/C borderline candidates.

Hence, it is the particular concern of the effect of using undemanding or demanding question prompts to B/C candidates in test administrations, which is the fundamental theme of the present research.

Thus, for all of the above reasons the task of '*supporting an opinion*' was selected for investigation.

This research query addresses the following hypotheses:

H_O There is no difference between speech samples elicited by different question prompts in the task of '*supporting an opinion*' in an oral proficiency interview test.

H_A There is a difference between speech samples elicited by different question prompts in the task of '*supporting an opinion*' in an oral proficiency interview test.

Thus the present research seeks to investigate the following question: Is there a difference in speech samples elicited by different question prompts in the task of '*supporting an opinion*' in an oral proficiency interview test?

The independent variable under investigation in this research is the use of various question prompts. The dependent variable is the kind of responses they elicit.

For the purposes of the present research, the term 'question prompt' has been used to describe spoken phrases and sentences employed by raters to elicit the task of *supporting an opinion* in an oral proficiency test. In fact, these phrases and sentences may also include statements intended to generate an opinion

response. Thus in the interests of brevity and comprehensiveness, I include statements in the definition of 'question prompts.'

Context of the study

The Canadian Public Service Commission's (PSC) occupational testing centre uses an ACTFL-variant oral proficiency test, the Second Language Evaluation: Oral Interaction (SLE:OI), to assess the oral abilities of its employees and potential employees. Moreover, the PSC is the body responsible for administering language tests in the staffing of bilingual positions throughout most of Canada's Federal Departments, its affiliated agencies, and the Canadian Armed Forces. Students of the Government language training school, operated by the PSC are also users of the SLE:OI test. The SLE:OI oral proficiency interview test exists in English and French formats. Second language tests for Canada as a whole are administered at the Montreal and Ottawa offices of the PSC, in face-to-face and telephone versions of the test. Candidates for employment in the Government must demonstrate L2 ability prior to employment whenever a position has been classed as bilingual.

Approximately thirty percent of all Canadian Federal Government positions have been identified as requiring bilingual ability in French and English, the two official languages of Canada. A second language reading, writing or oral proficiency requirement has been identified for each bilingual job description, and each of the three skills has been assigned a level of required L2 proficiency. The L2 proficiency requirements of bilingual Government positions correspond to the

rating scale levels accorded in the Government L2 language test battery, which includes the SLE:OI.

During the 4-month period of the present study, the Montreal office of the Public Service Commission administered 2,097 SLE:OI oral proficiency interview tests in English.[19] A total of 6,541 of these English tests were given in that office over the fiscal year of 2000-2001.[20]

Participants

In conformity with the ethical standards of McGill University, all participants in the present study signed an English or French authorization form 'Informed consent to participate in research,' or 'Consentement à participer à la recherche,' (see Appendices A and B, respectively). Additionally, the certificate of ethical acceptability appears in Appendix C.

All test candidates recorded on the audiocassettes were personally telephoned to obtain consent, the majority expressing positive interest in the present study.

Test candidates

Twenty candidates of the English SLE:OI who were tested over a 4-month period in 2000, participated in the study. Of the total candidates, 16 were already

---

[19] Statistics are not available for the precise number of English tests in the exact period of the study, since the data was collected from July 20 through to November 16, 2000; the reported number covers the *total* number of tests administered throughout the months of July to November, inclusively.

[20] The fiscal year in this case refers to the year dating from April 1, 2000 to March 31, 2001.

employed in the Government in various departments, and in the military. Four candidates were non-civil servants who were in the process of being evaluated in language and other staffing tests in the selection processes of competitions for various Government positions. None of the participants was enrolled in the Government language training school at the time of the study.

All of the candidates were French Canadians who spoke French as a first language. An even split of gender characterized the group; 10 were female and 10 were male. Eleven candidates, or just over half of the group had been tested in the telephone version of the test, while 9 others were tested in face-to-face test administrations. All had been selected for inclusion in the study solely because their performance in the SLE:OI fluctuated between intermediate score level B and advanced level C throughout the test.

The majority of the candidates held middle or higher management positions.[21] This may have occurred by happenstance since, as noted above, the sole criterion for candidate selection in the present study was by virtue of the fact that their tests had been flagged as comprising 'borderline' performances. Alternately, this may have occurred as a result of an element of sample selection bias (see below, Procedure: Phase 1, Initial Data Collection section of the present study), or in fact, for other reasons. In view of the fact that this was not the focus of the present research, this feature was noted but not pursued.

---

[21] The profile of candidate job standing was the following: 3 were in a higher clerical class; 8 were in middle management; 6 were in higher management; and the status of 3 was unavailable. These figures and job designations at the time of the data collection include both civil servants and non-civil servants.

Test Raters

Eight accredited SLE:OI rater-interviewers employed in the Montreal office of the PSC volunteered to participate in the study. Six identified and submitted the borderline test cases that were used in the study; six participated in an exercise in which discussions and data qualification into categories occurred (Workshop 1); and finally, six attended a training session and questionnaire completion exercise, to further qualify the data (Workshop 2). With one or two exceptions, the same raters participated in all of the above procedures. The raters acted as expert judges, since all were certified Canadian Government oral interaction assessors. (See Introduction.) The test raters' work experience varied from 1 to 17 years. In addition, of the eight rater participants, seven had previous experience as second language teachers.

English as a second language teacher

An English as a second language teacher participated by trialling the process of question prompt categorization in a small sample, in Trial 3 of the preparation to Workshop 1. The teacher had twenty-four years of teaching experience, had received training in and applied the communicative method of language teaching. At the time of the study the teacher was teaching English as a second language, language arts (ESL-LA) at a French secondary school in Montreal.

Instruments

The SLE:OI oral proficiency interview test

The SLE:OI test was the oral proficiency testing instrument under investigation in this study. The SLE:OI, an ACTFL-variant oral proficiency test, consists of a structured interview in a guided conversational format lasting approximately 30 minutes, serving as a speaking interaction in which candidate fluency, vocabulary, grammar and pronunciation are assessed in a global rating. One rater and one candidate are present during the interview test: the rater interviews and evaluates candidate performance concurrently. The SLE:OI is considered a *direct* test since it is audiotaped and rated concurrently. (*Semi-direct* tests are audiotaped for subsequent rating, usually by a separate rater, as in the case of the Simulated OPI test, the SOPI.)

The SLE:OI is rated by means of a rating scale consisting of 5 major bands of speech performance descriptors, which are graduated from the weakest to most competent performance in the following order: X, A, B, C, Ex (exemption from further testing[22]). The bands are relatively comparable to the Novice-low to Superior band descriptors used by ACTFL, as explained by Cole and Neufeld (1991).

Broadly-defined, the SLE:OI can be considered to be an occupational test in that its content topics are work-related in a general sense, and specific to individual candidates when possible, as is commonly the case with ACTFL-variant oral proficiency tests. Some of the functions used in the structure of the test are performing communicative activities such as asking questions, relating

---

[22] It is noteworthy that in the case of the SLE:OI, the assigned order of the proficiency levels is curiously inverse to that which normally would be expected.

events, giving explanations, and expressing and supporting opinions.

The SLE:OI is administered in both a face-to-face and a telephone format. For a description of the SLE:OI test band descriptors and task examples, see Appendix J.

The Question Prompt Categorization Grid

The present research sought to ascertain whether different question prompts used for the task of supporting an opinion, in an oral proficiency test elicited similar responses. Consequently, an exercise qualifying question prompts was accomplished in Workshop 1 of the present research, and this resulted in the creation of the 'Question Prompt Categorization Grid'.

I conducted Workshop 1, and the participants were with six SLE:OI raters acting as judges. A bank of 152 question prompts from the SLE:OI oral proficiency test were classed into categories by the judge-participants. Each of the six judges were given approximately five question prompts transcribed from actual tests (see Procedure). Working with the question prompt data, the judges identified five categories in which they could be placed. This process was repeated a total of three times to allow categorizations adjustments and modifications to be made. (For example, midway through the proceedings one participant added a sixth class to her particular categorization list.) All of this information was recorded on a flip-chart and discussed by the group in each of the three phases. Thus at the termination of the workshop a grid generated from the data had been created, identifying 35 discrete categories of question prompt. The

completed Question Prompt Categorization Grid appears in the presentation and discussion of the qualitative results (see Chapter 4, Table 5).

Additionally, the judges of Workshop 1 indicated in discussions that the most compelling factor differentiating the categories was task difficulty, or complexity. For the purposes of the present research, the terms 'task difficulty' and 'task complexity' have been used interchangeably.

The Criteria for Determining Task Difficulty document

Based on the classification exercise of Workshop 1, Workshop 2 sought to qualify the categories of question prompt identified in the Question Prompt Categorization Grid in terms of their relative difficulty. In preparation for Workshop 2, I reviewed the methodology of several researchers in determining task difficulty in L2 speaking tasks (Brindley, 1987; Brown & Yule, 1983; Anderson & Lynch, 1985; cited in Nunan, 1989, p. 141-3). These approaches were then modified to reflect the particular context of the SLE:OI oral proficiency test.

Following that, a taxonomy was produced in a document entitled 'Criteria for Determining Task Difficulty,' which is reproduced in Appendix D. The document served as the basis for an introductory discussion in Workshop 2 of the issue of how to establish task difficulty in the present context. The judges discussed issues of task difficulty, and I acted as moderator. The judges then completed the 'Question Prompt Category Complexity Questionnaire,' described below.

The Question Prompt Category Complexity Questionnaire

I created the Question Prompt Category Complexity Questionnaire based on the results of the Question Prompt Categorization Grid, which had been produced in Workshop 1. Thus, the Question Prompt Category Complexity Questionnaire was intended as a means for Workshop 2 judges to assign relative difficulty levels to the 35 previously-identified categories of question prompt. The Question Prompt Category Complexity Questionnaire appears in Appendix E. In creating the questionnaire, certain modifications were made to the categories originally identified in Workshop 1.

For example, of the total original categories, six paraphrased the same idea. Therefore, these were collapsed into single categories, leaving a final total of 30 separate categories of question prompt. The 30 categories were then broadly grouped under seven general headings for the purposes of clarification. Only one category, characterized by a judge as *Questions leading to opinion*, was omitted on the basis that it was too vague for the purposes of qualification.

Finally, the questionnaire included instructions to the respondents to assess the difficulty of each of the 30 remaining question prompt categories on a scale of 1 to 4, where 1 = *easy*, 2 = *somewhat easy* , 3 = *fairly difficult*, and 4 = *difficult*.

Procedure

This investigation consisted of two phases of qualitative and quantitative data collection, respectively. Thus, Phase 1 entailed a period of qualitative data

collection and categorization, accomplished in Workshops 1 and 2. Thus, Phase 1 consisted of the qualification of question prompts and categories of prompts.

Phase 2 quantitatively examined data taken from the former Phase 1, and new data was also generated in Phase 2. For example, in Phase 2, a consensus of the questionnaire response data was identified through quantitative means. Similarly, quantitative methods were employed along with discourse analyses in examining candidate question prompt response data.

Procedure: Phase 1

Initial Data Collection

Primarily, I had requested that SLE:OI raters identify tests administered over a 4-month period.[23] Additionally, I specified that the tests be those in which candidates had demonstrated markedly variable test performances, fluctuating between the score borders of B and C, known amongst raters as borderline tests, (or in this case B/C borderlines). As a result of inherent candidate variation between levels, these tests are known to be challenging to rate. Therefore, the selected SLE:OI tests had been more problematic to rate than is normally the case. Borderline cases were chosen since, with regards to test method effects, this population of test candidates were in the most tenuous position of any (see Figure 2).

[23] The selected tests were administered between July 20 and November 16, 2000. The choice of a 4-month period of data collection was partially made for operational reasons. Audiocassette recordings of SLE:OI tests are erased 4 months subsequent to test administrations. Consequently, none would have been available prior to the 4-month timeframe.

The final data selected were audiocassette recordings of samples of rater and candidate discourse in 27 SLE:OI oral proficiency interview tests administered over a period of 4 months in 2000. The test samples were included in the study solely in view of the fact that they qualified as B/C borderline tests administered over time. Consequently, it could be argued that their selection was to some degree random. According to Hatch and Lazaraton (1991) "You can achieve a random sample if everyone and everything has an equal and independent chance of being selected" (p. 43). In a sense this definition applies to the test samples in that they included *all* of the identified B/C borderline tests of that time period. However, the fact that they were in turn selected by the raters themselves suggests at least the possibility that some selection bias may have occurred. The tapes were volunteered, and therefore they cannot be considered to be a true random sample. Nevertheless, the research design can be said to approach that of 'two-stage sampling' as defined by Petersen, Kolen and Hoover (1989, cited in Linn), in that the sample tests were selected from a population (all B/C border tests over 4 months), and subsequently from these the selection was further refined on a qualitative basis (for more details on the latter, see 'Procedure: Phase II, Selection of question prompt samples from consensus data,' below.).[24]

Finally, no practical mechanism existed for a completely objective

---

[24] The authors' example of this case is the following: "In a norming study using two-stage sampling, schools might be selected in the first stage using cluster sampling, and students might be sampled from within schools as a second stage" (Petersen, Kolen & Hoover, p. 240, cited in Linn, 1989).

selection of B/C tests given that no formal registry of these cases existed. It was known that the raters did keep informal records of such cases. Consequently, it was decided that the best way to proceed was to request the raters themselves to select the sample tests.

Initially, it was anticipated that sampling error variance might be controlled by way of selecting sample tests on the basis of several additional constraints. These included having equal numbers of final score B and C tests, and face-to-face and telephone tests. In addition it was hoped that the candidate population sampled would: (a) work in similar types of jobs, (b) have jobs of similar rank, (c) have similar levels of education, and (d) have been tested by the same rater, (e) be actual *or* potential civil servants, and (f) have had previously taken the test the same number of times, *or* had never taken the test. Unfortunately, this plan had to be abandoned ultimately as it was not possible to meet these conditions in the present research context. In fact, practical realities imposed that the conditions of sample selection could be limited to the isolated fact that the tests would be B/C borderlines and would have been administered over a 4-month period. On the other hand there was a measure of control for sampling error variance, given the uniformity found in the following conditions:

1. All the candidates had previously been identified as having demonstrated fluctuating performances between score levels B and C throughout the test.
2. Close to half of the candidates were tested in telephone and face-to-face test administrations (the ratio was 11:9).

3. None of the participants were enrolled in the Government language training school at the time of the study.
4. French was the first language of all the candidates.
5. Half of the candidates were male, and half female.

In summary, the first condition occurred by design, while the others arose from the data as a fortunate result of chance.

Delimitation of question prompts

The first step in managing the data was to identify and delimit actual questions used to elicit the function of argumentation, or supporting an opinion. It is important to note at this juncture that the actual test function in its entirety includes both *supporting* and *defending* an opinion. It was decided to focus only on the former part of the task in order to more accurately distinguish question prompts that were comparable across test administrations. That is because I felt that there could be a great deal of variation in sub-questions whose aim it was to encourage candidates to elaborate on the initial topic, rather than to introduce a new one. It seemed logical that sub-questions might take any direction, therefore it would be best to avoid that which could not be equitably compared.

An additional threat to the integrity of the data was the fact that although one task of stating an opinion might occur in an oral proficiency test, in fact several questions may be employed in order to accomplish the task. Ideally, one question ought to be sufficient in order to do this, but in actual practice, several may be necessary. Moreover, questions are often rephrased or abandoned in favour of others. My concern was that critical data might be lost if an unduly

limited sample of only one or two question prompts were to be documented. For this and the reasons stated above, it was determined that all *independent question prompts* used in the sample oral proficiency tests would be included in the sample registry.

For the purposes of the present study, the term *independent question prompt* means a question prompt which by its structure stands alone in its role of attempting to elicit an opinion, as opposed to that which serves to prolong a previously elicited response or responses. The latter, prolonging prompts, will be called *non-independent question prompts*.[25]

When a one or two sentence preamble was essential to the understanding of an independent question prompt, this was also included in the data.
I felt it was preferable *not* to include those question prompts which were peripheral to the testing of the task of supporting an opinion. These were informal and usually very short, interjected comments clearly not formulated to elicit the fuller sample needed in the '*support an opinion*' task. For example, some question prompts occurred during the discounted warm-up or wind-down phases of the test. Thus while the questions served to ask the candidate's opinion on some matter, by their structure and placement in the test, they were plainly not intended to elicit a sample of the function of stating an opinion, so they were not noted or transcribed. Moreover, it was clear both from their initial and final placement, from their non-work-related topics, from their formulations, and from the

---

[25] A fictitious illustrative example of the latter would be a question such as "Oh yes, so you were saying just now, but why not?" Obviously, this query would not stand alone to qualify as an independent question prompt.

presence of opinion testing elsewhere in the test, that these were not intended for evaluation purposes.

Nevertheless, on the exceptional occasions when an opinion segment was clearly being tested within the structure of another function, such as a role-play, it was included in the data set.[26] When any of the above conditions were not clearly met, the question prompts were not included in the data.

It was noticed that on occasion a prompt which clearly did not appear to be intended to elicit an opinion, did in fact do so when opinions were volunteered by candidates. These were excluded since this structure did not conform to the elicitation question prompt structures under investigation. (These could be of interest however, in a research study of another nature.)

In general and as much as was possible, the responses were ignored at this stage of the research. (Responses were superficially reviewed when it was necessary to determine if the '*support an opinion*' task was being tested.) This was done in order to minimize possible bias in the upcoming qualification part of the research study. Thus I felt that I, as researcher should not have any preconceived attitudes regarding the type of response the selected questions extracted. Exceptions to this on the other hand, were those questions which seemed controversial, confidential, or to which in some instances the candidates themselves had objected. In these cases it seemed best to review the answers for the purposes of probable elimination. (In a similar attempt to avoid bias, I did not

---

[26] This practice is done occasionally and when the rater feels it is appropriate, in order to test the function while maintaining a conversational style in the interaction.

read reports of candidate test scores, which were available from the beginning of the research, until all of the data analysis of the study had been completed.)

In summary, it was felt that a principled protocol for question prompt identification was crucial for the success of the study. Table 1 illustrates these and other criteria used for inclusion and exclusion of question prompts in the data.

Protocol of question prompt transcription

Notwithstanding the role of repetitions and hesitations in characterizing communication, I decided to omit these features of the question prompts in the transcription process. This was done in view of the fact that they might detract from the overall message. Their exclusion would afford a measure of uniformity to the question prompts, given that the intention was to compare them in the upcoming qualification exercise of Workshop I. Thus, it was felt that the construct under investigation, (differences in question prompts), needed to be as clearly characterized as possible, and moreover since the original oral language would become written and transcribed data, omitting repetitions and hesitations would more effectively preserve the integral message.

Similarly, individual words or short phrases were *added* to question prompts when needed to provide an understandable context ensuring the comprehensibility of the message.[27] This was rarely necessary. The additions were made when substituting a preamble instead would have created an undesirably long sample.

---

[27] It is important to differentiate this case from that of 'Questions requiring deictic markers which were not provided in an immediate preamble,' as noted in Table 1. The latter were excluded since they required much context definition which was not specified in the preceding speech.

