Are We All On the Same Page?

An Exploratory Study of OPI Ratings across NATO Countries Using the NATO STANAG 6001 Scale

By

Julie J. Dubeau, B.A. (Hons.)

A thesis submitted to the Faculty of Graduate Studies and Research
In partial fulfillment of the requirements for the degree of
Master of Arts

School of Linguistics and Applied Language Studies
Carleton University, Ottawa, Ontario, Canada
September 2006

Your file Votre référence
ISBN: 978-0-494-18258-1
Our file Notre référence
ISBN: 978-0-494-18258-1

Abstract

Tests that provide military and civilian personnel with Standardised Language Profiles in English carry high stakes in NATO. Since no known study has investigated the inter-reliability of ratings among member countries, this exploratory research informs on the comparability of ratings assigned to oral proficiency interviews (OPI). One hundred and three participants from eighteen countries and two NATO units rated two English language OPIs against the NATO STANAG 6001 scale, the common metric. Results indicated that there were some differences in the ratings assigned from country to country, and differences in ratings within each country. Ratings were brought closer to the mean by introducing 'plus levels' into the scheme. The results also showed that experience alone is an insufficient condition for rating correctly and consistently, and the need for more scale training was evident, if score reliability is to be achieved. This study's findings have implications for testers, raters, testing managers and stakeholders within NATO.

Acknowledgements

Returning to university after a long hiatus poses many challenges to any student. With a busy family life and career, the addition of night courses, essays and a thesis, required not only organisation and discipline, but also a lot of help from other people! Thankfully, I was not alone.

First, the M.A. in Applied Language Studies at the Carleton University School of Linguistics and Language Studies (SLALS) proved to be a flexible programme suited to my needs. I am especially grateful to Professor Janna Fox for having guided me throughout the thesis writing process. Indeed, what began as a 'mess' of summaries of articles and books reviewed, voluminous data and findings, magically (Ha!) turned into a thesis under her patient advice. I would like to thank Professor Devon Woods, not only for being a terrific professor, but also for being a friend.

I am most indebted to the Canadian Forces Language School, particularly to Mrs Irene Copping, for having provided OPI samples. Without that contribution, this research could not have taken place.

Sincere thanks go to all of my contacts from the countries and organisations that agreed to participate in this research. I especially would like to thank the raters themselves for having taken time out of their busy schedules to contribute enthusiastically!

I would also like to acknowledge the exceptional support provided by Dr. Richard Monaghan and Mme Lucie Ratté of the Canadian Defence Academy. I was granted leave from my job to complete this thesis and their encouragement throughout the Master's degree undoubtedly contributed to my success.

Sincere thanks go to my colleague Jana Vasilj-Begovic for her support, professional insights and her friendship.

I would like to thank my parents for offering me needed respite from all of the pressures, my aunt Jacqueline for the profound influence she has had on my life, and my parents in-law, not only for their frequent (*but never frequent enough!*) trips to Ottawa but for truly being the best in-laws anyone could ever wish for.

Last but not least, I dedicate this thesis to my ever-so-patient and supportive husband Danny, and to our lovely daughters Zoé and Sasha. Hopefully the girls will come away from their *Maman's* experience with the desire to pursue their own goals and aspirations, whatever they may be.

Table of Contents

List of Tables

List of Figures

List of Appendices

List of Abbreviations

ACTFL – American Council on the Teaching of Foreign Languages
ATC – Air Traffic Control
BILC – Bureau for International Language Coordination
CAEL – Canadian Academic English Language (CAEL) Assessment
CASE – Cambridge Assessment of Spoken English
CDA – Canadian Defence Academy
CFLS – Canadian Forces Language School
EAP – English for Academic Purposes
EIL – English as an International Language
ESP – English for Specific Purposes
FSI – Foreign Service Institute
IELTS – International English Language Testing System
ILR – Interagency Language Roundtable
ITA – International Teaching Assistants
JSSG - Joint Services Sub Group
LSP – Language for Specific Purposes
LT – Language Testing
NATO – North Atlantic Treaty Organisation
NTG – NATO Training Group
OET – Occupational English Test
OJT – On-the-job Training
OPI – Oral Proficiency Interview
PARP – (PfP) Planning and Review Process
PfP – Partnership for Peace
PSC – Public Service Commission (of Canada)
SLA – Second Language Acquisition
SLP – Standardised Language Profile
STANAG 6001 – Standardization Agreement 6001
TLD – Target Language Domain
TLU – Target Language Use
TOEFL – Test of English as a Foreign Language