Table 1

Criteria for Identification and Selection of Question Prompts

Question prompts included in the data
• Independent question prompts: those preambles and/or statement or question prompts clearly formulated to elicit the function of stating an opinion.
Question prompts excluded from the data
• Non-independent question prompts: question prompts and short question fragments which by their structure do not stand alone in a role of attempting to elicit an opinion.
• Discourse not formulated to elicit the function of stating an opinion.
• Questions of opinion in the warm-up and wind-down segments of the test.
• Questions requiring deictic markers which were not provided in an immediate preamble.
• Questions whose format did not appear intended to elicit an opinion, but which did elicit one, (volunteered opinions).
• Questions and responses to questions which identified participants.
• Questions which elicited responses of a confidential nature.
• Questions which were highly controversial, which dealt with sensitive or politically sensitive topics, or which candidates indicated were inappropriate.

As noted previously, it was deemed necessary to arrive at as much uniformity as possible in transcribing the question prompts for subsequent comparison purposes. Initially this seemed an unlikely and unrealistic expectation. The data, however, proved to be easily standardized since the length of preambles and question prompts were surprisingly similar. The average length of question prompts was approximately twenty-five words.

Each candidate was assigned an alphabetical letter and each rater a Roman numeral designation for identification purposes and in order to preserve data confidentiality. Various information regarding the test was recorded with the initial transcription. This included the date of the test, candidate position, position applied for, test structure, the channel and locale of the test (in the case of telephone administrations.)

When this process was complete, 152 question prompts from the original 27 tests submitted by raters had been identified, and transcribed. The average number of independent question prompts per test was 5.63.

In anticipation of Workshop 1, the transcribed question prompts were formatted on 18 pages with between 5 and 10 questions on each. The pages were prepared using Word software, in table format in order to make them easily readable.

Workshop 1 Preparatory categorization trials: Piloting the methodology

In order to prepare for Workshop 1 and in the process, to familiarize myself with the process of categorization of the question prompt data scheduled

for Workshop 1, I decided to proceed with three trial categorizations. I carried out two of the trials, and an English as a second language (ESL) teacher accomplished a third.

The primary difference between Trials 1 and 2 and the categorization exercise proposed for the upcoming Workshop 1, was that in Trials 1 and 2 an attempt would be made to group *all* of the question prompts; in Workshop 1 the judges would each be given smaller samples of prompts to catalogue. Thus, I categorized the 152 question prompts on two occasions, in Trials 1 and 2. Following the experience of the trials, I felt that an earlier, preliminary decision to include more judges in the categorization process was valid. The subjective nature of the judgements made this option particularly compelling, since I felt that having more judges would likely lead to the production of a more accurate qualification of the data.

Trial 3 was attempted in order to further test the procedure. The choice of an experienced teacher of English as a second language (ESL) to catalogue the Trial 3 data was an afterthought; I realized that this might bring a fresh perspective to the exercise. (Additionally, the brief ESL component in the study would serve peripherally to reflect the fundamental link that ESL teaching has with LT, as a kind of nod to language teaching, which in many ways can be said to drive language testing.) Nevertheless, the principal objective of asking the ESL teacher to participate in Trial 3 was to establish how feasible it would be to require a professional in the field of second language education to catalogue a sample of 10 question prompts in a desired timeframe of 5 minutes.

Thus, the ESL teacher was asked to categorize a small sample of 10 question prompts into five categories. No constraints were placed on the teacher for the requested grouping choices; instead, it was suggested that he allow the data to generate the most appropriate categories. The teacher was given 5 minutes in which to categorize the 10 question prompts. Since the *primary* reason for holding Trial 3 was to ascertain its feasibility, the results of the Trial 3 exercise were reported and compared to those of Trials 1 and 2, but they were not extensively analyzed (see Chapter 4, Presentation and Discussion of Results: Qualitative Analyses).

Workshop 1: Protocol of question prompt categorization

The objective of Workshop 1 was to qualify the question prompt data into categories determined by six SLE:OI rater-judges, and with myself conducting the exercise. The following procedure was followed.

I began with a brief prologue to the workshop informing the judges that the areas of interest in the study were the prompts used in the test function of *supporting an opinion* and the discourse they elicit. The categorization exercise was then introduced. The participant judges were advised that they should freely allow the *data* to generate the categorization process, rather than it being a process of prescribing preconceived categories. For this reason I declined to give much detail of the kind of categories that might be determined.

After that, each judge was given a sheet on which 10 opinion task question prompts had been transcribed, for the purposes of categorization. The participants were given approximately 5 minutes in which to class the question prompt data,

working individually. Following the initial classification exercise, each judge's categorization results were recorded on a flip-chart. Their appropriateness were then considered and discussed amongst the researcher and judges as a group.

In a second and subsequent third attempt at classification of the data, the categories were further discussed and modified, as necessary. Sheets of 5 to 10 transcribed question prompts each were given out to the participants in the second and third trials. In this way the process was repeated in order to refine the categories, and to complete the process for the entire data bank of question prompts.

Workshop 1 lasted approximately 2.5 hours, and the proceedings were audiotaped. In a follow-up procedure after Workshop 1, the judges who participated in that exercise were asked to identify the particular question prompts which they had placed in their category designations, which they did. (This was done in order to be able to later match qualified prompts to the responses they had elicited in test administrations, after Workshop 2 had been completed.)

Workshop 2 preparations: Development of the Criteria for Determining Task Difficulty document

Preparations for Workshop 2 were based directly on the results of Workshop 1. Thus, I began the preparations by examining the Workshop 1 proceedings by reviewing an audiotape of the session. Certain sections were then transcribed when it was felt this would elucidate their meaning and validate my original impressions of what the judges had said in the proceedings. This effectively confirmed that the judges of Workshop 1 had identified question

prompt category difficulty/complexity as the primary characteristic differentiating the categories. Thus, I determined that Workshop 2 would have as its objective the qualification of question prompt category difficulty/complexity in the SLE:OI test task of *supporting an opinion.*

The original research question of the present study asked if there were differences in generated speech samples and in test scores when different question prompts were used in language tests. Consideration of this led me to question what characteristics language test tasks, and in particular language proficiency test tasks, should entail *ideally.* Furthermore, what *ideal* features of test tasks were reflected in desirable question prompts? What task features would be considered to be undesirable?

In the event that these questions might arise during the forthcoming Workshop 2 discussions, I felt it would be prudent to investigate the issue beforehand. Indeed, it was felt essential that the participant judges understand as clearly as possible the concept of task difficulty in the context of the SLE:OI test. In order to apply these concepts to the SLE:OI test context, I felt it would be helpful and instructive for the judges to be aware of accepted SLA approaches to assigning task difficulty.

In order to address the issue of ascertaining task difficulty, I examined Nunan's (1989) synthesis of several factors of the ideal general-skill language learning task.[28] I then modified Nunan's list to better reflect the context of

---

[28] Nunan (1989) based this work on second language tasks used in the L2 classroom, and his list is of a general nature, not specifically addressing L2 sub-skills such as speaking, listening, reading or writing.

SLE:OI *oral proficiency* test tasks, by eliminating and adding content. I subsequently created a customized list, taken from Nunan's L2 task characteristics, of the ideal oral proficiency test task.

Thus, based on Nunan (1989) it may be posited that the ideal proficiency test task should:

1. communicate clearly what is expected of the candidate
2. closely approximate the communicative skills candidates would be expected to use in the workplace
3. involve a sharing of information
4. activate background knowledge of the topic featured
5. enable candidates to manipulate specific features of language

Interestingly, the characteristics I had noted and cited above closely parallel those of the previous Workshop 1 judges. The judges' comments arose spontaneously when I had asked them what they felt differentiated the question prompt categories (see Chapter 4, Presentation and Discussion of Results, Phase 1: Qualitative analyses, Workshop 1: Question Prompt Categorization Grid data). This correspondence of Workshop 1 judges' views with those of SLA researchers in the five ideal task attributes list suggests the timeliness of this approach.

The above led to further refinement of the investigation into L2 task characteristics. Therefore, I investigated approaches which had previously been used to ascertain learner task difficulty in L2 *speaking* tasks. Brindley (1987); Brown and Yule (1983); and Anderson and Lynch (1985) delineated factors contributing to task difficulty in the communicative L2 classroom context (cited

in Nunan, 1989, p. 141-3). Despite the fact that these methodologies had been devised for qualification of tasks in the second language classroom, and not for language testing per se, they nonetheless closely resembled the Workshop 1 judges' approach to determining question prompt difficulty. Therefore, as in the previous exercise, they were modified to reflect the present performance testing context.

The above surveys represent what may constitute ideal L2 tasks in *general*, in *test tasks*, and in *speaking tasks*. All served to advance the development of a more precise understanding of the construct of task difficulty in anticipation of Workshop 2.[29]

Therefore, based on this review of the literature to determine how various researchers had ascertained L2 learner task difficulty, and on the previously mentioned review of the proceedings of Workshop 1, a taxonomy of characteristics for consideration when assigning SLE:OI test task difficulty/complexity was produced, in the form of a document entitled 'Criteria for Determining Task Difficulty.' It is found in Appendix D. The document was created to enhance comprehension and promote discussion of the issue of task, question prompt, and question prompt category complexity or difficulty in the forthcoming Workshop 2.

Workshop 2 preparations: Creation of the Question Prompt Category Complexity Questionnaire

---

[29] As anticipated, later in Workshop 2, some judges *did* question of what the ideal SLE:OI test task should be comprised.

During the previous Workshop 1, the Question Prompt Categorization Grid had been produced, which recorded the judges' categorization of 152 question prompts into 35 classes (see Chapter 4, Table 5).[30] This document served as a basis for the production of a questionnaire for use in Workshop 2, the 'Question Prompt Category Complexity Questionnaire,' (see section on Instruments, and Appendix E).

The categories of the Question Prompt Categorization Grid required some modifications before they could be listed in the Question Prompt Category Complexity Questionnaire. Originally, in Workshop 1, 35 question prompt categories had been identified in the Grid. However, I elected to omit the categories identified as *leading to opinion* and *miscellaneous.* This was done in view of my belief that they were too vague to allow for qualification in terms of their level of complexity. Consequently, I did not feel that these categories would be expected to be qualified with any precision.

In addition, of the total categories, six were paraphrased versions of the same concept. Consequently, in each of these cases, I opted to collapse the two paraphrased categories into a single category. The six paraphrased categories were the following:

1. *[To what] extent... questions*
2. *To what extent... quantitative questions*
3. *How 'adjective' is... evaluative adjective questions*
4. *How... is this... questions*

[30] Twenty-one of the prompts could not be included in this list (see Procedure: Phase II, Identification of 2-group data question prompts section).

5. *Descriptions (with a free rein in the response)*
6. *Questions leading more to a description*

In modifying the above, the first category was put into the second, and the name of the second one was conserved. The third and fourth categories were grouped together in a category which was renamed *How 'adjective' is... evaluative adjective, range questions (using degree-intensifying adjectives)*. Finally, the fifth and sixth categories were regrouped as *Description (free rein in response); questions leading to more of a description.*

In addition, 3 categories were collapsed and restated. Originally they were:

1. *Agree or disagree*
2. *Yes and no questions*
3. *Yes/no, little opinion required*

They were rephrased as *Saying yes, little opinion required, agreeing*, and *Saying no, little opinion required, disagreeing*. Additionally, the category identified as *Short and long questions*, was divided into 2 categories, (renamed *Short* and *Long*.)

After these measures were taken, the data consisted of 30 question prompt categories, which underwent slight modifications in order to be used in the questionnaire. After all, it was considered vital that the questionnaire be as clear and easy to read as possible in order to control for instrument bias. Therefore, the content of the categories was unchanged but the sentences describing them were restructured for uniformity. For example, all of the instances of formulaic

sentences were rewritten so that each of them would begin with its key interrogative word.

Additionally, my review of the categories resulted in the observance that patterns emerged from the data in the form of distinct types of categories. Given that, and because the qualification task was considered to be quite demanding, I decided to incorporate these category types as 'macro' headings in the questionnaire, for clarity. Upon further analysis of the data, it was decided to separate the fifth macro heading into 3 'micro' headings. The macro and micro headings are listed below:

1. A) Topic specification question prompts
2. B) Question prompts with an expected elicited response which is functional
3. C) Question prompts grouped by length or amount of detail in the expected response
4. D) Question prompts which use formulaic questions
5. E) Question prompts with an expected elicited response of a particular type:
    - Relating
    - Speculating
    - Other
6. F) Grouped by vocabulary used in question prompt
7. G) Grouped by syntax used in question prompt

This final question prompt category list was used to comprise the Category complexity questionnaire. (The end result, listing the macro and micro headings as well as the 30 question prompt categories is reproduced in Appendix H.)[31]

Once all of the above had been accomplished, the preparations for Workshop 2 were competed.

Workshop 2: Protocol of questionnaire administration

Workshop 2 sought to qualify the question prompt categories in terms of their relative difficulty. Workshop 2 was held 1 month after Workshop 1. As in Workshop 1, six SLE:OI rater-judges participated in the exercise.

In Workshop 2, I initially reviewed with the judges the results of the previous workshop, in some detail. I also reiterated judge comments from Workshop 1 identifying question prompt difficulty as a foremost feature differentiating the question prompt categories. Following that, the Criteria for Determining Task Difficulty document was distributed among the judges. They were then asked to read the document, which was subsequently used as a point of reference for a discussion concerning task difficulty, which ensued.

This was followed by the judges' reflecting on and discussing various approaches to determining question prompt category difficulty. As anticipated, one judge asked what factors would constitute the ideal question prompt. In response, I attempted to clarify some of the issues related to the ideal task or

---

[31] These macro headings appear in Appendix H, 'Workshop 1 and 2 results: Question prompts and headings,' and in the questionnaire in their original format listed from A to G.

question prompt formulation as discussed above. An attempt was made to respond in a precise yet *concise* approach in hopes of avoiding the introduction of bias. That is, the desired objective of the exercise was to stimulate discussion on the subject of task difficulty, whilst allowing judges the freedom to use their unbiased best judgement in the subsequent exercise of completion of the Question Prompt Category Complexity Questionnaire.

Once the issue of determination of task difficulty had been addressed in discussion, I asked the judges to provide their judgments of the relative difficulty of each of the question prompt categories arising from Workshop 1, using a ranking scale from 1 (*easy*) to 4 (*difficult*) to complete the Question Prompt Category Complexity Questionnaire (see Appendix E). The judges worked individually to complete the questionnaire.

Workshop 2 was concluded in approximately 1 hour. The proceedings were recorded on audiotape.

Analysis of the Question Prompt Category Complexity Questionnaire responses

The Question Prompt Category Complexity Questionnaire was created for the purposes of qualifying the category data in terms of their relative complexity or difficulty. The questionnaire required participant judges to determine task complexity for each category of question prompt. This had been done by employing a scale of 1 to 4.[32]

[32] In the scale 1 represented a question prompt considered to be *easy*, 2 signified *somewhat easy*, 3 signified *fairly difficult*, and 4 denoted *difficult*.

Once the questionnaire had been administered in Workshop 2, the audiotape of the proceedings was reviewed, and as had been done in the previous workshop, some of the judges' comments were transcribed. The effect of this review was that it was resolved that 4 of the questionnaire item responses be eliminated from the data, so as to avoid biasing the results. The justification for this action is as follows.

Firstly, it was found that category 'E) 1) *Comparing, asking for qualities*' was problematic due to the evidence of the audiotape of the Workshop demonstrating that one of the participants had clearly misinterpreted the intended meaning of the instructions. This was evident in her comments which had gone unnoticed during the proceedings' general discussion. Certainly this judge, and possibly others had not understood the meaning of the item, therefore it was decided to omit this item from the data.

Secondly, problems had arisen when there had been confusion in the discussion pertaining to whether categories C) *Short* and *C) Long* referred to the question prompts themselves or to the expected responses to them. Again, in reviewing the session audiotape, I determined that this issue had not been adequately clarified in the course of the workshop. Therefore, I decided that the response to the categories itemized as C) *Short* and *C) Long,* be discounted from the questionnaire data as well.

Finally, four reasons were found justifying the omission of the category called *E) 4, Question Prompts with an Expected Elicited Response of a Particular Type, Other, Elaborating or wrapping-up, elicited from a statement,* ' which were

(a) 2 out of the 6 judges rejected this item as 'Not Applicable,' (N.A.), (b) it appeared from the audiotaped evidence of Workshop 2 that the judges (including the one who had originally proposed the category), found the category definition to be obtuse, (c) the category made reference to a way of wrapping up the test task rather than of eliciting it, as did the other categories, thus it did not 'fit' the data set. Therefore, due to all of the above it was decided to omit this item.

In conclusion, after omissions, the original 30 categories of the Question Prompt Category Complexity questionnaire data were reduced to 26. The next step in the examination of the questionnaire results showed that a consensus on the level of question prompt category difficulty had been reached by the judges (see Chapter 5, Phase 2, Identification of consensus of Question Prompt Category Complexity Questionnaire responses).

The fact that a very clear consensus was found regarding question prompt categories the participants had identified as *easy* and *difficult*, meant that it was possible to advance to the next stage in the procedure. This next step was to select candidate responses from each of the *easy* and *difficult* groups in order to identify candidate responses to the qualified questions. Following that, the candidate responses would be transcribed, in preparation for further analysis.

Procedure: Phase 2

Identification of question prompts issuing from 2-group consensus

On the basis of the consensus of *easy* and *difficulty* question prompt categories noted above, it was possible at this point to isolate questions from the data bank which were included in the two consensus groups. To summarize, a

large bank of specific question prompts had been categorized; each category had been qualified as to its level of difficulty; therefore, it was possible to identify and select those questions included in the sets of *easy* and *difficult* categories at this point.

Several factors resulted in diminishing the set of deemed *easy* and *difficult* questions (hereafter called the *easy* group and the *difficult* group). Some of these were due to participant withdrawals from the study for personal reasons. The *easy* and *difficult* question groups were further condensed by virtue of the fact that four of the category qualification questionnaire items had been found to be problematic and had been deleted from the data bank (see Chapter 4 for a detailed discussion of this). In addition, analysis of the data from Workshop 1 indicated that some of the judges had left some questions uncategorized.

The end result of all of the above was that the final set of question prompt data was decreased. The *easy* group question prompt number went from 19 to 11, while the difficult group diminished from 13 to 10 questions.

Following this, 11 question prompts from the *easy* group, and 10 from the *difficult* group were identified. Subsequently, the candidate responses to these question prompts were analyzed through the procedures of discourse analysis (see Chapter 5).

Transcription of candidate responses: the response idea unit (RIU)

One of the challenges of transcribing oral discourse is in delineating boundaries to speech acts. In the case of elicited responses to question prompts, it was necessary to review the literature and to determine a principled method with

which to do this. Crookes (1990) surveyed several language segmentation units used in second language discourse analysis. For example, the *utterance* has been chosen as a discrete unit of speech. It has been defined as:

> a stream of speech with at least one of the following characteristics:
>
> 1. Under one intonation contour
> 2. Bounded by pauses
> 3. Constituting a single semantic unit.
>
> (Crookes & Rulon, 1985, as cited in Crookes, 1990, p. 187)

Correspondingly, the notion of an *idea unit* has been defined by Kroll (1977) as:

> A chunk of information which is viewed by the speaker/writer cohesively as it is given a surface form... related... to psychological reality for the encoder. (Kroll, 1977, as cited in Crookes, 1990, p. 184)

It was decided to incorporate the approaches of the utterance and the idea unit, and furthermore, it was decided to border the units by pauses *and/or* intonation changes.[33] This was done in order to accommodate the variable French L1 intonation and pauses in the candidate responses. Therefore, the candidate response boundaries were delineated using what will be called, for the purposes of the present study, the *response idea unit* (*RIU*). It is defined as:

---

[33] This feature reflected a concern about the difficulty of establishing units in the present context where some candidates demonstrated heavy L1 interference, and also in view of the fact that Tarone (1985) had been "unable to analyse some of her recorded speech samples because it was so dysfluent, there were so few complete sentences (sic) and so much hesitation and repetition" (cited in Foster, Tonkyn, & Wigglesworth, 2000, p.360).

> A segment of information which is a single semantic unit, bounded by pauses and/or intonation changes, and in which the speaker speaks cohesively with the purpose of relating the message to psychological reality for the encoder.
>
> (Adapted from Crookes & Rulon, 1985; and Kroll, 1977, as cited in Crookes, 1990, p. 187, 184.)

Therefore, it follows that response idea units are delimited by topic shift boundaries. The *RIU* proved to be effective. Its application in delineating the response data was unexpectedly undemanding, suggesting that it was a valuable tool for the present purposes.

The coding protocol used in transcribing the *RIU*s is illustrated in Appendix F.

Analysis of candidate responses: Discourse analysis protocol

Fluency protocol selection rationale

The *type-token ratio* (*TTR*) is an equation revealing the number of separate words per total number of words in a text. Several researchers have used the *TTR* to measure the fluency of discourse in second language speaking tests (Douglas, 1994; Tomiyama, 2000; Wigglesworth, 1997b). In addition, Crookes (1989) used a TTR in an SLA study of L2 interlanguage.

On the other hand, Lennon (1990) has defined the temporal aspect of fluency in second language speakers as "speech at the tempo of native speakers, unimpeded by silent pauses and hesitations, filled pauses... self corrections, repetitions false starts and the like" (cited in Cucchiarini, Strik, & Boves (2000,

webp. 2). Accordingly, some researchers have evaluated L2 discourse fluency by investigating the frequency of total *unfilled pauses*, *self-repetitions*, *self-repairs*, and other features (Foster & Skehan, 1996; Tomiyama, 2000; Wigglesworth, 1997b). Given that this methodology looks at *fluency feature frequency*, I will call it the *FFF* method.

In the present study it was decided to incorporate and adapt both the *TTR* and the *FFF* approaches in order to arrive at a comprehensive, and more accurate estimation of response discourse fluency.

Fluency: Type-token (TTR) measure

The *TTR* fluency analysis in the present study was accomplished in the manner of that of Douglas (1994). Douglas described the *TTR* protocol in the following way, "The ratio is an indicator of the number of words produced, *discounting false starts and repetitions* [italics added] (type), as a function of the total number of words produced for the item (token)" (p.131).

In addition, in the present study it was decided to omit repetitions in the *RIU* of lexical items the interlocutor (the rater-interviewer), had said. Thus the total number of *types* and *tokens* per RIU, as well as the *type-token ratios* were tabulated accordingly. This was repeated for each candidate response in the *easy* and *difficult* question prompt groups. Next, the fluency as measured in the *TTR* was compared in both groups using a *Shapiro-Wilks test for normality*, followed by a *t-test*, and a *Wilcoxon Two-Sample test*. This, and the most of the other statistical analyses were done using Statistical Analysis System (SAS) software;

exceptionally, two *Chi square analyses* of discourse complexity data were done using AB-STAT software.

Fluency: Fluency feature frequency (FFF) measure

In order to measure *FFF*, occurrences of *repetitions*; *self-repairs*; *silent pauses*; and *filled pauses* were taken for each *RIU* in both groups. For the purposes of the present study, a silent pause is defined as a silent speech hesitation of 1 second or more. *Filled pauses* include gaps filled by sounds such as 'um' and 'uh.' It is noteworthy that both *silent* and *filled pauses* have the effect on fluency of briefly suspending speech.

Since the *RIU*s were of course not of uniform length, the frequency counts of the 4 speech qualities were converted to percentages for comparison. A *Chi square contingency table analysis* was performed, however, using the frequency counts of the data.

Alternately, the total number of silent pauses in seconds was calculated, and subsequently both a *Wilcoxon Two-Sample Test*, and a *t-test* were done.

Accuracy measurement

Prior to analyzing the accuracy and complexity measures, it was determined to further subdivide the *RIU* divisions in conformity with standard practices in discourse analysis of this sort. This would also afford a standard of greater uniformity across measures, and it would facilitate the process of identifying discrete grammatical features within *RIU*s. Since the data involved speech samples, the written textual sentence unit was considered ineffective to the analysis task. A more appropriate oral speech division was sought.

Thereafter, the coding conventions of Foster et al (2000) were reviewed, and in some measure incorporated in the present study, as well as were those of Berman and Slobin (1994).[34] The *RIU* sections were divided into an adaptation of Foster et al's *Analysis of Speech Unit* (*AS-Unit*), *Level 3*. Thus, the conventions of what I shall call the *Simplified Analysis of Speech Unit* (*AAS-Unit*) appear in Appendix G.

Discourse accuracy was measured by frequency counts of target or nontarget forms of various grammatical components, within the *AAS-Units* of the *RIUs*. The selection of forms for examination were adapted from Wigglesworth (1997b). Thus verb morphology accuracy was addressed by assessing the following bound morphemes: *subject-verb agreement*; the presence of *an obligatory subject and/or verb*; and appropriate *tense marking*. The accuracy of a lexical form was assessed by means of examining the *common, compound, and abstract noun usage*.

Frequency counts of the presence of target and nontarget grammatical forms were used to perform a *Chi square contingency table analysis* to compare the measures in the *easy* and *difficult* groups.

Complexity measurement

Discourse complexity has frequently been measured through analysis of *clause subordination* (for example, Crookes, 1989; Foster & Skehan, 1996; Foster et al, 2000; Wigglesworth, 1997b). Following the coding of Foster et al, the *RIU*

---

[34] I am indebted to Dr. Gillian Wigglesworth for acquainting me with the former work of Foster, Tonkyn and herself, and also to Dr. Ruth Berman for suggesting I consult the latter.

data from the two groups was divided into *independent, subordinate, and subclausal units*, as the latter has been defined by Quirk, Greenbaum, Leech and Svartvik (1985, p.838-853, cited in Foster et al).

The frequency count procedure was followed by the administration of a *Chi square contingency table analysis*, comparing the two groups.

This chapter has delineated in detail the qualitative and quantitative procedures of Phases 1 and 2 of the present study. In Chapter 4 the qualitative results of Phase 1, and in Chapter 5 the quantitative results of Phase 2, are presented and discussed.

Chapter 4

Presentation and Discussion of Results: Phase 1: Qualitative analyses

Workshop 1 preparatory Trials 1, 2 and 3: Piloting the methodology

Prior to Workshop 1, it was necessary to ascertain how feasible it would be for judges to categorize the transcribed question prompts. In order to do this, three trial categorizations were done. I performed the first two trials myself, and a third was done by an ESL teacher.

In Trial 1, I categorized the entire data bank of 152 question prompts. No basis for their classification was used other than by allowing myself to be guided by any first impressions the data might bring out. These impressions were based on the familiarization I had garnered following the process of transcribing the bank of question prompts. I attempted to identify five categories of data prompt. However, as this number proved to be too limiting given the vastness number of prompts, I ultimately identified seven categories.

It was intended that 5 minutes be spent on the exercise in order to limit reflection and to encourage an impressionistic approach to the task. However given the large number of question prompts involved, ultimately 10 minutes were spent in the categorization exercise.

The results of Trial 1 led to four classifications by sentence structure or form, and three by general topic. Those question prompts categorized by the former included prompts in the form of *statements intended to elicit a reaction*; formulaic structures such as *statements followed by 'do you agree?,'* and *'to what*

extent... ' formulations; as well as prompts which were *repeated, including paraphrased forms of other questions.* In addition, the latter three categories were classed by *general topic*, which were technology in general, specific technologies, and issues surrounding gender in the workplace. The results of Trial 1 are found in Table 2.

Table 2

Trial 1: Categorization of 152 question prompts after 10 minutes

	Category 1	2	3	4	5	6	7
Researcher Judge	Statements for reaction	Statements + 'do you agree?'	'To what extent' questions	Repeated, including paraphrased questions	General technology questions	Specific technology questions	Gender in the workplace questions

In Trial 2, I again categorized the complete 152 question prompt data bank. However on this occasion the process was done more systematically, taking approximately 3 hours to carefully categorize the voluminous data bank. Unlike the impressionistic approach of Trial 1, the approach used in Trial 2 was methodical and exacting. This methodology seemed to have interesting implications for the trial results since the sole basis found to accommodate all of the question prompts into categories was by classing them all by *general topic.* Ultimately, five categories of general topics were identified in the trial. The results of Trial 2 are found in Table 3.

Table 3

Trial 2: Categorization of 152 question prompts after 3 hours

	Category 1	2	3	4	5
[redacted]	Personal characteristics; and profession-related, employee relations, and gender questions	Opinions on issues of how to work, working environment, and teleworking	Technological advances and the uses of technology, including communications	How the public or media portray or perceive others	Miscellaneous topics

In Trial 2, I had effectively spent more time and classified far more question prompts than would any of the study participants. It appears possible that the sheer numbers of these data constrained the outcome of the exercise. By way of illustration, it might be considered hypothetically that a space satellite could overview a broad geographical area of the earth and categorize parts of that area into towns. However, a bird watcher viewing a section of one of those towns with the use of binoculars might in addition, notice several species or categories of birds, possibly perceiving the gender of some, whether some are young offspring or adults, and so on. Thus, in this case closer inspection would have increased and refined the categorization process.

Since the objective of the categorization process of the present research was to arrive at an accurate classification of question prompt data and not solely on a broad overview of them, and in view of the results of Trials 1 and 2, it was resolved that the best approach to classification would involve several sequential categorization exercises using limited numbers of question prompts in each attempt.

I was also quite concerned about the possibility that the Workshop 1 judges unwittingly exercise bias in their categorization decisions, given the familiarity they have with the ACTFL testing tradition of placing a great deal of emphasis on question *topics*. Certainly, the categorization task was intended to allow participant-judges to categorize the data on any basis they chose. The process of classifying small groups of question prompts appeared to allow for more freedom in the categorization process, as was evident in the results of Trials 1 and 2. For this reason as well as those stated above, I therefore endorsed the procedure of categorization of small groups of question prompts to foster a more careful inspection of the data.

Trial 3 was an attempt to further test the procedure from the perspective of another judge (see Chapter 3, Purpose and design of the study, Participants), and in particular to establish if the desired exercise timeframe of 5 minutes would be feasible. Accordingly, an ESL teacher was asked to spend five minutes to categorize 10 data bank question prompts into 5 categories. The teacher had no difficulty in accomplishing the assignment in the time allotted. The results of Trial 3 are in Table 4, below.

Table 4

Trial 3: Categorization of 10 question prompts after 5 minutes

	Category 1	2	3	4	5
ESL Teacher Judge	Questions regarding support for employees	Questions referring to personality traits	Questions about employees' abilities	Questions making reference to the appropriateness of a situation	Questions dealing with issues to be promoted or encouraged

The results of Trial 3 were that 3 of the question prompts were classed in categories of *employee-related issue topics*, and 2 of them were classed according to other *general topics*. The former category of employee-related issue topics, is interesting in view of the fact that it reflects a basic design feature of ACTFL-variant interview tests. That is, that their content is intended to closely reflect candidate interests (see Introduction). Thus the categories of employee-related issue topics included questions regarding support for employees; questions referring to personality traits; and questions about employees' abilities. On the other hand, the general topic categories were comprised of question prompts making reference to the appropriateness of a situation; and those dealing with issues to be promoted or encouraged.

Trial 3 more closely approximated the projected categorization exercise of the upcoming Workshop 1, than had the previous trials. Trial 3 had a small sample size of 10 question prompts, and its duration was 5 minutes. Conversely, the categorization exercise of Trial 3 did not include successive categorization attempts whereby the categories might be adjusted or modified, which was the proposed procedure of Workshop 1. This aspect of the exercise was not deemed as necessitating a trial, so it was not carried out.

Nevertheless, Trials 1 and 3 served to illustrate that it would be feasible to ask participant judges to categorize ten question prompts, and that they could be expected to do so in a time duration of between 5 and 10 minutes.

To summarize, Trial 1 looked at a large data sample, the entire data bank. Yet it is possible the categorization results may have been constrained by the

vastness of the data bank reviewed, by the impressionistic approach followed, and by the short, 10 minute time allotted for the exercise. Therefore, Trial 2 sought to more methodically look at the complete data bank, over a period of 3 hours.

Possibly the Trial 2 results were also constrained by the very fact that it failed to look closely at small samples of the data. There is evidence of this in view of the fact that the categorization results were of one sort exclusively, that of general topic. Hence, some limitations of the overview perspective may have come in to play, given that the more precise 'binocular' view was not present in Trial 2.

Finally, Trial 3 sought to incorporate elements of the first two trials, but this time in a categorization exercise involving the closer inspection afforded a smaller sample of 10 question prompts, categorized in the shorter time of 5 minutes. Trial 3 seemed to more successfully approximate the anticipated approach of Workshop 1, yet it too was in all likelihood constrained by the fact that it failed to allow for modification of the categories in subsequent attempts.

Thus, it was decided that the best approach to follow in Workshop 1 would be to proceed by asking the judges to categorize a small sample of 10 question prompts in a limited time of between 5 and 10 minutes, and to modify these categories in subsequent attempts. It was determined that in Workshop 1 it was possible for the six judges to categorize the complete data bank of 152 question prompts in three attempts.

Workshop 1: Production of the Question Prompt Categorization Grid

The purpose of Workshop 1 was the categorisation of 152 question prompts by six SLE:OI rater-interviewers acting as judges.

Given the option of working individually, or of conferring as a group concerning their category choices, the judges chose to individually categorize the prompts, and then to consult as a group after each attempt, citing a general consensus that consultation would facilitate any needed modification of categories. Accordingly, after the initial and each subsequent attempt, I recorded the categories identified by each participant on a flip chart at the front of the room, and these were examined in a group discussion.

In the initial categorization attempt five judges were each given 10 question prompts to categorize, and a sixth judge, of her own volition categorized 20 question prompts. The initial categorization procedure took 8 minutes.

In the second categorization attempt five judges were each given 10 question prompts to categorize, and one judge received 5 question prompts for categorization. The second and third attempts were completed in approximately 10 minutes on each occasion.

In the third attempt three judges were given five question prompts to categorize, and two judges received six question prompts for the purpose. In the third attempt one judge did not receive question prompts to categorize, but participated by reviewing the prompt data given to another judge.

Workshop 1 took a total of two hours, most of which was spent in a group discussion led by myself. The central focus of the talks involved the relevance of

the question prompt categories identified. Although the categorization exercise was done in three stages, there was little modification of the original categories identified. (For example, Judge VI added a fourth category in the second attempt.) The exception to this was the case of Judge IV, who created a new list of categories in the second attempt. The results of the Workshop 1 categorizations are found in Table 5.

Judge I categorized the sample question prompts as *problem-specific*, those intended to elicit opinions about a specific case; *relational, how one fact relates to another fact*; questions about *choices or options*; those involving *description (with a format of "free rein" in the response)*; and lastly, those in which candidates would be required to *agree or disagree.*

Judge II identified the first two categories as '*surface*' and '*deep*' questions, in relation to the amount of complexity required in their responses. This judge qualified the *surface* and *deep* categories in the following comment, "They're [candidates are] having to compare and contrast as well as say why something is important. To me that seems as if it requires more depth in the response."

Judge II called a third category *confused, possibly multiple questions, and meandering questions*. This category was almost unanimously endorsed as one that raters wished to avoid. There was some discussion, however, of the contrary supposition; comments were made to the effect that questions that take longer to frame and are more slowly phrased sometimes appear to elicit a more voluminous

Table 5

Workshop 1 Question Prompt Categorization Grid: Final categorization of 152 question prompts

	Category 1	2	3	4	5	6
Rater Judge I	Problem-specific (case study)	Relational (how one fact relates to another fact)	Choice/options	Description (with a free rein in the response)	Agree or disagree	
II	'Surface' questions	'Deep' questions (Compare/ contrast/say why kinds of questions)	Confused/ possibly multiple questions/ meandering questions	Repetition of a key vocabulary element in the question	Short/long questions	
III	Solution-seeking questions (could lead to explanation)	Presenting different points of view	Justifying points of view by generalizing	How *adjective* is... evaluative adjective questions	Pick one out of a series (most important quality etc.)	
IV	[What] do you think... questions	[To what] extent... questions	Would you say... questions	Statements, with no direct opinion word used	Miscellaneous	
	What sort of person... questions, (asking for qualities, could be comparative)	How ... is this... questions, /Evaluative, range questions)	Yes/no questions	Devil's advocate questions, (seeking a response)	'Recipe for a solution' questions, for a solution, (e.g. "How do you strike the balance between *x* and *y*...")	
V	Speculative questions about outcomes	To what extent... quantitative questions	Yes/no, little opinion required	Listing questions (where the response may list)	Statement questions to elicit elaboration or wrap-up	Questions leading to more of a description
VI	Job specific	Leading to opinion	Leading to explanation	Suggesting a point a point of view or speculating on one, to get a reaction		

sample than they otherwise might. It was suggested that this could be advantageous to some candidates who may benefit from having additional processing time.

A fourth category identified by judge II were those questions involving *repetition of a key vocabulary element in the question.* An example question prompt was quoted in which the phrase 'in your opinion' was repeated twice. The overt signalling of the task was considered by certain judges to be unfairly advantageous to some candidates. Others suggested that with some test candidates this was a necessary instructional component, needed to increase the likelihood that candidates understood the task requirements. Finally, judge II categorized some question prompts as *short or long.* While there was some discussion regarding the advantages of either, no consensus was reached by the group as to which kind would be preferable.

Judge III categorized some question prompts as *solution-seeking, which could lead to an explanation.* These were generally considered to be problematic since they could elicit a sample of an explanation rather than the intended one of *supporting an opinion.* Secondly, judge III identified categories of question prompts which were effectively *presenting different points of view*, for the purposes of obtaining a reaction, or either agreement or disagreement on the part of the candidate. In a similar vein, this judge identified a category of question prompt formed by *justifying points of view by generalizing.* This kind of question is also intended to elicit a reaction of some sort.

A further category judge III documented was *evaluative adjective questions,* with the example given of *'How adjective is....x?'* This kind of formulaic question prompt uses degree-intensifying adjectives that require a qualified answer. (An example of this would be a question asking how important something is.) Similarly, this judge identified a category of question wherein candidates would be asked to *pick one out of a series, (such as the most important quality of someone or something).*

Initially, Judge IV classed the question prompt samples in three categories based on formulaic sentence structure. These were [*What*] *do you think...questions*; [*To what*] *extent questions*; and *Would you say...questions.* This judge also identified a category of *statements where no direct opinion word was used,* and a *miscellaneous* one. However as noted earlier, judge IV was the only one to modify the original list to create a new one in the second categorization attempt.