CHAPTER ONE: INTRODUCTION

1.1North Atlantic Treaty Organization (NATO) Language Testing Background

In order to provide context for the present study, Chapter 1 provides a brief outline of the issues relevant to language testing in the NATO context, describes the origins of the NATO STANDARDIZATION AGREEMENT (STANAG) 6001 Language Proficiency Levels, and informs on the language requirement for member countries. This Chapter also explains the aims, rationale and research questions of the study.

1.1.1 Language Testing in the NATO Context

Although both English and French hold official status at NATO, English has become the operational language, and the teaching and testing of the English language within the NATO community have gained importance in the last few years due to the addition of new countries in 2004 and a large number of joint taskings such as peace-support operations. English language teaching and testing carry high stakes not only for individual military members, but also for countries aiming to meet language goals, or force goals (Partnership for Peace (PfP) Planning and Review Process (PARP) Presentation, Bureau for International Language Co-ordination (BILC) Sofia, 2005). This is true for new NATO countries, for countries aspiring to join NATO or for any military/civilian members seeking positions within NATO. Furthermore, members must have a

Standardised Language Profile (SLP) based on the NATO STANDARDIZATION AGREEMENT (NATO STANAG) 6001 Language Proficiency Levels, the common scale used in this sphere. For some countries this means allocating a large number of human and financial resources to language training, and ensuring a significant percentage of the force has achieved the prescribed SLPs through national testing systems. For individual members, careers may hang in the balance, if because of unsatisfactory test results they are denied advancement possibilities or financial gains linked to various operational or NATO postings. All SLPs are based on tests in the listening, speaking, reading and writing skills, which have as criteria the STANAG 6001 scale (hereinafter referred to as 'the STANAG')[1].

In my present function as National Standards Officer for Language Programmes with the Canadian Defence Academy, a training and education unit within the Department of National Defence, I participate in professional seminars, conferences and specialised working groups on various projects, under the auspices of the Bureau for International Language Co-ordination (BILC). The BILC is NATO's advisory body for language training issues and policy. One of our mutual goals is to share foreign language-related practices, experiences and materials.

[1] Within NATO, there are numerous **STAN**DARDIZATION **AG**REEMENT documents. The only 'STANAG' referred to in this thesis is the STANAG 6001.

Comprehensive language testing systems, be they governmental, educational or private such as the Test of English as a Foreign Language (TOEFL), the International English Language Testing System (IELTS), the Cambridge Assessment of Spoken English (CASE), or the Canadian Academic English Language (CAEL) Assessment, to name a few, generally have a central testing authority responsible for test development, on-going validation, training of developers and testers, administration and monitoring of its testing instruments and results. I am not aware of any standardized language testing contexts where different groups of testers will need to arrive at a score reported on a common scale, but where each 'unit' or in this case each country, acts independently. The STANAG testing system presents itself as standardized, but in fact, each country must establish its own training structure, design its syllabi and teaching materials, implement a testing framework, develop tests, and monitor training outcomes. In an effort to share the load, many countries collaborate and exchange materials and best practices; however since the BILC is an advisory body, it offers consultations on language training and testing issues, but does not impose teaching or testing practices on participating nations. Apart from the Oral Proficiency Interview (OPI)[2] in use in some countries, there are no standardized testing instruments administered throughout countries, and while common testing practices are suggested, there are no mechanisms in place to ensure that the SLPs reported from one country, are equivalent to the scores from the next.

[2] The OPI is a standardized testing instrument, whose validity is based on empirical data. However, modifications made to the standardized test mechanism, to rater training as well as to rating protocols, could potentially alter validity.