Thus, judge IV's second set of categories included *questions asking for qualities, such as 'What sort of person....'* Judge IV explained that this could be used as a comparative line of argumentation. As in the case of judge III, this judge also categorized *evaluative questions* which use degree-intensifying adjectives, which in this case were termed '*range questions*' since they questioned the breadth of certain issues.

The remainder of judge IV's categories focussed on the kind of expected response they would elicit. These were *yes or no* [response] *questions*; those framed as *devil's advocate- styled questions, seeking a response*; and what judge

IV termed *solution-seeking questions*, such as those in which a type of "recipe" for a solution was sought. The judge gave an example of a question in the latter category, which was '*How do you strike the balance between x... and y...?*'.

Judge V identified 6 question prompt categories altogether. These included those of *speculative questions about outcomes*, and formulaic questions which this judge called *quantitative questions, such as 'To what extent....'*

As in the case of judge IV, judge V also based some categorization on the intended response. One of these categories was *yes or no* [response] *questions.* Judge V noted that for these types of questions, little opinion was required in the response. In addition, judge V included what was called *listing questions*, since the elicited responses could consist of lists intended to support the argument. *Statements* intended to elicit an *elaborative wrap-up* were also categorized. These would occur solely in the final stages of the 'opinion' task in the test. Judge V also identified a category called *questions leading more to a description* in the response. In the second categorization attempt this judge commented that the latter category bore some similarity to that which judge III had called *evaluative adjective questions.*

Judge VI categorized *job-specific* types of questions; and in categorizing, differentiated between those that were *leading to an opinion* as opposed to those that were *leading to an explanation.* In addition, this judge identified a category of question prompts which was called *suggesting a point of view or speculating on one, in order to get a reaction.*

During Workshop 1 discussions the judges were asked what they felt differentiated the identified categories. Several factors were discussed, including the following:

1. response complexity required
2. amplitude of possible responses
3. length of question prompt
4. lexical complexity of question prompt
5. presence of overt lexical signalling of the task
6. confused question prompt formulation
7. ephemeral nature of current question topics (which may tap, or fail to tap candidate background knowledge as topics fall in and out of favour over time)
8. prerequisite cognitive (as opposed to linguistic) ability required

All of the above have an impact on the complexity or difficulty of question prompts and categories of question prompt. The judges also indicated in the workshop discussions that the concept of question prompt difficulty should include variables related to the candidate. For example, due to variation in candidate background knowledge, some questions could be difficult for some candidates, while the same questions might be easier for other candidates. This viewpoint to some extent echoes that of Bachman and Palmer (1996), who include individuals' personal characteristics, topical knowledge, affective

schemata, and language ability among factors affecting performance in language tests.[35]

Notwithstanding the reality of the facet of candidate factors, clearly the judge participants of Workshop 1 determined that task complexity was the *principal* factor differentiating the identified question prompt categories. This accords with the fact that the issue under investigation in the present research is that of question prompt complexity. Consequently, candidate facets and other factors influencing test performance, while important, are beyond the scope of the present research and will not be addressed.

Interestingly, one judge suggested that if the present research were interested in task complexity, then the categories identified in Workshop 1 might be more pertinent to the investigation than would that of topic. Similarly, I had postulated that the procedure of categorizing small groups of question prompts would foster a more careful inspection of the data. In fact, this was the rationale behind holding the workshop. Thus, the comment of the judge supported my perception of the results of Trials 1,2 and 3 as discussed earlier.

Following the collection of the categorization data and the discussions of Workshop 1, I asked the judges how many of the questions they had recognized. This was done because I was concerned that the judges' categorization decisions

---

[35] In Bachman and Palmer's (1996) conceptual framework of language use as it relates to specific language test uses, the authors define personal characteristics as age, sex, and native language; topical knowledge as the real-world knowledge that individuals bring to the testing situation; affective schemata as the affective or emotional correlates of topical knowledge, and language ability as the particular construct specific to the testing situation (64-66).

might be biased in the event that they might recognize question prompts they themselves had formulated in the tests.[36] The response to the question was reassuring. With one exception, none of the judges recognized any of the 152 question prompts presented in Workshop 1. This suggests that in this data collection exercise, bias based on recognition of the data was negligible. In addition, it validates the protocol used to transcribe the bank of question prompts. The protocol had been intended to standardize the data across question prompts; sentence structures approaching uniformity (and thus less recognizable), had been expected to be more easily compared and categorized (see Chapter 3, Procedure: Phase 1, Transcription protocol of selected question prompts).

Finally, it was noted that in the Question Prompt Categorization Grid there was considerable redundancy in the categories the judges had identified. It followed that the next step would be to organize these data in order to see if patterns might emerge from the identified categories.

As noted previously, the judges of Workshop 1 had identified prompt complexity/difficulty as the primary characteristic qualifying the categories. Review of the audiotape of the Workshop 1 discussions reconfirmed this. Hence, it followed that the next step would involve the qualification of the categories in terms of their relative complexity, in Workshop 2.

---

[36] The majority of the judges were SLE:OI raters those who had previously selected the tests used in the present study. Therefore, the judges had themselves formulated the question prompts used in the data bank. Similarly, Nunan (1989) conducted a task qualifying workshop with ESL teachers in which the teachers' task descriptions were rendered unrecognizable for the purposes of qualification of task difficulty.

Workshop 2: The Question Prompt Category Complexity Questionnaire

Phase 1 concluded with the administration of The Question Prompt Category Complexity Questionnaire in Workshop 2.

Phase 2 began with the examination of the questionnaire response data and the new data generated from it. These data are presented and discussed in Chapter 5.

Chapter 5

Presentation and Discussion of Results: Phase 2: Quantitative analyses

It will be recalled that the present research consisted of two phases of qualitative and quantitative data collection, respectively. Phase 1 involved the collection of qualitative data, including categorization and the identification of *easy* and *difficult* categories. This was accomplished in Workshops 1 and 2.

Phase 2 examined data taken from the former Phase 1, through quantitative means. New data was also generated in Phase 2, and quantitatively analyzed. Phase 2 begins with the presentation of the data that follows, 1) the identification of a consensus of questionnaire response data, followed by 2) the presentation of the results of discourse analyses which examine the nature of candidate responses to question prompts from the *easy* and *difficult* groups.

1) Question Prompt Category Complexity Questionnaire: Consensus identification

It will be recalled that in each category the judge-respondents had indicated their determination of its level of difficulty. This was done using a scale of 1 to 4. Following that, an analysis of the participant responses was done in order to determine if any unequivocally identifiable consensus as to difficulty level had been reached.

In order to distinguish the level of consensus amongst the judges, a binary protocol was used. Thus the 4-point scale of level difficulty was divided into 2 sections; those scaled as *1, easy*, and *2, somewhat easy*; and those scaled as *3,*

fairly difficult, and *4, difficult*. A consensus was considered to exist when all of the respondents elected to respond in only one of the 2 sections. Thus, a consensus was determined to occur in category 8 (*'surface' questions*), since all of the responses to this item were found in points *1, easy* and *2, somewhat easy*.

Similarly, a lack of consensus was found to exist when the respondents selected scale points occurring on both sides of the binary division. Thus, a lack of consensus was found in category 5, (*description, (free rein in response); questions leading more to more of a description*), since the responses occurred on both sides of the binary division, in points *1, easy*; *2, somewhat easy*; and *3, fairly difficult*.

A majority consensus was considered to exist when it was established that there was a consensus of responses falling on either side of the binary division, and a majority of those responses occurred on one of the two scale points in that section. For example, in category *25, Repetition of key vocabulary element in question*, a majority consensus was determined since all of the responses were on one side of the binary division, in points *1, easy* and *2, somewhat easy*; and because the majority of these fell in one scale point, *2, somewhat easy*. The questionnaire analysis results are illustrated visually in Figure 3. More detailed reporting of the response data is found in Appendix I. The results of the questionnaire data analysis indicated that there was no group consensus for 15 of the categories qualified. This fact is interesting in itself since it illustrates the intricacy of determining prompt, and in turn, task difficulty. This is especially evident given the fact that the respondents of the present study were highly

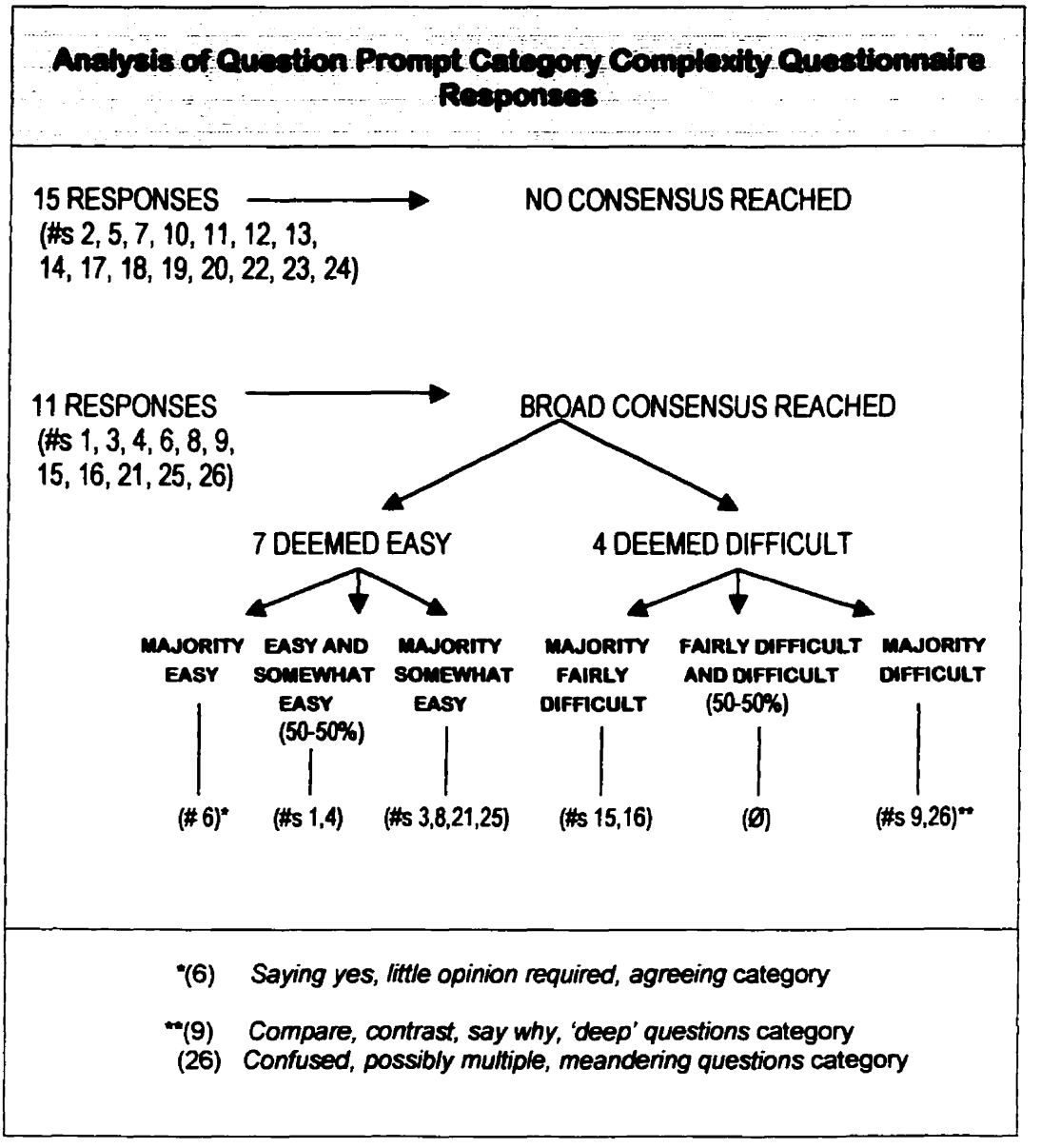

Figure 3 Analysis of Question Prompt Category Complexity Questionnaire responses qualifying category difficulty, where a consensus is determined to exist when all respondents select either scale points 1 and 2, or 3 and 4; and where a majority consensus was considered to exist when there was a consensus of responses falling on either side of the binary division, and a majority of those responses occurred on one of the two scale points in that section.

trained and experienced judges who moreover, dealt professionally with issues of question prompt difficulty on a daily basis. The fact that they could not arrive at a consensus in determining prompt or task difficulty underscores the complexity involved in creating parallel prompt forms, particularly in the context of the 'conversational' nature of ACTFL-variant tests.[37]

A consensus of a broad nature was found, however in 11 categories. Seven categories were qualified by consensus as *easy*, while four were qualified by consensus as *difficult*. Of the former, only one was qualified with a majority consensus as *easy*; two were equally qualified as *easy* and *somewhat easy*; and four were qualified by majority consensus as being *somewhat easy*. In terms of the latter qualification, two categories were qualified with a majority consensus as *fairly difficult*; none were qualified as both *fairly difficult* and *difficult*; and two were deemed by majority consensus to be *difficult*.

The objective of this exercise was to identify categories of question prompts that had been very clearly qualified as *easy* and *difficult*. It can be seen from Figure 3 and from the discussion of the results, above, that the categories with a majority consensus of *easy* and *difficult* meet this requirement. Thus, not surprisingly, category *6, Saying yes, little opinion required, agreeing* was deemed by the majority to be *easy*. And on the contrary, categories *9, Compare, contrast, say why, 'deep' questions*, and *26, Confused, possibly multiple, meandering questions* were qualified by majority consensus as *difficult*. Interestingly, of these,

---

[37] Alderson, Clapham and Wall (1995) observe that "It is to be hoped, of course, that equivalent versions will be of a similar level of difficulty and have a similar spread of scores" (97).

the judges' qualifications in categories six and nine, in percentages were exactly opposite, being *1, 83%; 2, 17%; 3, 0%; 4, 0%* and *1, 0%; 2, 0%; 3, 17%; 4, 83%,* respectively.

Once these categories of question prompt had been very clearly identified as *easy* and *difficult*, the task remained to seek out the candidate responses they had elicited in the actual test administrations. The audiocassette recordings had been conserved for all of the tests from which question prompts had been used in the present study. Therefore, it was possible to locate the audiocassette recordings of the respective responses to the question prompts from categories 6 (the *easy* group); and 9 and 26 (the *difficult* group).

From the complete data bank of 152 question prompts, 18 occurred in the *'easy'* category *6, (Saying yes, little opinion required, agreeing)*. Due to participant mortality, this number was reduced to 11. The question prompts in the easy group were numbered E1 through E11.

Alternately, 13 questions occurred in the *'difficult'* categories, *9, (Compare, contrast, say why, 'deep' questions)*, and *26, (Confused, possibly multiple, meandering questions)*. Of the 13 questions, 3 were omitted due to participant mortality, leaving a total of 10. The question prompts in the difficult group were numbered D1 through D10.

2) Discourse analysis

In the final act of processing the data resulting from Workshop 2 in the form of the Question Prompt Category Complexity Questionnaire, the 11 *'easy'* question prompts from category 6, and the 10 *'difficult'* question prompts from

categories 9 and 26 were matched with their respective candidate responses, and the latter were transcribed. The transcribed responses were then examined, and analyzed across the *easy* and *difficult* groups. Subsequently, the candidate responses from the *easy* and *difficult* groups were analyzed through the procedures of discourse analysis of which the results are presented below.

Analysis of fluency: Type-token ratio

It will be recalled that it had been decided to delineate the question prompt responses by means of *response idea units* (*RIUs*). The fluency of responses in the *easy* and *difficult* groups was first analyzed by means of the *type-token ratio* (*TTR*) occurring in each *RIU*. It was found that there was little difference in *TTR*s in both groups. An alpha level of .05 was used for this and all statistical tests. However, significance levels (*p values*), are reported for all statistical tests performed using Statistical Analysis System (SAS) software. The *TTR* results of both groups is reported in detail in Table 6.

Table 6

Type-token ratio per response idea unit of the *easy* and *difficult* groups

Easy group	*Type-token*	*Ratio*	*Difficult* group	*Type-token*	*Ratio*
E1	37/48	.77	D1	27/57	.47
E2	66/103	.64	D2	31/54	.57
E3	43/60	.72	D3	118/329	.36
E4	57/105	.54	D4	37/59	.63
E5	57/103	.55	D5	23/28	.82
E6	80/145	.55	D6	64/141	.45
E7	58/114	.51	D7	92/210	.44
E8	61/125	.49	D8	91/213	.43
E9	85/208	.41	D9	60/97	.62
E10	40/64	.63	D10	37/61	.61
E11	94/190	.49			

A *Shapiro-Wilks test of normality* was done for each of the *easy* and *difficult* groups.[38] For the two sample groups, the *Shapiro-Wilks* statistics were close enough to unity to suggest that the distributions were normal; the *easy* and *difficult* test statistics were $W = 0.9481, p = 0.6196$, and $W = 0.9243, p = 0.3942$, respectively. The means and standard deviations of the *TTR* in the two groups are reported in Table 7. In the *easy* group, the mean *type-token ratio* (*TTR*) in the candidate responses to question prompts E1 through E11 was 0.57. The mean *TTR* in the difficult group was 0.54.

Table 7

Type-token ratio group *means* and *standard deviations*

	M	*SD*	*N*
Easy group	0.57	0.11	11
Difficult group	0.54	0.14	10

In conclusion, these results show that there was no significant difference in fluency as indicated by the presence of number of words produced (*types*) in relation to total number of words produced per *RIU* (*tokens*), in the two groups sampled.

Subsequently, an *independent t-test* was administered. The results

[38] A value of the *Shapiro-Wilks test* statistic close to unity coupled with a large *p value* indicates that the hypothesis of a normal distribution of type-token ratios should not be rejected, as it approaches normality (H_0); conversely, a low *p value* indicates a lack of normality.

indicated that there was no significant difference between the mean *type-token ratios* in the *easy* and *difficult* groups, ($t = -0.62$, $df = 19$, $p = 0.5451$). In addition, the same comparisons were made using a nonparametric *Wilcox two-sample test* and the results vindicate the values of the *t-test*, ($T = -0.7754$, $p = 0.4381$).

Analysis of fluency: Fluency frequency features

The *FFFs* examined in the present study consisted of *silent pauses, filled pauses*, *repetitions*, and *self-repairs*. These results are seen in Tables 8 and 9.

Table 8

Fluency frequency features per response idea unit: *Easy* group

	Occurrences in number and (percentage of features in total words)			
Response	*Silent pauses*[a]	*Filled pauses*[b]	*Repetitions*	*Self-repairs*
E1	2 (4.2%)	1 (2.1%)	0 (0%)	0 (0%)
E2	0 (0)	4 (3.9)	1 (1.0)	3 (2.9)
E3	2 (3.4)	12 (20)	0 (0)	1 (1.7)
E4	0 (0)	3 (2.9)	7 (6.7)	0 (0)
E5	1 (1.0)	6 (5.8)	2.5 (2.4)	1 (1.0)
E6	0 (0)	2 (1.4)	11 (7.6)	0 (0)
E7	1 (0.9)	20 (17.6)	2 (1.8)	0 (0)
E8	2 (1.6)	24 (19.2)	2 (1.6)	0 (0)
E9	3 (1.4)	41 (19.7)	13.5 (6.5)	1 (0.5)
E10	0 (0)	11 (17.2)	4 (6.3)	1 (1.6)
E11	1 (0.5)	24 (12.6)	4 (2.1)	2 (1.1)

[a]*Silent pauses* are defined as those silent speech hesitations of 1 second or more, in the coding conventions of the present study.
[b]*Filled pauses* are defined as those speech hesitations which are filled expressions such as <uh> and <um>, in the coding conventions of the present study.

The *FFFs* of candidate responses in the two groups were analysed and the results recorded. Additionally, these data were converted to percentages in order to account for differences in response length.

Table 9

Fluency frequency features per response idea unit: *Difficult* group

	Occurrences in number and (percentage of features in total words)			
Response	*Silent Pauses*	*Filled pauses*	*Repetitions*	*Self-repairs*
D1	0 (0%)	1 (1.8%)	1 (1.8%)	0 (0%)
D2	2 (3.7)	6 (11.1)	2 (3.7)	1 (1.9)
D3	1 (0.3)	16 (4.9)	18 (5.5)	1 (0.3)
D4	4 (6.8)	15 (25.4)	7 (11.9)	0 (0)
D5	0 (0)	1 (3.6)	0 (0)	0 (0)
D6	2 (1.4)	16 (11.3)	11 (7.8)	2 (1.4)
D7	0 (0)	12 (5.7)	8 (3.8)	2 (1)
D8	0 (0)	8 (3.8)	9 (4.2)	3 (1.4)
D9	0 (0)	0 (0)	4 (4.1)	0 (0)
D10	1 (1.6)	6 (9.8)	0 (0)	0 (0)

The results of a *Chi square contingency table analysis* revealed that there were strong, significant differences in the *FFFs* between the *easy* and *difficult* groups, χ^2 (3, $N = 376$) = 13.32, $p = 0.004$. However, since this test did not identify which feature effect accounted for the differences, the data was further investigated with the use of a statistical procedure of increased power.

A secondary *Chi square contingency table analysis* of simultaneous categories was done for each of the two groups. For example, the frequency of *silent pause* events was compared to the frequency of all of the *filled pause*, *repetition*, and *self-repair* events. Thereafter, a fluency feature of each of the two

groups was isolated and compared with the remaining three collapsed features until all four features had been compared. The results of the collapsed *Chi square* for the feature '*silent pauses*' are reported in Table 10.

Table 10

Chi square contingency table for *silent pause* effect

	Silent pauses	*Filled pauses, repetitions* and *self-repairs*	Total
Easy group	12	204	216
Difficult group	10	150	160
Total	22	354	376

χ^2 (1, $N = 376$) = 0.0805, $p = 0.7767$, ns

Another manner of regarding the Table 10 silent pause feature results is by noting that among the easy group, 12 of 216 events (5.6%), were characterized as silent pauses, while among the difficult group, 10 of 160 (6.3%) were so characterized. This difference in proportions is not statistically significant, χ^2 (1, $N = 376$) = 0.0805, $p = 0.7767$, ns.

The results of the collapsed *Chi square* for *filled pauses* found a strongly significant effect of this fluency feature, χ^2 (1, $N = 376$) = 12.3595, $p = 0.0004$. Indeed, 148 of 216 events (68.5%) in the *easy* group resulted in *filled pauses*, but only 81 of 160 events (50.6%) occurred in the *difficult* group.

The collapsed *Chi square test* also revealed differences between the groups in the *repetitions* feature, χ^2 (1, N = 376) = 11.1860, p = 0.0008. In fact, this effect was slightly stronger than that of *filled pauses*. Again, it is noteworthy that in the easy group 47 of 216 events (21.8%) resulted in repetitions, while 60 of 160 events (37.5%) resulted in repetitions in the difficult group. The differences are highly significant.

Finally, the results of the investigated *self-repair* feature in the two groups showed no significant differences, χ^2 (1, N = 376) = 0.4289, p = 0.5125, ns. In the easy group, 9 of 216 events (4.2%) were characterized as self-repairs, and 9 of 160 events (5.6%) occurred in the *difficult* group. These differences were not significant.

To summarize, the results of the *FFF* show that no significant differences existed between the *easy* and *difficult* groups in terms of *silent pause* and *self-repair* effects. However, strong significant differences between the two groups were evident following the investigation of the *filled pause* effect. In that case, the *easy* and *difficult* groups resulted in *filled pauses* in 68.5% and 50.6% of the total fluency events, respectively, (p = .0004). Thus, significantly more *filled pauses* occurred in the group tested with easy question prompts than in the group tested with difficult ones.

This would suggest that the *easy questions* group were functioning at a lower level of L2 oral proficiency as demonstrated by their response performance. Furthermore, since those candidates tested with difficult questions used *filled pauses* significantly less often, it is possible that these individuals accomplished

the task with more ease than had the other group. All of this leads to the possibility that the test raters had noticed the differences in proficiency, and had divided the groups prior to administering the opinion test function. It also suggests that the consequence of their observations could have been that they had selected qualitatively different kinds of question prompts (easy or difficult), as a result.

Similarly, strong significant differences were found in the *easy* and *difficult* groups in the occurrences of *repetitions* in their responses. However, in the case of *repetitions*, significantly fewer occurred in the *easy question* group (21.8%), in contrast to the *difficult question* group (37.5%), (p = .00008). This data suggests that the use of difficult question prompts may have affected candidate fluency in terms of *word repetitions* in the sample studied. Possibly the greater cognitive demands of the difficult questions resulted in hampered fluency in this regard. However, these results in some measure challenge those of the *filled pauses* since the latter indicate that the greater cognitive requirements of the difficult questions resulted in fewer *filled pauses* in the *difficult* versus the *easy* group.

Next, in an exercise related to computation of frequency, interval data was analyzed in the form of the total *pause time* in seconds, in the *RIU*s of the *easy* and *difficult* groups. The *total pauses* in each *RIU* of the group samples were measured in seconds. The *means* and *standard deviations* of the *pause time* data in the two groups are reported in Table 11. *Shapiro-Wilks tests for normality* were performed on the distributions of the two groups. The results indicate that while

Table 11

Total *pause time* in seconds: Group *means* and *standard deviations*

	M	SD	*N*
Easy group	2.27	2.33	11
Difficult group	1.80	3.08	10

the *easy* group distribution was barely normal ($W = 0.8591, p = 0.0561$), the *difficult* group distribution was not normal ($W = 0.6464, p = 0.0002$). In view of the fact that normality was not achieved, a nonparametric *Wilcoxon two-sample test* was administered, and its results indicate no significant difference in the two groups ($T = -0.8095$, normal approximation $p = 0.4182$; *two-tailed t-test* approximation $p = 0.4277$; ns). The same data was subjected to a *t-test* for which the result is less valid due to lack of normality. The results of a *t-test* of equal variances yielded the same conclusion of no significant difference between the groups ($t = -0.40, df = 19, p = 0.6944$, ns).

It is evident that the *easy* and *difficult* groups are similar in *pause times* in light of the fact that their respective *means* differ by only .47, or close to half a second. There is more variation in the groups' *standard deviations*, which is accounted for by the presence of an outlier, the candidate response to question prompt D4. This candidate demonstrated considerable weakness in fluency in responding to the test question, stopping to pause four times in the *RIU*. The *pause time* in seconds of each *RIU* in the *easy* and *difficult* groups appears in Table 12.

Table 12
Total *pause time* per *response idea unit* of the *easy* and *difficult* groups

Easy group	*Pauses* in seconds	*Difficult* group	*Pauses* in seconds
E1	2	D1	0
E2	0	D2	2
E3	6	D3	1
E4	0	D4	10
E5	3	D5	0
E6	0	D6	3
E7	1	D7	0
E8	4	D8	0
E9	6	D9	0
E10	0	D10	2
E11	3		

Finally, the results of the analysis of the *pause time* per *RIU* have clearly demonstrated that in the samples examined, no significant differences in pause time exist between the two groups.

Analysis of accuracy

For the purposes of the present study, a binary distinction was made between target and nontarget usage forms. Target forms will be said to include those of a proficient speaker, while nontarget forms will include all others, such as those in evidence in the interlanguage of much of L2 discourse. Therefore, the discourse accuracy in the *easy* and *difficult* groups was measured by frequency counts of target or nontarget forms of *verb morphology* and *noun usage*. The former were addressed by assessing bound morphemes of *subject-verb agreement*, in the presence of *an obligatory subject and/or verb*, and of appropriate *tense marking*; while of the latter *common, compound, and abstract noun usage* was examined. The results of *verb morphology* frequency analysis for the *easy* and *difficult* groups appear in Tables 13 and 14, respectively.

Table 13
Accuracy of *verb morphology* in occurrences per *response idea unit* in the *easy* group

	S-v agreement		*Tense marking*		*Obligatory s, v*	
Response	Target	Nontarget	Target	Nontarget	Target	Nontarget
E1	6	0	7	0	12	2
E2	19	3	18	4	16	2
E3	11	2	11	2	11	2
E4	12	3	10	5	7	8
E5	14	9	18	5	22	1
E6	21	5	20	6	24	2
E7	15	2	14	1	11	0
E8	17	0	12	5	12	5
E9	27	3	28	2	27	3
E10	3	3	4	2	1	5
E11	20	8	23	5	24	4
Totals	165	38	165	37	167	34

Table 14
Accuracy of *verb morphology* in occurrences per *response idea unit* in the *difficult* group

	S-v agreement		*Tense marking*		*Obligatory s, v*	
Response	Target	Nontarget	Target	Nontarget	Target	Nontarget
D1	7	2	8	1	9	0
D2	10	0	9	1	9	1
D3	32	13	44	1	40	5
D4	11	1	8	4	8	4
[a]D5	3	0	3	0	3	0
D6	24	3	25	2	22	5
D7	34	5	30	9	34	5
D8	25	11	29	7	30	6
D9	10	5	14	1	13	2
D10	4	3	7	0	3	4
Totals	160	43	177	26	171	32

[a]Response D5 was atypically much shorter than the others at 28 words, which may have contributed to its absence of nontarget forms.

A *Chi square contingency table analysis* of simultaneous categories was performed, isolating one accuracy feature for comparison purposes across target and nontarget usage. The results showed no significant effects for *subject - verb agreement*, as shown in the contingency table, Table 15.

Table 15

Chi square contingency table for *subject - verb agreement* effect

	Target	Nontarget	Total
Easy group	165	38	203
Difficult group	160	43	203
Total	325	81	406

χ^2 (1, $N = 406$) = 0.3856, $p = 0.5346$, ns

These results show similarities between the two groups in this instance; among the easy questions, 165 of 203 *subject-verb agreement* events (81.3%) were found to be target events; in the *difficult* group 160 of 203 events (78.8%) were found to be target events. Correspondingly, nontarget events in the *easy* and *difficult* groups were 38 of 203 (18.7%); and 43 of 203 (21.2%), respectively. Thus, there were no significant differences in *subject - verb agreement* effects in the *easy* and *difficult* groups. Interestingly and coincidentally, the total events of *subject - verb agreements* in the responses was equal at 203, in both groups.

The same procedure produced a result of no significant effects for *tense marking*, χ^2 (1, $N = 405$) = 2.3392, $p = 0.1262$, ns. In this case, in the *easy* group

target forms accounted for 165 of 202 events (81.7%), while 177 of 203 (87.2%) occurred in the *difficult* group. Nontarget forms accounted for 37 of 202 (18.3%) in the *easy* group compared to 26 of 203 (12.8%) in the *difficult* group.

The presence of *obligatory subjects and, or verbs* were computed in the same manner, resulting in no significant effects found in the *easy* and *difficult* groups, χ^2 (1, $N = 404$) = 0.098, $p = 0.7542$, ns. Target events in the *easy* group amounted to 167 of 201 (83.1%), compared to 171 of 203 (84.2%) of the *difficult* group. Nontarget usage in the *easy* group was found to be 34 of 201 (16.9%), and 32 of 203 (15.8%) in the *difficult* group.

Common, compound, and abstract noun usage was also not found to be significantly different across the *easy* and *difficult* groups, χ^2 (1, $N = 243$) = 1.2841 $p = 0.2571$, ns. Detailed results of the lexical analysis for the two groups appears in Table 16.

In conclusion, it was found that there were no significant differences in the *easy* and the *difficult* candidate test responses in terms of output accuracy. Moreover, the results of the accuracy discourse analysis demonstrate close similarities in the data of the two groups. This is in part due to the fact that two of the features under investigation, *subject - verb agreement* and the presence of *obligatory subject or verb elements*, were themselves closely associated. Therefore, it follows that investigations of these features can be expected to render results of a similar nature. However, in the other features examined, (*appropriate tense marking*; and *common, compound, and abstract noun usage*), there was a great deal of similarity of form use across the two groups. These

results lead to a conclusion that the *easy* and *difficult* question group performed in a homogeneous manner in terms of the output accuracy in the features investigated.

Table 16

Accuracy of a *lexical form* in occurrences per *response idea unit* in the *easy* and *difficult* groups

	Common, compound and abstract noun usage				
Easy group response	Target	Nontarget	*Difficult* group response	Target	Nontarget
E1	5	2	D1	1	0
E2	7	2	D2	4	0
E3	10	0	D3	23	4
E4	13	2	D4	4	0
E5	4	2	D5	4	1
E6	14	3	D6	7	5
E7	7	3	D7	13	2
E8	11	5	D8	18	7
E9	15	1	D9	9	1
E10	5	0	D10	5	1
E11	21	0			
Totals	112	20		88	23

Analysis of complexity

Discourse complexity in the candidate responses was analyzed by means of *clause subordination* in the two groups under investigation. In this process, *AAS units* were used to separate the oral language into discrete units of sentence-like structures. Foster, Tonkyn and Wigglesworth's (2000) definition of clause subordination was used for this purpose in the present study. Foster et al. note that

a subordinate clause "will consist minimally of a finite or non-finite verb element, plus at least one other clause element (Subject, Object, Complement or Adverbial)" (p.366).

Thus, the clauses in AAS units were examined for subordination, and the results computed. The results of discourse complexity analysis for the *easy* and *difficult* groups are found in Tables 17 and 18.

Tables 17 and 18 illustrate the close similarity of the two groups in discourse structure and length. Not surprisingly, using *Chi square tests*, discourse complexity was not found to be significantly different in total clauses across the *easy* and *difficult* groups, χ^2 (1, $N = 258$) = 0.016 $p < 0.05$, ns). Similarly, in the same test, clause subordination was not found to differ significantly in the two groups, χ^2 (1, $N = 44$) = 2.273 $p < 0.05$, ns).[39]

Using *Chi square tests for contingency tables*, in the *easy* group, 128 of 251 events (51.5%) were the *total number of clauses*, while in the *difficult* group 130 of 242 events (53.7%) represented the *total clause* number. Again, the *easy* group had 27 of 251 events (10.8%) as the number of *subordinate clauses*, and the *difficult* group had 17 of 242 events (7.0%) counted as total *subordinate clauses*.

In the *easy* group target the number of *AAS units* in the RIUs of the *easy* group and *difficult* groups did not differ significantly, χ^2 (2 $N = 2.1299$) = 2.1299 $p = 0.3447$, ns.

---

[39] These *Chi square tests* were the only statistical tests done using AB STAT software. Consequently, exact *p values* were not available for these results.

Table 17

Syntactic analysis of *clause structures* in the *easy* group

	Number of *AAS-units* and *clauses*			
Responses	*AAS-units*	*Total clauses*	*Subordinate clauses*	*% subordinate clauses* per *total clauses*
E1	5	7	1	(14.3)
E2	11	14	5	(35.7)
E3	3	4	1	(25.0)
E4	7	11	4	(36.4)
E5	6	11	3	(27.3)
E6	12	19	6	(31.6)
E7	9	9	0	(00.0)
E8	10	11	0	(00.0)
E9	16	22	6	(27.3)
E10	3	3	0	(00.0)
E11	14	17	1	(05.9)
Totals	96	128	27	(*M* 18.5%)

Table 18

Syntactic analysis of *clause structures* in the *difficult* group

	Number of *AAS-units* and *clauses*			
Responses	*AAS-units*	*Total clauses*	*Subordinate clauses*	*% subordinate clauses* per *total clauses*
D1	4	6	2	(33.4)
D2	3	5	2	(40.0)
D3	26	33	5	(15.2)
D4	7	8	1	(12.5)
D5	2	3	0	(00.0)
D6	8	12	3	(25.0)
D7	15	20	1	(05.0)
D8	16	26	2	(07.7)
D9	9	12	1	(08.3)
D10	5	5	0	(00.0)
Totals	95	130	17	(*M* 14.7%)

Finally, these results show that there were no significant differences between the *easy* and *difficult* groups in terms of clause complexity. Moreover, the close similarity of candidate responses in the two groups in terms of the complexity data results would suggest that the population sampled came from a very homogeneous group. This in turn suggests that from the point of view of output complexity, these candidates exhibited very similar qualities in their test performances.

In conclusion, the quantitative analyses of the Phase 2 results presented in this chapter show some paradoxical trends. A clear consensus was shown in the results of the Question Prompt Category Complexity Questionnaire responses, yet a considerable lack of consensus was demonstrated as well. Similarly, discourse analyses of candidate response fluency in the *easy* and *difficult* groups showed significant differences for filled pauses and repetitions, but no significant between-group differences for the other fluency features measured.

However, discourse analyses of response accuracy and complexity in the *easy* and *difficult* groups, indicated that there were no significant differences in these speech characteristics. In Chapter 6 following, these findings will be addressed in greater detail.

Chapter 6

Conclusions

Introduction

The present research has raised several questions related to method effects in the task of *supporting an opinion*, in an ACTFL-variant oral proficiency test, the SLE:OI. Specifically, it sought to investigate the kind of discourse generated from the use of different question prompts. The question of method effects was of particular interest given that SLE:OI raters have considerable latitude in the choice of question prompts (and therefore test methods) available to them for use in the same, and across different test administrations.[40] This liberty is due in large measure to the conversational formant of oral proficiency interview tests in general.

The issue of parallel test forms

The effect of task variation particularly on reliability in oral proficiency tests was of interest in the present study in view of the hypothesis that the employment of question prompts that were profoundly different in quality would result in non-parallel test forms. Certainly the outcome of administering non-parallel tests is that unfair advantages or disadvantages to some candidates may result. Of particular concern in the present research was the case of borderline candidates (whose test performance straddled the rating border between

---

[40] It will be recalled that in the SLE:OI, the rater is also an interviewer, administering the test independently.

intermediate B-level, and advanced C-level, termed B/C borderline cases). It was hypothesized that weaker (or B/C) candidates might fail to accomplish the task of *supporting an opinion* when unduly difficult question prompts were used (see Chapter 3, Figure 2).

The data of the Question Prompt Category Complexity Questionnaire results suggest the inherent difficulties involved in any exercise of qualifying question prompts, or question prompt categories in this case. This was evident in view of the fact that the respondent-judges were unable to reach a consensus in over half (15 out of 26) of the categories surveyed. This underscores the intrinsic challenges faced by oral proficiency test administrators in seeking to ask questions of equal value over tests. Alderson, Clapham and Wall (1995) recommend that "... it should be emphasised that that the interview needs to be carefully structured so that the aspects of the test which are considered important are covered with each student, and each student is tested in a similar way" (p. 62). In addition, Douglas (2000) has suggested that "the rhetorical form of the message is often as important as the content, and should reflect the norms of the target language use situation" (p. 61).