This is not to imply that such a situation is automatically and intrinsically problematic. After all, there are many contexts where 'qualifications' can be arrived at by seemingly parallel means. For example, universities hand out certificates and diplomas to graduates, and while criteria may vary somewhat from school to school, even from professor to professor, equivalence is assumed. Military institutions world-wide have established criteria for promotion of their members to higher ranks, and rank equivalence from one country to the next is by no means mistrusted or questioned, (e.g. a Major is a Major). Nonetheless, in the language testing discipline, particularly in a high stakes context, score equivalence, also referred to as score reliability, is generally considered to be a necessary condition for large testing programmes (Bachman 1990, Kane et al. 1999). In this sense, the NATO language framework may be viewed as atypical.

1.1.2 NATO Language Issues

The BILC was established within the NATO Training Group (NTG)/Joint Services Subgroup (JSSG) as a consultative and advisory body for language training matters in NATO. The JSSG as a multinational group of joint training experts supports the NTG by providing a forum to discuss and develop joint individual training. The BILC has the following responsibilities:

- To disseminate to participating countries print and multimedia instructional materials, tests and information on developments in the field of language training.

- To review the work done in the coordination field and in the study of particular language topics through the convening of an annual conference and seminar for participating nations.
- To act as a clearinghouse for the exchange of information between participating countries on developments in the field of language training.
- To provide the sponsorship of STANAG 6001, Language Proficiency Levels (Retrieved from www.dlielc.org/bilc/Constitution2004.doc May 5th, 2006).

Witnessing how the teaching of the English language within the NATO community has gained importance in the last few years, and how costly and time consuming that endeavour has been, the context of language testing in particular has become more and more complex. "The single most important problem identified by almost all partners as an impediment to developing interoperability with the Alliance has been shortcomings in communications" (BILC Report, October 2001). In 2004, seven countries from the Partnership for Peace (PfP)[3] programme joined NATO, for a total of 26. There are presently 20 PfP countries. Language training is central within armed forces due to the increasing number of peace-support operations, and is considered as having an important role in achieving interoperability among the various players.

1.1.3 The NATO STANAG 6001 Scale - Origins

In order to understand the present situation, a brief history of the STANAG 6001 scale is in order. The topic of language scales will be dealt with more thoroughly in Chapter 2.

[3] "The Partnership for Peace is a programme of practical bilateral cooperation between individual Partner countries and NATO." (Retrieved from www.nato.int June 4th, 2006).

In 1976, NATO adopted a language proficiency scale related to the Interagency Language Roundtable (ILR)'s 1968 document whereby language proficiency descriptors had been elaborated. In the 50's, the United States Government needed to inventory the language ability of Government employees, but no standardized system existed at that time in the academic community; therefore, the Government had to develop its own (Wilds, 1975). The Foreign Service Institute (FSI) formed an interagency committee that formulated a 'language' scale ranging from levels 1 to 6, but that scale did not make a distinction among the four skills, as it does today. The scale was eventually standardized to six base levels, ranging from 0 (= no functional ability) to 5 (= equivalent to an educated native speaker), (Herzog, n.d.). In 1968, several agencies jointly wrote formal descriptions of the base levels in four skills—speaking, listening, reading, and writing. By 1985, the U.S. document had been revised under the auspices of the Interagency Language Roundtable (ILR) to include full descriptions of the "plus" levels that had gradually been incorporated into the scoring system. Since then, the official Government Language Skill Level Descriptions have been known as the "ILR Scale", "ILR Guidelines" or the "ILR Definitions" (Herzog, n.d.).

When the STANAG 6001 was first adopted in 1976, it responded to a NATO-wide need to define language proficiency, and to have a common yardstick among countries that would ensure a shared understanding of the language proficiency of members. What was considered necessary at that time was a scheme that was objective, applicable to all languages and that could be used by many different

countries. The descriptors had to be specific enough so that positions, be they military or civilian, could be matched up to proficiency levels, and at the same time, it was important that they remain general enough so that they described the progression considered to be typical of learners, but not specifically related to any particular language or language curriculum. This approach was adopted to respond to these needs due to the great diversity of positions, tasks and roles of military and civilian personnel.