Conversely, in ACTFL and ACTFL-variant oral proficiency tests such as the SLE:OI, it is common practice to allow for sizeable variation in both form and content of question prompts. Moreover, the findings of the present study show that raters succeeded in arriving at a very clear consensus of three question prompt categories, placing them in bipolar *easy* and *difficult* groups.

The raters judged *easy* those question prompts requiring candidates to *say yes, with little opinion required, agreeing*; and *difficult* those question prompts requiring candidates to *compare, contrast, say why ('deep' questions)*, and *confused, possibly multiple, meandering questions* (see Chapter 4).

If the premise *were* to be accepted that oral proficiency tests should without compunction, include wide content variation in the form of their question prompts, then it must also be assumed that the various question prompts *would* constitute parallel test forms. This contradicts the findings of the present study in which judges in consensus found qualitative differences in question prompt categories. Yet if the above premise *were* accepted, it would follow that in the present study, there is no qualitative difference between the *easy* and *difficult* question prompt groups. And if this argument is carried a logical step further, may it may be assumed that there is no difference between the two kinds of question prompts identified by judges as constituting the *difficult* group, those which required candidates to *compare, contrast, say why ('deep' questions)*, and those that were considered to be *confused, possibly multiple, meandering questions*? Could these two types of *difficult* questions really be considered parallel test forms? Of course, only with empirical evidence could test forms be determined with any accuracy to be equivalent forms.

My argument here is intended to be somewhat fanciful. It is intended to illustrate the importance of giving serious consideration to controlling as much as possible for question prompt variation in oral proficiency interview tests (see Implications and recommendations). Moreover, in my professional experience as

an SLE:OI rater-interviewer, I know of no SLE:OI trainer who would seriously entertain the above premise. SLE:OI trainers give coordinated attention to training raters-interviewers to maintain as much uniformity in test content as possible. The assiduity of test trainers in the instance of the SLE:OI is fortunate. Nonetheless, ACTFL and ACTF-variant oral proficiency test development leaves a test design loophole allowing for a plethora of non-parallel test forms to flourish.

The research question and the research findings

> Revisiting the research question of the present study, it was:
>
> Is there a difference in speech samples elicited by different question prompts in the task of *supporting an opinion* in an oral proficiency interview test?

The findings support the premise that it is possible that *in general*, question prompts used in the task of *supporting an opinion* elicit discourse of comparable accuracy and complexity. Thus, candidate response accuracy as measured by *verb morphology* and *lexical accuracy* was not affected in the groups tested with *easy* and *difficult* questions. Similarly, response complexity as measured by *clause subordination* was not affected in the sample groups. However, *particular* results of the present research with regard to discourse fluency features demonstrated different and varied effects.

The fluency of responses in the two groups under investigation showed no effects for *type-token ratio*, or for *silent pause* or *self-repair* frequency. Additionally, there was no effect for total *pause time* between the two groups.

However, there were strong significant effects for *filled pauses* and *repetitions* in the groups tested with *easy* and *difficult* question prompts.

Unexpectedly, more *filled pauses* were found to occur in the group tested with *easy* question prompts than in the group tested with *difficult* ones. This could be explained by the suggestion that the group given *easy* questions had previously demonstrated a lower level of L2 oral proficiency than the *difficult* question group, (who displayed *filled pauses* significantly less often in response to the opinion task). By extension, this would also indicate that raters had already noticed this demonstrated weakness in proficiency in the B/C group prior to testing the *supporting an opinion* task, since they had elected to ask this group easier questions. Given that SLE:OI rater training discourages the use of question prompts of inconsistent complexity, it is likely that the choice of two groups of question prompt was unconscious on the part of raters.[41]

Additionally, those candidates demonstrating stronger B/C proficiency were shown to have less fluency impediment when asked *difficult* questions, as demonstrated by their significantly fewer exhibits of *filled pauses*. Thus, the evidence suggests that those candidates perceived by raters as having stronger L2 abilities are more likely to be asked *difficult* questions. If this were indeed the case, then theoretically it could be expected that strong borderline (B/C) candidates would have an advantage over less proficient (B/C) candidates; they

---

[41] The rater behaviour in question is qualified as unconscious in view of my knowledge of SLE:OI rater training and professional attitudes, and from the evidence of Workshops I and II in which raters reiterated that in test administrations, their intention is to ask test questions of equal difficulty.

would not both be tested with question prompts of the same order of difficulty. Thus, the stronger candidates would have a better chance of succeeding in the task probe, since it would be less demanding for them than it would be for the weaker candidates.[42] This was the original hypothesis illustrated in Figure 2 (See Chapter 3).

Therefore, it can be concluded that there was no evidence of method effects in the two groups when response discourse accuracy, complexity, or several features of fluency were analyzed. However, method effects were found when *easy* and *difficult* question prompts were used as demonstrated by two fluency features, *filled pauses* and *word repetitions*. The latter is not surprising in view of recent research where method effects have been found (for example, [Ellis, 1987; Smith, 1992; Tarone, 1979, 1988; cited in Bachman & Cohen, 1998], Turner & Upshur, 1995, Upshur & Turner, 1999). Additionally, Bachman and Cohen have noted that "different tasks can elicit different accuracy rates" (p. 83). Similarly, Norris, Brown, Hudson and Yoshioka (1998) recommend that in second language performance test development, a question which should be asked is "What are the difficulty levels of the tasks in terms of human performance?" (p. 141).

Some researchers, for example Lumley and Brown (1996), found rater behavioural factors that appeared to affect the level of interaction difficulty in oral

---

[42] In ACTFL and ACTFL-variant oral proficiency test, a 'probe' indicates testing at a higher level than candidate ability, in order to determine a ceiling of proficiency. The task of *supporting an opinion* is considered a probe of intermediate-high, and high-level candidates.

proficiency testing (cited in McNamara, 1997). Among the factors which increased the difficulty of test tasks were passivity, interrupting, and the use of sarcasm; on the other hand, task difficulty was eased by the degree interlocutors attended to factual questions, or tailored the questions to simpler forms to help candidates.

Conversely, if candidates were mistakenly perceived by raters as having stronger L2 abilities and were asked difficult questions, when these candidates were *in actuality* of weaker (B/C) ability, the result could be candidate inability to accomplish the test task. Thus, problems of test fairness would result. These results underscore the necessity of maintaining a bank of question prompts of parallel difficulty.

McNamara (1995, 1996) has shown that even highly-trained raters may under or over-rate subjective performances to a measurable extent (though not necessarily to the extent that test scores are influenced). Certainly, rater perceptions of candidate proficiency can never be perfectly correct in all cases. Furthermore, rating is particularly problematic in cases of demonstrated borderline performance. This is the case in language testing, and indeed it is the case in all subjective testing.

Implications and recommendations

The implications for rater training are clear. Not only should ACTFL-variant test raters be trained to 'bias for best' (Swain, 1995), they should also 'give the benefit of the doubt' in assigning questions of an easier order when

weaker B/C candidates could be disadvantaged by unduly difficult ones, *at the very least.*[43]

However, a far better alternative would be to ensure that question prompts have been determined to be of equal value by establishing a bank of question prompts which have been empirically determined to be parallel forms. Adopting either of these measures would prevent placing weaker borderline candidates at a disadvantage.

In the present study, the results of the discourse analysis of response *word repetition* showed strong significant differences between the *easy* and *difficult* groups. As anticipated, it was found that significantly *fewer* repetitions occurred in the *easy* question group in contrast to the *difficult* question group. These data suggest that the two groups responded quite differently to the *easy* and *difficult question prompts* they encountered. This may be explained by the assumption that the greater cognitive demands of the *difficult question* task impaired fluency in this regard as candidates sought to process the more complex content of the prompts. It is well known that people often resort to repetition while attending to complex ideas. Then again, it is possible that *filled pauses* may be employed to the same effect. It can only be said with certainty that the results of the present study demonstrate that some differences in candidate response occurred as a result of the use of *easy* or *difficult question prompts*. Further speculation would require analysis of the cognitive processes influencing the use of these fluency features, which is beyond the limitations of the present research.

---

[43] Indeed, Bachman and Palmer (1996) have advised that test rubrics "should be designed with the least proficient test takers in mind" (p. 141).

The original research question pertained to the effect of using different question prompts in the *supporting an opinion* test task. Related to this, an associated question arises as to the overall suitability of the task of *supporting an opinion* in the SLE:OI (see Chapter 3). How accurately does the task of *supporting an opinion* reflect the *TLU* of test takers?

The ACTFL Proficiency Guidelines (1986) for speaking ability contend that in the advanced-high L2 speaker "there is emerging evidence of ability to *support opinions*," and that the superior speaker "can *support opinions*" (ACTFL, online, retrieved June 20, 2000). The validity of basing second language assessments on this kind of a priori determination appears to be outmoded, as has been observed by many researchers since the publication of the Guidelines nearly twenty years ago (see Chapter 2). Certainly current knowledge about *TLU* domains as defined by Bachman and Palmer (1996) has transformed the LT community's approach to language test development, basing it on evidence rather than on intuitive judgements.

Recent research in Languages for specific purposes (LSP) testing has increased our collective understanding of how to more accurately tailor tests to specific candidate circumstances (Douglas, 2000). Consequently, there can be no doubt that determination of test content validity is now better served by fitting test tasks to empirically measured workplace *TLU* domains.[44]

---

[44] My anecdotal impression is that the kind of candidates truly comfortable with the opinion task is generally limited to lawyers, whose work *TLU* clearly and closely corresponds to the test task of *supporting an opinion*.

Bailey (1998) pointedly illustrates the advantages of basing oral performance testing on *TLU*, in suggesting three ways an air flight crew might best be tested for L2 oral proficiency: by a paper and pencil test, by an oral proficiency interview test, or by all passing "an authentic test of oral English communication in an air-to-ground radio setting using topics based on recordings of actual conversations between air traffic controllers and airline pilots" (p. 208). Surely, the third option would be the most compelling; it effectively matches *TLU* with test task.

It appears that the use of the ACTFL and ACTFL-variant test task of *supporting an opinion* is clearly problematic since in many instances of actual work duties, *supporting an opinion* does not occur. The danger of not matching professional *TLU* to test tasks runs the real risk of testing an unused and arbitrary construct. This is one weakness of the ACTFL and ACTFL-variant testing tradition.

For these reasons, and in view of the variation in results demonstrated in the present research, suggesting the presence of possibly disadvantageous method effects in the task of *supporting an opinion*, I conclude that the necessity of including this task in the SLE:OI, be reviewed. A review of this sort would serve two purposes. Firstly, it would demonstrate if the *supporting an opinion* task actually does reflect *TLU* in the SLE:OI population of test candidates. Secondly, if this were found not to be the case, further statistical study could be done to determine if test scores would be influenced by its exclusion from the test.

If upon completion, the task review exercise proved the task to be redundant, the outcome could be the modification of the SLE:OI to a shorter, and therefore more economical test. Additionally, it would relieve the test of what may be its primary source of unreliability.

In the meanwhile, I would additionally recommend that the following steps be taken:

1. that the officials responsible for SLE:OI testing instigate a study in order to establish appropriate norms of task difficulty in the task of *supporting an opinion.*
2. that a bank of parallel question prompts be created in order to ensure that all candidates would be tested with equal or parallel test forms (as is currently the procedure followed in the CASE test [Lazaraton, 1996]).
3. additionally and essential to 2. above, that the question prompts in the bank of parallel test forms be empirically determined through statistical means to be of equitable difficulty levels, prior to making any assertions that they represent equal forms.

Limitations of the study

Limitations of the present study have included the low number of candidate participants (21) who were determined to have been asked qualitatively easy and difficult questions. Further research using a larger sample size would allow for an investigation of what influence on test scores might be incurred as a

result of method effects of the use of *easy* and *difficult question prompts*, in an ACTFL-variant test.

With regards to *TTR*, several researchers have used the protocol in discourse analysis, and their lack of censure would indicate that they were satisfied with the *TTR*. (Crookes, 1989; Douglas, 1994; Tomiyama, 2000; Wigglesworth, 1997b). Yet Vermeer (2000), while calling the *TTR* the "most famous," device of lexical measurement, has nonetheless raised various doubts as to its usefulness in identifying lexical richness (p.65). Vermeer argues in favour of basing lexical measures not on the *TTR*, but rather on the "degree of difficulty of the words used, as measured by their (levels of) [sic] frequency in daily language input" (p. 65). Vermeer's contention that the *TTR* may be "the worst measure of lexical richness" is worrisome in view of the fact that in the present study group differences in *TTR* were expected but not achieved (p. 69).

On the other hand, the present study's analysis of *common, compound* and *abstract noun usage*, used to measure L2 accuracy, is effectively a measure of *lexical* accuracy. Thus, to an extent it serves the same purpose as that proposed by Vermeer. (Interestingly, the results of the present research indicated that in both the *TTR* and the examination of *common, compound* and *abstract noun usage,* no differences between the *easy* and *difficult* groups were found).

Furthermore, the methodology of qualitatively categorizing question prompts may have in itself to an undetermined extent influenced the categorization outcomes. This may be partly due to the judges' unfamiliarity with the protocol. McNamara (1997) comments on a psychometric view of this, in

noting that "Linacre (1989), in a brilliant discussion, has shown that allocation of instances to categories by judges is a probalistic, not a deterministic phenomenon" (p. 456).

Moreover, in the present research Phase 1 served as the basis of Phase 2. Therefore, any inaccuracies arising from the Phase 1 data would ultimately influence the Phase 2 results.

In the case of the responses to the Question Prompt Category Complexity Questionnaire, task unfamiliarity would not be expected to have measurably influenced responses since the judge-respondents were SLE:OI raters, accustomed to making qualifications about question prompt task difficulty in their daily professional lives. Nonetheless, the response data may have to some extent been influenced by the subjective nature of the task. Furthermore, it is possible that there may have been a small delayed time effect, which may have influenced the judges' familiarity with the categorization data, due to the fact that the questionnaire was administered some weeks subsequent to the first question prompt categorization exercise of Workshop 1.[45]

Finally, by virtue of combining qualitative and quantitative methods to second language research such as the present study entails, the results may afford a more comprehensive view of the ever-elusive truth. Indeed, Boland (1992) compared the two approaches in the following:

---

[45] This effect is qualified as small in view of the fact that extensive efforts were made in Workshop 2 to re-familiarize the judges with the categorization material of the previous Workshop 1.

> Qualitative methods lend themselves to discovering meanings and patterns while quantitative methods seek causes and relationships .... Researchers in the qualitative mode seek understanding through inductive analysis, moving from specific observation to the general. Quantitative analysis, on the other hand, employs deductive logic, moving from the general to the specific, i.e. from theory to experience. (Boland, 1992, p. 1-2)

The introduction of bias is an ongoing threat to the design and accomplishment of any study, and it is particularly the case when qualitative research is undertaken. This is in part due to the fact that a greater subjective element exists in the data collection than would be the case with quantitative methods, (though this threat exists in any kind of research.) For example, in the present study the Phase 1 data collection protocols involved quite subjective qualification exercises.

It was intended and hoped that this variable of subjectivity would be diminished by the fact that the chosen participants were highly trained judges, familiar with evaluating question prompt appropriateness. The study was also limited to some extent by the fact that it could not have included more qualitative and quantitative analyses. This was unfortunately beyond the possibility of the present research.

Suggestions for further research

I lend my voice to the many previous calls in the field of language testing, for more qualitative and quantitative research into the discourse generated in oral proficiency interview tests. Moreover, it is important to further study the constructs occurring in the target language use domains of ACTFL and ACTFL-

variant test candidates in order to more accurately tailor tests to individual test candidates, by restricting test tasks to more appropriate domains.

Concluding remarks

Messick (1989), in his seminal chapter on validity in educational testing, has determined that "content validity provides judgemental evidence in support of the domain relevance and representativeness of the content of the test instrument, rather than evidence in support of inferences to be made from test scores" (p. 17). Messick's message would appear to corroborate the need for language tests more closely based on *TLU* than is the current practice of ACTFL and ACTFL-variant tests.

Furthermore, McNamara advocates the incorporation of empirical evidence in test design, in the following, "Validity is not automatically achieved through test design alone; there must be a subsequent empirical demonstration of this relationship through investigation of data from actual performances, in test trials and under operational conditions" (p. 456).

Similarly, in her study of the discourse elicited under circumstances of task variation in an oral proficiency test, Wigglesworth (1997a), underscored the need for "routinely subjecting test data to rigorous discourse analysis, and to integrating discourse analysis into the process of test validation" (p. 47).

The present research has sought to respond to the appeals of common sense as well as to those of the language testing community. Its methodology also demonstrates the combination of qualitative and quantitative analysis to reach an end. It contributes to language testing literature by presenting empirical evidence

of the kind of discourse generated in an ACTFL-variant oral performance test. In addition, and with the generous support of SLE:OI test officials, it incorporates the use of empirical analysis into SLE:OI testing practice.

References

AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGES (1986). *ACTFL Proficiency guidelines.* Retrieved June 20, 2000 from the World Wide Web: http://www.actfl.org/

AMERICAN COUNCIL ON THE TEACHING OF FOREIGN LANGUAGES (2001). *Proficiency Testing.* Retrieved April 19, 2001 from the World Wide Web: http://www.actfl.org/

Alderson, J.C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: CUP.

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* Oxford: OUP.

Bachman, L.F. & Cohen, A.D. (Eds.) (1998). *Interfaces between second language acquisition and language testing research.* Cambridge: CUP.

Bachman, L.F., Davidson, F. & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13* (2), 125-150.

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

Bachman, L.& Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: OUP.

Bachman, L.F. & Savignon, S.J. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *The Modern Language Journal, 70* (iv), 380-390.

Bailey, K.M. (1998). *Learning about language assessment*. Pacific Grove: Heinle & Heinle.

Barnwell, D. (1989). 'Naïve' native speakers and judgements of oral proficiency in Spanish. *Language Testing, 6* (2), 152-163.

Boland, P. (1992). *Qualitative research in student affairs. Eric Digest.* Retrieved February 20, 2001 from the World Wide Web: http://www.imtcsamba.hct.ac.ae/assessment/ltrfile/ltr.html

Breiner-Sanders, K., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL Proficency Guidelines – Speaking, revised 1999. *Foreign Language Annals, 33* (1), 13-18. Retrieved April 23, 2001 from the World Wide Web: http://www.actfl.org/

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

Canale, M. & Swain, M. (1981). A theoretical framework for communicative competence. In A. Palmer, P. Groot & G. Trosper (Eds.). *The construct validation of tests of communicative competence.* (p.31-36). Washington D.C.: TESOL.

Clapham, C.M. & Corson, D. (Eds.). (1997). *Encyclopedia of language and education, volume 7: language testing and assessment.* Dordrecht, The Netherlands: Kluwer Academic Publishers.

Clark, J. (1988). Validation of a tape-mediated ACTFL/ILR-scale based test of Chinese speaking proficiency. *Language Testing, 5* (2), 187-205.

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

Cole, G. & Neufeld, D. (1991). Les tests d'évaluation de langue seconde de la Fonction publique du Canada. *Actes du colloque AQEFLS-McGill, 12* (3-4), 47-63.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition, 11*, 367-383.

Crookes, G. (1990). The utterance, and other basic units for second language discourse analysis. *Applied Linguistics, 11* (2), 183-199.