The STANAG from 1976 did not undergo revision until a few years ago. Participants in the 1996, 1997 and 1998 PfP Seminars identified inconsistencies among NATO nations' STANAG ratings. Discussions revealed possible reasons for these: limited details in the original STANAG left room for varying interpretations, pressure to inflate officers' ratings to qualify them for assignments, as well as different testing approaches (achievement, job-performance, proficiency) yielded different results for the same levels of proficiency (BILC Report, October 2001). With the impending accession of a dozen or so countries in the late 1990's, and because the STANAG 6001 was the criteria against which the proficiency of large numbers of military and civilian personnel was to be measured, there was an opportunity to update the scale. In 1999, a committee consisting of expert members from eleven participating countries reinterpreted the descriptors of the original 1976 STANAG. In 2000, the BILC Steering Committee approved the trial of the draft interpretation and the scale was trialled in 2000 and 2001 with participants from 15 countries who attended the first two

instalments of the Language Testing Seminar, in Garmisch-Partenkirchen, Germany (BILC Report, October 2001). The NATO Standardizing Agency integrated the updated interpretation and published Edition 2, in 2003. In 2005, another similar international committee effort led to the development of 'plus levels' which were added as an optional component to the six base level document, in 2006 (BILC Steering Committee Minutes, June 2006). A plus level in this context is defined as proficiency that is more than halfway between two base levels, and as proficiency which substantially exceeds the base skill level but does not fully or consistently meet all of the criteria for the next higher base level (NATO STANAG Plus Level Descriptors – Draft Document, 2005). Seeing as there are potentially 46 countries using the STANAG 6001 as criteria, how the scale is interpreted and used from country to country is a matter of interest to stakeholders.

1.1.4 The Language Requirement

Presentations given at BILC Conferences and BILC Professional Seminars, stress that the need is for personnel in units operating or liaising with NATO land forces, air forces, naval elements or headquarters to be able to communicate in English. "Competency in English language skills is a pre-requisite for participation in exercises, operations and postings to NATO multinational headquarters. The aim is to improve English language skills of all personnel who are to cooperate with NATO forces in NATO-led PfP operations, exercises and training, or with NATO staffs. These individuals must be able to communicate effectively in English,

with added emphasis on operational terminology and procedures" (NATO Partnership Goal (Example) PG G 0355, Language Requirements, 2004). Such 'goals' state that nations should "not only address special measures to increase in general the language proficiency of current officers and NCOs[4] but also the integration of adequate language training as part of their normal career development to ensure adequate language proficiency for future officers and NCOs" (NATO Partnership Goal (Example), PG G 0355, Language Requirements, 2004).

There are some salient points of discussion among BILC-member countries that are linked to this study. First, it is unknown how the language levels were assigned to the positions. Second, STANAG labels may be deceptive. And third, goals may have been set too high for many countries to reach. Regarding the designated position levels, it is highly likely that well meaning individuals took out the STANAG 6001 descriptions and made what they thought were reasonable decisions (BILC member #1, personal communication, 2006). Since the BILC was not consulted in the setting of standards for posts, it may be that language requirements for positions have been set without the help of language professionals. The issue of Standardized Language Profiles (SLP) for job requirements, having been possibly haphazardly assigned may be one of the factors driving the training and testing towards level 3, which in the STANAG scale is called 'Minimum Professional Proficiency'. This situation may be compounded by diverse interpretations as to what 'Minimum Professional

[4] Non-Commissioned Officers.

Proficiency' means and what constitutes professional language ability, even though the STANAG 6001 (Edition 2) level descriptors reflect that level 3 speakers are highly proficient speakers[5]. The descriptors set high expectations of proficiency for this level, and training to this level of proficiency in countries where there is no established English learning tradition and where the political past has not provided many opportunities to learn English, is a highly challenging undertaking, to say the least.

A second recurring issue is linked to the STANAG 6001 level 'labels', including Edition 2. According to a BILC member, a threshold Level 2 is too low for most of the officer jobs as laid out in PG G 0355 but Level 3 is too far a reach for most of the language programs in NATO/PfP countries (BILC contact #1, personal communication, 2006). Correspondence with another BILC member supports the comments above:

> As you know, job descriptions have a language requirement stated as "SLP". SLPs are normally determined by the supervisors of those jobs and/or by [personnel][6] without the assistance of language experts, and are either based on tradition (the SLP given to the job during the last review) or can be an attempt by the [HR person] to match up the job requirements with statements from STANAG 6001. Often added to that is the feeling that there should be a relationship between rank and proficiency level (NCOs can be 2, most officers a 3, and higher ranking officers a 4.) I can imagine how [someone] who is not a language expert, reacts to the level titles:

[5] The full NATO STANAG 6001 Language Proficiency Descriptors Ed. 2 can be found in Appendix A.
[6] Editorial changes in this text were made to specific title references and are indicated by square brackets.

Level 1- Elementary
Level 2 - Fair (Limited Working)
Level 3 - Good (Minimum Professional)
Level 4 - Very Good (Full Professional)

> I'm sure the natural tendency is to select at least Level 3 for staff officer positions because it is "Good" and "Professional" (BILC contact #2, personal communication, 2006).

At the BILC conference in June 2006, a committee was asked to review the current labels for each STANAG 6001 language level and recommended that five out of the six be changed to better reflect the level standards (BILC Steering Committee minutes, June 2006).

Furthermore, some BILC members have been questioning whether the level of expected proficiency set forth in the Partnership Goals has been set too high (BILC Conference - Budapest, June 06). Some argue that not all positions require officers with level 3 across the four skills, and this may be an unnecessarily high standard for some nations to meet. This push towards level 3 may cause a two-fold washback[7] effect; the first is a possible drift upwards of the standard in order to meet the goals, the second is personnel who are reported as having proficiency levels that may or may not be adequate to fulfil their military mandate. Although these issues are not directly investigated here as they are outside the scope of this study, they are relevant because their impact can be felt at all stages of the training/assessment loop.

[7] Washback is a term used to refer to the impact that testing (and I would add policies to the definition) has on language teaching and learning (Alderson and Wall, 1993).

1.2 Rationale for Study

Many of the more established NATO countries that were operating on the previous STANAG[8] chose to adapt their testing system to the new interpretation (e.g. Canada), while the newer NATO countries set up their teaching and testing programmes based on the STANAG 6001 (Edition 2) from the outset. With the help of the U.K., the U.S.A and Canada, advances were made in PfP and 'new' NATO countries regarding curricular and assessment programmes. Also, many countries teamed up by attempting to reach a common understanding of the scale, by trialing items, or piloting new test versions on each other's students. However, despite these and other efforts by the U.S. Defense Language Institute, the British Council's Peacekeeping English Project, and the BILC sponsored Language Testing Seminar, various administrators, testers, and teachers who are members of the BILC have commented that there still appear to be differences among countries in the interpretation of the STANAG levels (personal communications, BILC Conference June 04, and BILC Seminar October 05). But what is the nature of these differences? Are they attributable to the tests in place in some countries? Are they related to the application of the scale, to the way tests are designed, conducted, rated? One can only speculate since as to date, there has been little systematic research in this area.

[8] The STANAG originally promulgated 21 October 1976.

It is clear however, that from the standpoint of assessing second or foreign language speaking in particular, there are numerous questions that come to mind regarding the people involved in this challenging activity. While it is recognized that it is valuable to have people to judge peoples' language ability, it is also acknowledged in the language testing literature that human judgement is inevitably going to be subjective (Bachman 1990; McNamara 1996). Both the examinee and the trained professional will have been part of the equation, and it is clear that the sums of the parts do not always add up to the same score. Under the best of conditions, assessing language ability by means of performance testing (in its broad sense) can lead to measurement error; for example, there can be differences or variance among interviewers in the way they elicit their sample, in the choice and difficulty of tasks presented to the examinee, and in the way raters interpret the rating scale and apply the criteria, all of which may impact the ratings assigned to examinees. In a high stakes testing context, such as NATO, these variables are of interest not only to researchers and practitioners, but also to end 'users'[9].