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America, 107* (2), 989-999. Retrieved May 9, 2001 from the World Wide Web: http://lands.let.kun.nl/Tspublic/strik/a67b.html

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

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: CUP.

Educational Testing Service Network. (1999). *What is ETS?* Retrieved July 30, 2001 from the World Wide Web: http://www.ets.org/aboutets/visitors.html

Ellis, R. (1997). *Second Language Acquisition.* Oxford: OUP.

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

Foster, P., Tonkyn, A. & Wigglesworth, G. (2000). Measuring spoken language: A unit for all measures. *Applied Linguistics, 21* (3), 354-375.

Fulcher, G. (1995). Variable competence in second language acquisition: A problem for research methodology? *System, 23* (1), 25-33.

Fulcher, G. (1996a). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13* (2), 208-238.

Fulcher, G. (1996b). Testing tasks: issues in task design and the group oral. *Language Testing, 13* (1), 23-52.

Fulcher, G. (1997). The testing of speaking in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education.* (p. 75-85). Netherlands: Kluwer Academic Publishers.

Fulcher, G. (1999). [Review of the book *A history of foreign language testing in the United States: From its beginnings to the present*]. *Language Testing, 16* (3), 389-393.

Gatbonton, E. (1978). Patterned phonetic variability in second language speech: a gradual diffusion model. *Canadian Modern Language Review, 34*, 335-347.

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

Hatch, E. (1992). *Discourse and language education.* Cambridge: Cambridge University Press.

Hatch, E. & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics* New York: Newbury House.

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

Jennings, M., Fox, J., Graves, B. & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16* (4), 426-456.

Jones, R. & Spolsky, B. (Eds.) (1975). *Testing language proficiency.* Arlington, VA: Centre for Applied Linguistics.

Kenyon, D. (1998). An investigation of the validity of task demands on performance-based tests of oral proficiency. In A.J. Kunnan (Ed.), *Validation in Language Assessment: Selected Papers from the 17th LTRC, Long Beach*, (p.19-40). Mahwah, NJ: Lawrence Erlbaum.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16* (2), 163-188.

Lantolf, J.P. & Ahmed, M.K. (1989). Psycholinguistic perspectives on interlanguage variation: A Vygotskyan analysis. In Gass, S., Madden, S., Preston, D. & Selinkjer, L. (Eds.). *Variation in second language acquisition, Volume II: Psycholinguistic issues.* (p.93-108). Avon, England: Multilingual Matters.

Lantolf, J.P. & Frawley, W. (1985). Oral-proficiency testing: a critical analysis. *The Modern Language Journal, 69* (iv), 337-345.

Lantolf, J.P. & Frawley, W. (1988) Proficiency: Understanding the construct. *Studies in Second Language Acquisition, 10,* 181-195.

Lantolf, J. & Frawley, W. (1992). Rejecting the OPI – again: A response to Hagen *ADFL Bulletin 23*, (2), 34-37.

Lazaration, A. (1992). The structural organization of a language interview: A conversational analytic perspective. *System, 20*, 373-386.

Lazaraton, A. (1996). A qualitative approach to monitoring examiner conduct in the Cambridge assessment of spoken English (CASE). In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment; Selected papers from the 15th Language testing research colloquium, Cambridge and Arnhem.* (p.18-54). Cambridge: Cambridge University Press.

Lewkowicz, J. (2000). Authenticity in language testing: Some outstanding questions. *Language testing, 17* (1), 43-64.

Lowe, P. (1983). The ILR oral interview: origins, applications, pitfalls, and implications. *Die Unterrichtspraxis, 16*, 230-244.

Madsen, H.S. & Jones, R. L. (1981). Classification of oral proficiency tests. In A. Palmer, P. Groot & G. Trosper (Eds.) *The construct validation of tests of communicative competence.* (p. 15-30). Washinton, D.C.: TESOL.

Matthews, M. (1990). The measurement of productive skills: doubts concerning the assessment criteria of certain public examinations. *ELT Journal, 44* (2), 117-121.

McNamara, T.F. (1995a). Modelling performance: Opening Pandora's box. *Applied Linguistics, 16* (2), 159-179.

McNamara, T. (Speaker). (1995b). *Measuring performance in second language oral tests.* (Videocassette Recording). Ottawa, Can: Language Training Canada and the Personnel Psychology Centre, Government of Canada.

McNamara, T. F. (1996). *Measuring second language performance.* Essex, England: Addison Wesley Longman Limited.

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

McNamara, T. (2000). *Oxford introductions to language study; Language testing.* Oxford: Oxford University Press.

McNamara, T.F., & Adams, R.J. (1991). Exploring rater characteristics with Rasch techniques. (ERIC Document Reproduction Service No. 345 498)

McNamara, T.F. & Lumley, T. (1995). Rater characteristics and rater bias: implications for training. *Language Testing, 14* (1), 54-71.

NcNamara, T.F. & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14* (2), 140-156.

Messick, S. (1989). Validity. In R. L. Linn (Ed.). *Educational measurement* (p.13-103). New York, NY: American Council on Education and Macmillan Publishing Company.

Norris, J., Brown, J.D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu: University of Hawai'i Press.

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge: Cambridge University Press.

Peckham, R. (undated). *The interagency language roundtable scale.* Retrieved November 9, 2000 from the World Wide Web: http://fmc.utm.edu/~rpeckham/ilrhome.html

Peterson, N., Kolen, M. & Hoover, H. (1989). Scaling, norming, and equating. In Linn (Ed.), *Educational measurement* (p. 221-262). New York, NY: American Council on Education and Macmillan Publishing Company.

Public Service Commission of Canada/Commission de la fonction publique du Canada. *Second language evaluation: Oral interaction test.* Retrieved October 19, 2000 from the World Wide Web: http://jobs.gc.ca/ppc/en_sle_pg__04_a.htm

Richards, J.C., Platt, J. & Platt, H. (1992). *Longman dictionary of language teaching & applied linguistics.* Essex, England: Longman Group UK Limited.

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

Salaberry, R. (2000). Revising the revised format of the ACTFL Oral Proficiency Interview. *Language Testing, 17*, (3), 289-310.

Serson, S. (2001). Opening remarks for delivery by Scott Serson, President, Public Service Commission to the Public Accounts Committee. Retrieved May 9, 2001 from the World Wide Web: http://www.psc-cfp.gc.ca/speech/ss030501_e.htm

Shohamy, E. (1988). A proposed framework for testing the oral language of second/foreign language learners. *SSLA, 10,* 165-179.

Shohamy, E.(1990). Language testing priorities: a different perspective. *Foreign Language Annals, 23* (5), 385-392.

Spolsky, B. (1995). *Measured words.* Oxford: OUP.

Stansfield, C.W. & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*, (3), 347-364.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.). *Input in second language acquisition.* (p.235-53). Rowley, MA: Newbury House.

Tarone, E. (1998). Research on interlanguage variation: Implications for language testing. In L. Bachman & A. Cohen (Eds.). *Interfaces between second language acquisition and language testing research.* (p.71-89) Cambridge: Cambridge University Press.

Tomiyama, M. (2000). Child second language attrition: A longitudinal case study. *Applied Linguistics, 21*(3), 304-332.

Turner, C. & Upshur, J. (1995). Some effects of task types on the relation between communicative effectiveness and grammatical accuracy in intensive esl classes. *TESL Canada Journal, 12* (2), 18-31.

Upshur, J. & Turner, C.E. (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.

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

Wigglesworth, G. (1997a). Task variation in oral interaction tests: Increasing the reality. *Prospect, 12* (1), 35-49.

Wigglesworth, G. (1997b). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14* (1), 85-106.

Wright, B.D. & Masters, G.N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Young, R. (1989). Ends and means: Methods for the study of interlanguage variation. In S. Gass, C. Madden, D. Preston & L. Selinker (Eds.). *Variation in second language acquisition. Psycholinguistic issues 2.* (65-90) Philadelphia: Multilingual Matters.

Young, R. (1991). *Variation in interlanguage morphology.* New York, NY: Peter Lang Publishing Inc.

Young, R. (1992). Expert-novice differences in oral foreign language proficiency. (ERIC Document Reproduction Service No. FL 020- 514)

Young, R. (1995a). Conversational styles in language proficiency interviews. *Language Learning 45*, (1), 3-42.

Young, R. (1995b). Discontinuous interlanguage development and its implications for oral proficiency rating scales. *Applied Language Learning 6*, (1,2), 13-26.

Appendices

Appendix		Page
A.	Informed Consent to Participate in Research	153
B.	Consentement à participer à la recherche	155
C.	Certificate of Ethical Acceptability	157
D.	Criteria for Determining Task Difficulty Document	159
E.	Question Prompt Category Complexity Questionnaire	161
F.	The Response Idea Unit Transcription Coding Protocol	167
G.	Conventions of the Simplified Analysis of Speech Unit	169
H.	Workshop 1 and 2 results: Question prompts and headings	173
I.	Analysis of responses to Question Prompt Category Complexity Questionnaire	177
J.	SLE:OI Test Information	183

Appendix A

Informed Consent to Participate in Research

INFORMED CONSENT TO PARTICIPATE IN RESEARCH

This is to state that I agree to participate in the research project entitled "An Investigation into the Second Language Evaluation: Oral Interaction", and conducted by Christian Colby-Kelly, with Dr. Carolyn Turner, supervisor, McGill University Department of Second Language Education.

Purpose and Procedures: This research will look at excerpts of some of the oral language produced in the Canadian Government's Second Language Evaluation: Oral Interaction (SLE:OI) test. Short samples of oral speech from the SLE:OI will be transcribed and later analyzed. They will be used solely for research purposes. All participants, (testers and test candidates), are asked to give their written consent. The names of all participants will not be published; instead participants will be referred to by a confidential code whereby they will be identified by a designated number.

Conditions of Participation: The only request of test candidates is that they give their written consent of participation in the project. Participants may appreciate that in choosing to give their consent, they are supporting their own language testing milieu; this research is designed to contribute to providing Government test users with quality testing services.

All involved SLE:OI testers are asked to consent to participate in the project. Some testers will also be asked to give expert judgements relating to data classification in a workshop session. Participation of this kind is expected to be professionally enriching since it is expected to enhance testers' awareness of certain aspects of the SLE:OI test.

Participants may withdraw from the project at any time without penalty or prejudice. They will be contacted by phone by Christian Colby-Kelly to ensure that all the conditions of this agreement are well understood prior to signing the Consent Form.

- I understand the purpose of this study.
- I understand how confidentiality will be maintained.
- I understand that I am free to withdraw at anytime from the study without any penalty or prejudice.

I have carefully studied the above and understand my participation in this agreement. I freely consent and voluntarily agree to participate in this study.

Name (please print) _____

Signature _____ Date _____

Appendix B

Consentement à participer à la recherche

CONSENTEMENT A PARTICIPER A LA RECHERCHE

Par la présente j'atteste que j'accepte de participer au projet de recherche intitulé "Etude sur l'évaluation de langue seconde: Test d'interaction orale", mené par Christian Colby-Kelly, avec le Dr. Carolyn Turner, surveillante de projet, Département de l'Enseignement en langue seconde de l'Université McGill.

But et procédures: Cette recherche portera sur des échantillons d'entre-vue produites dans le cadre d'evaluation de langue seconde: Test d'interaction orale (ELS: IO), du gouvernement canadien. De courts extraits du ELS:OI seront transcrits et analysés. Ils serviront uniquement à des fins de recherche. Tous les participants, (les candidats(es) à l'examen et les évaluteurs(trices) de l'examen) sont priés d'accorder leur consentement par écrit. Les noms des participants ne seront pas publiés. Les participants seront identifiés par un code numérique confidentiel.

Conditions de participation: Tous les candidats(es) du test doivent signer le document intitulé "Consentement à participer à la recherche". Ces derniers comprendront qu'en acceptant d'y participer ils contribuent à améliorer les conditions d'administration des examens de langue. Cette recherche vise à contribuer à fournir des services de qualité aux usagers des examens du gouvernement.

Tous les évaluateurs(trices) de l'ELS:IO impliqués dans le projet sont également priés(es) de signer le consentement à participer. Certains(es) entre eux participeront à un atelier sur la classification des données. Nous croyons que cette participation sera une expérience enrichissante sur le plan professionnel et contribuera à approfondir les connaissances de certains aspects de l'examen d'ELS:IO.

Les participants peuvent se retirer du projet à n'importe quel moment sans pénalité ou préjudice. Christian Colby-Kelly communiquera avec eux par téléphone pour s'assurer que toutes les conditions de cette entente sont bien comprises avant la signature du formulaire de consentement.

- Je comprends le but de cette étude.
- Je comprends de quelle façon sera assuré la confidentialité lors de project de recherche.
- Je comprends que je suis libre de me retirer à n'importe quel moment sans pénalité ou préjudice.

J'ai soigneusement étudié le texte ci-dessus et je comprends ma participation dans cette entente. Je consens librement et j'accepte volontier de participer à cette étude.

Nom (en lettres moulées svp) _____

Signature _____ Date _____

Appendix C

Certificate of Ethical Acceptability

Appendix D

Criteria for Determining Task Difficulty Document

Criteria for Determining Task Difficulty

Adapted from Brindley (1987), Brown and Yule (1983), and Anderson and Lynch (1985), in Nunan (1989)

FACTORS TO BE TAKEN INTO CONSIDERATION IN DETERMINING TASK DIFFICULTY

Easier ⟶	**More difficult**
Task	
low cognitive complexity	cognitively complex
simple syntax	complex syntax
specific vocabulary	generalized vocabulary
has few steps	has many steps
familiar topic	unfamiliar topic
familiar context	unfamiliar context
much context provided	no context provided
interesting/involving	boring/non-involving
does not require grammatical accuracy	requires grammatical accuracy
does not require cultural knowledge	requires cultural knowledge
narratives/instructions	opinion/explanation
Text	
is short, not dense (few facts)	is long and dense (many facts)
clear presentation	presentation not clear
information is explicit	info. requires inferences
repetition of message occurs	no repetition of message
synonyms used	no synonyms used
familiar content	unfamiliar content
many contextual clues	few contextual clues

Appendix E

Question Prompt Category Complexity Questionnaire

Question Prompt Category Complexity Questionnaire

Introduction: The following are the categories of question prompts which you as a group have identified at the last workshop. You have also indicated in discussions at that workshop that the key factor differentiating them is that of complexity.

Instructions: In order to re-familiarize yourself with the work of the last workshop, please read over all the question prompt categories. Then indicate the level of complexity you would assign each one by circling the appropriate number using the following scale:

1 easy
2 somewhat easy
3 fairly difficult
4 difficult

A) Topic Specification Question Prompts:

Job specific	1	2	3	4
Problem-specific (e.g. a case study)	1	2	3	4

B) Question Prompts with an Expected Elicited Response which is Functional:

Solution-seeking questions (could lead to *explanation*)	1	2	3	4
Leading to *explanation*	1	2	3	4
Description (free rein in response); questions leading to more of a *description*	1	2	3	4

C) Question Prompts Grouped by Length or Amount of Detail in the Expected Response:

Short	1	2	3	4
Long	1	2	3	4
Saying yes, little opinion required; agreeing	1	2	3	4
Saying no, little opinion required; disagreeing	1	2	3	4
'Surface' questions*	1	2	3	4
'Deep' questions**	1	2	3	4

D) Question Prompts which use Formulaic Questions:

"[What] do you think..." questions	1	2	3	4
"To what extent..." quantitative questions	1	2	3	4
"How *adjective* is..." evaluative adjective, range questions (Using degree-intensifying adjectives)	1	2	3	4
"Would you say..." questions	1	2	3	4
"How do you..." questions (e.g. strike the balance between..., etc.; 'recipe' questions which seek a solution)	1	2	3	4

* 'Surface' questions = Don't lock the candidate into a deeper response.

** 'Deep' questions = Lock the candidate into a deeper response; compare/contrast/say why... kinds of questions.

E) Question Prompts with an Expected Elicited Response of a Particular Type:

1) Comparing

- Asking for qualities ("What sort of person...")

N.B. "This could be comparative." 1 2 3 4

2) Relating

- Relational

(how one fact relates to another fact) 1 2 3 4

3) Speculating

- Speculative questions about outcomes 1 2 3 4
- Suggesting a point a point of view;

or speculating on one, to get a reaction 1 2 3 4

- Presenting different points of view 1 2 3 4
- Devil's advocate questions,

seeking a response 1 2 3 4

- Choice/options 1 2 3 4

4) Other

- Elaborating or wrapping-up,

elicited from a statement 1 2 3 4

- Listing questions;

where the response may include a list 1 2 3 4

• Justifying points of view by generalizing	1	2	3	4
• Picking one out of a series,				
(e.g. most important quality, etc.)	1	2	3	4

F) Grouped by Vocabulary Used in Question Prompt:

Statement, no direct opinion word used	1	2	3	4
Repetition of key vocabulary element in question	1	2	3	4

G) Grouped by Syntax Used in Question Prompt:

Confused/possibly multiple questions				
/meandering questions	1	2	3	4

N.B. One category, 'Leading to *opinion*', was omitted due to its being too vague for the purposes of determining complexity.

Appendix F

The Response Idea Unit Transcription Coding Protocol

The Response Idea Unit Transcription Coding Protocol

Response idea units (RIUs) are bordered by topic shift boundaries, defined in the following way:

> A segment of information which is a single semantic unit, bounded by pauses and/or intonation changes, and in which the speaker speaks cohesively with the purpose of relating the message to psychological reality for the encoder.
>
> (Adapted from Crookes and Rulon, 1985; and Kroll, 1977, as cited in Crookes, 1990, p. 187, 184.)

Furthermore, RIUs were transcribed using the following transcription coding:

<um>, <uh>	- filled pauses
<uh huh>	- encouragers [Note these occurred exclusively in rater speech, and never in that of test candidates]
<x sec>	- unfilled pauses of 1 second or more
<?>	- inaudible or incomprehensible sounds
<...>	- false starts, voice trails off or is interrupted
[ ]	- indicated words left out, such as repetitions omitted to create greater uniformity in responses
{ }	- enclosed deictic clauses or one or more word responses which paralleled those of questioner
.	- end of a clause or sentence unit
,	- end of a clause within a sentence unit
?	- end of question forms, and sentence unit ending in which intonation rises
Gee, OK	- one-word exclamations in English
Ouf, bof, bien	- one-word exclamations in French

Contractions were counted as separate words. Sentences or clauses beginning with 'and' and 'because' were accepted since this is a common French Canadian syntactic structure, and was found to produce meaningful, if interfered, speech.

Appendix G

Conventions of the Simplified Analysis of Speech Unit

Conventions of the Simplified Analysis of Speech Unit

Foster, Tonkyn and Wigglesworth (2000) have defined their unit of speech, the Analysis of Speech unit (AS-unit), in the following terms:

> An AS-unit is a sinle-speaker's utterance consisting of *an independent clause, or sub-clausal unit*, together with any *subordinate clause(s)* associated with either [italics in original].
>
> (p. 365)

The focus of the present study has dealt with oral test response fluency, accuracy and complexity. An extensive inquiry into these discourse characteristics was beyond the scope of the present research. Therefore I decided to adapt the Foster, Tonkyn and Wigglesworth (2000) AS-unit, simplifying it to meet the more modest needs of the current study. The new Simplified Analysis of Speech (SAS) unit simply parses the response data into *independent* and *subordinate clauses*, excluding the *sub-clausal unit*.