1.3 Aim of Study & Research Questions

It has been reported by Green and Wall (2005) that although considerable progress has been achieved in many Central and Eastern European countries regarding testing practices, some members of national testing teams apparently

[9] A user in this context refers to people or organisations that make decisions, based on the individuals' proficiency results.

find that the STANAG 6001 scale is difficult to use, and that interpretation is perceived to vary from country to country. These differences are said to pertain not only to new members, but also to more established NATO countries. In the past 30 years since the adoption of STANAG 6001 by NATO, there have been few, if any, studies investigating language testing issues in the wider NATO sphere[10], aside from the Green and Wall (2005) report outlining some of the challenges various countries have and are facing. In occupationally related testing contexts, Wesche (1992) remarks that relatively little professional literature is published outside of the academic network, probably since most of the test development is done for practical purposes and is operational.

Perceptions regarding discrepant interpretations of the STANAG are anecdotal since no official data have been gathered on inter-country rating practices; however, important questions do remain un-investigated: Are there significant differences from country to country? Are these differences any more significant than the ones between raters within a country? In view of the lack of existing research literature on language testing in the military context, the following exploratory study was undertaken with the aim of contributing to the understanding of rating practices among various NATO countries.

The overarching research question was: **How comparable or consistent are OPI ratings across NATO raters and countries?**

[10] There may have been studies commissioned internally by countries, but these have not been publicly circulated.

Research questions were organised as follows:

1. Research questions pertaining to the ratings:

- How do ratings of the same oral proficiency interviews (OPIs) compare from rater to rater?
- Would the use of plus levels increase rater agreement?
- How do the ratings of the OPIs compare from country to country?
- Are there differences in scores within the same country?

2. Research questions pertaining raters' training and background:

- Are there differences in ratings between raters who have received varying degrees of tester/rater training and STANAG training?
- Did experienced raters score more reliably than inexperienced ones? Are experienced raters scoring as reliably as trained raters?
- Are there differences in ratings between participants who test part-time versus full-time, are native or non-native speakers of English, and are from 'Old' and 'New' NATO countries?

3. Research questions pertaining to the rating process and to the scale:

- Do differing rating practices affect ratings?
- Do raters appear to use the scale in similar ways?
- What are the raters' comments regarding the use and application of the scale?

In sum, this study investigated the inter-rater reliability of oral proficiency scores among raters across 18 countries and 2 NATO units, some long-standing and some newly included, through the analysis of qualitative and quantitative data.

The study examined rating practices of the various raters as well as their interpretations of the STANAG, by asking participating tester/raters to rate two sample oral proficiency interviews (OPI) that were elicited in live tests, in Canada in 2005. It explored the efficacy of 'plus levels' or plus ratings. Would plus levels have been deemed helpful by the raters in attributing their scores, had their rating protocol made use of them? In addition, the study explored the ways in which raters use the various STANAG statements and inherent rating factors to arrive at their ratings.

It is hoped that the findings will inform decision makers and raters, generate useful discussion among participating countries and ultimately, promote international benchmarking efforts. This study is a small first step in providing empirical information on the status of the NATO STANAG 6001 testing standards beyond the present anecdotal rhetoric.

1.4 Organisation of Thesis

Chapter 1 of this thesis situated the study within the NATO context and defined issues relevant to the STANAG scale. The literature review in Chapter 2 is divided into two sections. The first section investigates the literature in language testing research on the construct of testing speaking as viewed through the lens of the five eras in language testing, identified by Shohamy (1996). They are the organizing principle in synthesizing the concepts, and in presenting the most

relevant approaches in relation to the NATO context. The second section of the literature review presents studies whose findings have shed light on rater variance or bias by exploring interviewer variation, rater/scale interaction, rater training, the native speaker rater, and some paralinguistic features of assessment. These studies have focused on language testing in educational settings, some have looked at work-related or occupational contexts and others have been conducted in government. The research methodology of the study is presented in Chapter 3, while Chapter 4 of the thesis presents and discusses the findings. Chapter 5 offers conclusions and acknowledges limitations of the study. It also discusses implications and makes a few recommendations for the NATO language testing context.