The following Foster, Tonkyn and Wigglesworth (2000) definition of these clauses was employed in the AS-unit and the SAS unit:

> *An independent clause* will be minimally a clause including a finite verb.
>
> *A subordinate clause* will consist minimally of a finite or non-finite Verb element plus at least one other clause element (Subject, Object, Complement or Adverbial)
>
> (p. 366)

Moreover, the SAS unit relies on the coding conventions of the AS-unit, level 3. (Level 3 was designed by Foster, Tonkyn and Wigglesworth [2000] expressly for standardization of discourse such as that of OPI test candidates, in which units of a certain completeness would be required for comparison purposes; such is the case in the present research.) Level 3 is illustrated below:

> Excluded are:
>
> - One-word minor utterances
> - Echo responses which are verbatim
> - Verbless elliptical AS-units and SAS units involving ellipsis of elements of the interlocutor's speech
> - AS-units and SAS units involving substitution of clause, predicate, or predication level units of interlocutor's speech
> - One or two-word greetings and closures
>
> (p. 370-371)

In addition, the following conventions were followed:

- False starts were defined as "where the speaker repeats previously produced speech" (p. 368).
- False starts were excluded from the data
- French words were not counted in the TTR
- Contractions were counted as separate words
- Silent and filled pauses occurring within the confines of false starts were omitted *unless* they abutted the boundaries of included discourse, in which case they were considered to be a part of the SAS unit
- Self-repair was defined as self-correction, "when the speaker identifies an error either during or immediately following production and stops and reformulates the speech" (p. 368).

The RIU data was divided into SAS units using the following coding:

| |	- upright slashes indicate SAS unit boundaries
: :	- double colons indicate a clause boundary within the SAS unit
{ }	- curled brackets surround false starts, functionless repetitions, and self-repairs
—	- excluded data was struck out, rather than deleted from the final analysis (see example above)

Appendix H

Workshop 1 and 2 results: Question prompts and headings

Workshop 1 and 2 results: Question prompts and headings

Workshop I:

Assessor I –

Problem-Specific (Case Study):
Relational (How One Fact Relates to Another Fact):
Choice/Options:
Description (Free Rein in Response):
Agree or Disagree:

Assessor II –

"Surface" Questions:
"Deep" Questions, (Compare, Contrast, Say Why):
Confused/Possibly Multiple Questions/Meandering Ones:
Repetition of Key Vocabulary Element in the Question:
[Short or] Long:

Assessor III –

Solution-Seeking Questions (Could Lead to Explanation), "How do you...":
Presenting Different Points of View, "Some people...other people," either/or:
Justifying Points of View by Generalizing:
Evaluative Adjective Questions, "How *Adjective* is...":
Pick One out of a Series (Most Important Quality, Etc.):

Assessor IV –

[What] Do You Think...Questions:
[To What] Extent...Questions:Assessor IV:
No Direct Opinion Word Used (Statements):
Would You Say...Questions:
Yes/No Questions:
Asking for Qualities ("What Sort of a Person...", This Could be Comparative):
Range Questions ("How ...is this?") *Evaluative*:
Devil's Advocate Questions, Seeking a Response:
Recipe Questions, ("How do You e.g. Strike the Balance Between...") for a Solution, Sol. Ques.:

Assessor V –

Yes/No, Little Opinion Required:
Speculative Questions about Outcomes:
Listing Questions (Where the Response may List):
Quantitative Questions ("To What Extent..."):
Questions Leading More to a Description:
Quantitative Questions ("To What Extent..."), + Questions Leading More to a Description:
Statement Questions to Elicit Elaboration or Wrap-up:

Assessor VII –

Job-Specific Questions:
Questions Leading to Opinion, + Evaluative Adjectives:
Questions Leading to Explanation:
Questions Suggesting a Point of View or Speculating on One in Order to Get a Reaction:

* 'Surface' questions = Don't lock the candidate into a deeper response.
** 'Deep' questions = Lock the candidate into a deeper response; compare/contrast/say why... kinds of questions.

Workshop 2:

A) Topic Specification Question Prompts:

1. Job specific
2. Problem-specific (e.g. a case study)

B) Question Prompts with an Expected Elicited Response which is Functional:

3. Solution-seeking questions (could lead to *explanation*)
4. Leading to explanation
5. *Description* (free rein in response); questions leading to more of a *description*

C) Question Prompts Grouped by Length or Amount of Detail in the Expected Response:

6. Saying yes, little opinion required; agreeing
7. Saying no, little opinion required; disagreeing
8. 'Surface' questions*
9. 'Deep' questions**

D) Question Prompts which use Formulaic Questions:

10. "[What] do you think..." questions
11. "To what extent..." quantitative questions
12. "How *adjective* is..." evaluative adjective, range questions (Using degree-intensifying adjectives)
13. "Would you say..." questions
14. "How do you..." questions (e.g. strike the balance between..., etc.; 'recipe' questions which seek a solution)

E) Question Prompts with an Expected Elicited Response of a Particular Type:

E1) Relating

15. Relational (how one fact relates to another fact)

E2) Speculating

16. Speculative questions about outcomes
17. Suggesting a point a point of view; or speculating on one, to get a reaction
18. Presenting different points of view
19. Devil's advocate questions, seeking a response
20. Choice/options

E3) Other

21. Listing questions; (where the response may include a list)
22. Justifying points of view by generalizing
23. Picking one out of a series, (e.g. most important quality, etc.)

F) Grouped by Vocabulary Used in Question Prompt:

24. Statement, no direct opinion word used
25. Repetition of key vocabulary element in question

G) Grouped by Syntax Used in Question Prompt:

26. Confused/possibly multiple questions /meandering questions

Appendix I

Analysis of responses to Question Prompt Category Complexity Questionnaire

Analysis of responses to Question Prompt Category Complexity Questionnaire

The level of complexity was assigned by circling the appropriate number, using the following scale:

5	easy
6	somewhat easy
7	fairly difficult
8	difficult

A) Topic Specification Question Prompts:

1. Job specific	50%	easy
	50%	somewhat easy
	0	fairly difficult
	0	difficult
2. Problem-specific (e.g. a case study)	20%	easy
	60%	somewhat easy
	20%	fairly difficult
	0	difficult

B) Question Prompts with an Expected Elicited Response which is Functional:

3. Solution-seeking questions (could lead to *explanation*)	0	easy
	100%	somewhat easy
	0	fairly difficult
	0	difficult
4. Leading to *explanation*	50%	easy
	50%	somewhat easy
	0	fairly difficult
	0	difficult
5. *Description* (free rein in response); questions leading to more of a *description*	33%	easy
	33%	somewhat easy
	33%	fairly difficult
	0	difficult

C) Question Prompts Grouped by Length or Amount of Detail in the Expected Response:

6. Saying yes, little opinion required; agreeing	83%	easy
	17%	somewhat easy
	0	fairly difficult
	0	difficult
7. Saying no, little opinion required; disagreeing	67%	easy
	17%	somewhat easy
	17%	fairly difficult
	0	difficult
8. 'Surface' questions*	40%	easy
	60%	somewhat easy
	0	fairly difficult
	0	difficult
9. 'Deep' questions**	0	easy
	0	somewhat easy
	17%	fairly difficult
	83%	difficult

D) Question Prompts which use Formulaic Questions:

10. "[What] do you think..." questions	0	easy
	60%	somewhat easy
	40%	fairly difficult
	0	difficult
11. "To what extent..." quantitative questions	0	easy
	33%	somewhat easy
	0	fairly difficult
	67%	difficult

* 'Surface' questions = Don't lock the candidate into a deeper response.

** 'Deep' questions = Lock the candidate into a deeper response; compare/contrast/say why... kinds of questions.

12. "How *adjective* is..." evaluative adjective, range questions (Using degree-intensifying adjectives)	17%	easy
	0	somewhat easy
	17%	fairly difficult
	67%	difficult
13. "Would you say..." questions	17%	easy
	50%	somewhat easy
	33%	fairly difficult
	0	difficult
14. "How do you..." questions (e.g. strike the balance between..., etc.; 'recipe' questions which seek a solution)	0	easy
	50%	somewhat easy
	17%	fairly difficult
	33%	difficult

E) Question Prompts with an Expected Elicited Response of a Particular Type:

E1) Relating

26. Relational (how one fact relates to another fact)	0	easy
	0	somewhat easy
	67%	fairly difficult
	33%	difficult

E2) Speculating

16. Speculative questions about outcomes	0	easy
	0	somewhat easy
	83%	fairly difficult
	17%	difficult
17. Suggesting a point a point of view; or speculating on one, to get a reaction	0	easy
	17%	somewhat easy
	50%	fairly difficult
	33%	difficult
18. Presenting different points of view	0	easy
	50%	somewhat easy
	33%	fairly difficult
	17%	difficult

19. Devil's advocate questions, seeking a response	0	easy
	33%	somewhat easy
	17%	fairly difficult
	50%	difficult
20. Choice/options	0	easy
	80%	somewhat easy
	20%	fairly difficult
	0	difficult

E3) Other

21. Listing questions; (where the response may include a list)	0	easy
	100%	somewhat easy
	0	fairly difficult
	0	difficult
22. Justifying points of view by generalizing	0	easy
	17%	somewhat easy
	67%	fairly difficult
	17%	difficult
23. Picking one out of a series, (e.g. most important quality, etc.)	50%	easy
	33%	somewhat easy
	17%	fairly difficult
	0	difficult

F) Grouped by Vocabulary Used in Question Prompt:

24. Statement, no direct opinion word used	0	easy
	17%	somewhat easy
	83%	fairly difficult
	0	difficult
25. Repetition of key vocabulary element in question	33%	easy
	67%	somewhat easy
	0	fairly difficult
	0	difficult

G) Grouped by Syntax Used in Question Prompt:

26. Confused/possibly multiple questions /meandering questions	0	easy
	0	somewhat easy
	40%	fairly difficult
	60%	difficult

Appendix J

SLE:OI Test Information

Second Language Evaluation: Oral Interaction Test

Description | Evaluation
Test Results | Candidate Feedback Sheet
Testing Information | Tips
Examples of Language Tasks at Levels A, B, and C | Examples of the Characteristics of Performance at Levels A, B, and C

Description

The Oral Interaction Test assesses your ability to speak and understand French as your second official language. (There is also an Oral Interaction Test available to assess second language oral proficiency in English). The evaluation takes the form of a conversation with an assessor about work-related matters and lasts approximately 30 minutes. To provide a record of the test, the conversation is recorded on an audio cassette.

The Oral Interaction Test assesses your overall ability to communicate in French in a work context, based on the language tasks you can perform and the accuracy with which you communicate your message. Language tasks involve performing communicative activities such as asking questions, relating events, giving explanations, and expressing and supporting opinions. Accuracy refers to the degree to which fluency, grammar, vocabulary, and pronunciation affect your communication. At each level of proficiency, the assessor will evaluate your ability to perform specific language tasks by asking you questions or engaging you in a dialogue similar to a conversation you might have in the course of your work. Examples of the language tasks required at each proficiency level may be found after the section on Tips.

While the specific topics discussed during the conversation will vary, the subject matter remains work-related. At the A level, topics are relatively simple and concern such things as hours of work, office procedures, and routine work tasks. At the B and C levels, you will be asked to deal with more complex topics such as projects you have worked on, problems encountered, or your views on an issue affecting your work. It is important to note, however, that you will be evaluated on how well you communicate in French, and not on the factual content or ideas expressed. Since the test is protected, you may express your own views, which need not necessarily reflect those of your organization.

Evaluation

The language tasks and the degree of accuracy required become more demanding as one progresses from Level A to Level C. When assigning a level to your performance, the assessor will evaluate the overall degree of clarity, ease and precision with which you communicate your ideas. Your final result is a **global evaluation** of your ability to perform language tasks in a variety of work-related contexts with the appropriate accuracy. Based on your test performance, you will obtain Level A, B, or C, or receive an exemption in Oral Interaction. If you do not meet the minimum requirements for Level A, this will be indicated by an X on your result sheet. A description of some of the characteristics of performance at Levels A, B and C may be found at the end of this document. Exemption from further testing is granted to C-level candidates who obtain a high enough rating that they need not be tested again.

Test Results

The result you have obtained will be sent to you soon after you have taken the test. If you have questions about the level you have been assigned, the Candidate Feedback sheet is designed to clarify them for you. As described in the section entitled Candidate Feedback Sheet, the feedback sheet gives you an indication of the areas you need to improve to enable you to effectively communicate in French at the next level. If you do not receive a Candidate Feedback sheet, please contact the responsible officer in the organization that requested your test.

Candidate Feedback Sheet

Many requests are received from candidates who wish to know more about how to improve their oral interaction skills in French. On the feedback sheet you will find a series of comments for each of the levels of oral interaction that may be assigned following the test. Feedback about your performance on the test will be indicated by the checked boxes on this sheet which will be sent to you once you have taken the test. Your feedback give you an indication of the areas you will need to work on to enable you to effectively communicate at the next level. To give you an idea of what is expected at each of the three levels, please consult the chart, which provides an outline of the more common characteristics of performance for all the levels.

We would ask you to keep in mind that the Oral Interaction Test was not designed to give feedback in terms of specifics errors. Because the rating is global, the feedback you receive suggests only the general areas on which you should focus to improve your oral interaction skills. If you would like to know more specifically how to improve your linguistic performance, you may wish to consult a language teaching organization. They are equipped to further diagnose your language abilities and determine the method and materials suitable for improving your areas of weakness.

Testing Information

- Bring one piece of identification with your signature on it, and your Personal Record Identifier (PRI) if you are a government employee.

- If you need special test arrangements because of a disability, please notify the responsible officer in the organization that requested your test so that appropriate arrangements can be made.

- During the OI interview the assessor will ask you questions about your work, your work experiences, or perhaps your professional training if you are just entering the workforce. In addition, you may also be asked to discuss other topics that are related to the work environment.

- Please inform the assessor if any of the questions asked during the test are sensitive for personal or security reasons, or if you do not know enough about a particular topic to be able to talk about it. This will have no effect on your test result. However, you should know that all the information on the recording of the test is protected.

- The OI Test is designed to assess how well you communicate in your second language. During the test the assessor is not assessing how well you know your job or a specific topic, but rather how well you can communicate what you do know. It is the communication that is important and not the factual content of the topics discussed.

- Assessors recognize that some candidates may feel nervous during a test. Your assessor will try to help you feel at ease during your OI Test.

- If you should feel indisposed before or during the test, tell the assessor or the officer in charge. Otherwise you must accept the test result and the retest restrictions.

- Sometime after the test you will receive your result from the organization that requested your test. Unless you have been exempted from further testing, along with the test result you will also receive a feedback sheet, indicating what areas you would need to improve if you wish to reach the next level of oral proficiency. If you have just met the requirements of a level, a box will be checked at the end of the feedback information, indicating that you will need to practice your oral language to maintain the assigned level.

- If you do not receive your feedback sheet or if you have any questions about the test, you should get in touch with the contact person in the organization that requested your test.

Tips

Prepare for the test.

- Try to speak and listen to French as much as possible before taking the test. You can do this by listening to the radio, watching television, and speaking French with your colleagues and friends.

Arrive on time and speak French from the beginning.

- Arriving on time and speaking French as soon as you meet the assessor will help you adjust more quickly to the testing session.

Do not be overly worried about making mistakes.

- If you can't think of a certain word, use a simple substitute to explain the meaning and continue with the conversation. If you are aware that you are making mistakes and would feel better if you corrected them, go ahead and do so. However, remember that frequent corrections may disrupt the flow of the conversation.

Tell the assessor if the topic is sensitive.

- If any of the questions posed by the assessor concern a topic that is sensitive for personal or security reasons, inform the assessor and he or she will move on to another topic.

Don't be discouraged if parts of the test seem difficult.

- At various times the assessor will use more complex questions to give you the opportunity to perform at your maximum level of proficiency. However, testing at this higher level of proficiency will not take the entire testing session.

Pay no attention to the cassette recorder.

- It is used to provide a record of your test. Concentrate on talking to the assessor instead.

Answer questions as fully as possible.

- In order to give the assessor a sufficient sample to evaluate, expand on your answers by giving details, explaining points or developing your thoughts.

Examples of Language Tasks at Levels A, B, and C

Level A

Ask and answer simple questions.

- A machinist asks a colleague where and how a certain tool may be obtained, or a personnel officer answers an employee's questions about the time allowed for a particular test.

Give simple directions or instructions.

- A receptionist directs a visitor to the cafeteria, or a manager gives a new clerk instructions on how to handle the travel arrangements for an upcoming trip.

Handle simple work-related situations.

- A secretary tells a visitor that the director is out of town and therefore unavailable for a meeting.

Level B

Give simple explanations.

- An administrative officer explains to a manager the standard procedures for hiring a term employee.

Give factual descriptions (of people, places or things).

- An officer describes to a manager the design, colour and dimensions of the information brochures that have been ordered.

Narrate events (past, present, future).

- A security officer relates to the supervisor the events of a break-in.

Handle work-related situations with a complication.

- A clerk resolves the problem of an incomplete supply order with the person responsible for filling out the order.

Level C

Give detailed explanations and descriptions

- A secretary explains to another secretary a complex system of keeping track of ministerial correspondence.

Handle hypothetical questions

- A unit head explains to a superior what would happen to the work output if a compressed work week were adopted by the unit.

Support an opinion, defend a point of view, or justify an action

- A supervisor defends the opinion that flexible hours for the unit should be permanently adopted.

Counsel and give advice.

- A librarian helps a colleague make a decision about an employment option.

Handle complex work-related situations

- The head of a unit discusses with a junior employee the problem of that employee's frequent absences and tardiness, and the effect this has on the rest of the work unit.

Examples of the Characteristics of Performance at Levels A, B, and C

Level	A	B	C
Ability to converse	can sustain a simple question and answer exchange can produce new sentences (not simply repeat memorized material)	can sustain an informal conversation on concrete topics is able to paraphrase when lacking the exact vocabulary	can participate effectively in discussions on a broad variety of topics can expand on topics with ease
Ease in using the language	delivery may be slow can form sentences with some hesitations	speaks with some spontaneity may hesitate when using more complex sentences	has a natural delivery seldom hesitates except to look for ideas
Clarity of communication	has basic vocabulary for routine work-related topics can talk about facts in the present	has concrete vocabulary for less routine work-related topics can situate facts and events in time (i.e., has good mastery of simple	has precise vocabulary to convey exact meaning can link sequences of facts and events in time (i.e., has solid

	can link words to form simple sentences may ask for repetition or rephrasing of some questions can generally be understood if the listener pays close attention	verb tenses) can link sentences together into longer passages has few difficulties understanding the assessor can be understood by most people but repetition may sometimes be required	complex verb forms) can link sentences effectively to convey complex ideas can readily and accurately interpret what the assessor says can be easily understood; pronunciation does not interfere with communication

X	Your performance does not meet the minimum requirements for Level A.
Exemption	You have been exempted from further testing in oral interaction because your performance contains no major weaknesses. Since you can handle most situations in French with excellent control of the language and a high degree of ease, it can be expected that you will maintain level C.