CHAPTER TWO: LITERATURE REVIEW

This Chapter, the literature review, is presented in two sections. The first section reviews literature related to language testing constructs, reviewing the five main eras as identified by Shohamy (1996). The second section of the Chapter introduces literature related to rater variance in assessment.

2.1 Language Testing Constructs

2.1.1 Introduction

The central preoccupation of language testing (LT), a subfield within applied linguistics, is the measurement of language knowledge. But what exactly is being measured? And how do we go about measuring it?

Researchers have long attempted to answer 'what it means to know a language' a question put forward by Spolsky (1973); and in the last thirty years, many theoretical models have been proposed in an attempt to identify the traits underlying language knowledge. Shohamy writes: "It is the complexity of the language trait that creates a need for a special discipline called language testing, for there is still no full understanding of what is involved in knowing a language" (Shohamy 1996, p.143). When trying to characterize the manifestation of competence, or to define ability during a performance on a language test, there are many issues and factors that come into play; issues of construct, methods, validity

and reliability, standards, scales, ethics, etc. Bachman (1990) points out that the interpretations of test scores are limited because our observations of performance are indirect, incomplete, imprecise, subjective, and relative. He adds that in order to minimize the effects of these limitations, and to maximize the reliability of test scores and the validity of test use, we should follow three fundamental steps in the development of tests: (1) provide clear and unambiguous theoretical definitions of the abilities we want to measure; (2) specify precisely the conditions, or operations that we will follow in eliciting and observing performance, and (3) quantify our observations so as to assure that our measurement scales have the properties we require (p.50). This, he goes on to say, is how we may provide the essential linkage between the abilities we want to measure on the one hand and the observations of performance on the other, and thus form the basis for both test development and the interpretation and use of test scores.

The difference of opinion surrounding the meaning of the term 'language ability' is unresolved in the LT literature to this day, with terms such as 'performance', 'competence' and 'proficiency' having taken on a variety of meanings. These terms will be defined from the various perspectives and approaches to testing language knowledge presented below.

2.1.2 Approaches

Shohamy (1996) identifies five phases or eras in the evolution of language testing: the discrete point era, the integrative era, the communicative era, the performance

testing era and the alternative assessment era, 'each reflecting the different definitions of language knowledge of the time and the specific measurement procedures that went along with it' (p.143). The first two, the discrete point and integrative eras, as well as the last one, the alternative assessment era will be described only briefly below. The communicative era and the performance testing era, will be discussed thoroughly in relation to the underpinnings of the construct of speaking as it is represented in the literature and how it relates to testing speaking in the context of interest for this study.

2.1.2.1 The Discrete Point Era

The first of the phases, the discrete point era, focused mainly on the assessment of isolated items, such as grammar or vocabulary, and test formats such as true-false or multiple-choice tests were commonly used. Even productive skills of speaking and writing were tested without requiring that examinees actually 'produce' language.

2.1.2.2 The Integrative Era

In the integrative era, global language samples were used so that language tasks on tests were contextualised, not isolated. The cloze test is an example of this method, where words are deleted from larger texts, and test-takers have to fill-in the blanks. Oller (1975) was a proponent of this approach, claiming that "it tapped integrative language and reflected a unitary notion of language which underlies the language knowledge based on the learner's pragmatic *grammar of*

expectancy" (italics in original, Shohamy, 1996, p.144). Test tasks that combine two skills such as listening to and writing notes on a lecture, for example, are also described as integrative (Davies et al, 1999). Integrative tests were perceived as more valid than discrete point tests, in which language elements are de-contextualised, and came closer to representing processes necessary to perform communicative tasks. Although these methods became popular and are still in use today, criticism of this approach centered on the fact that these tests lacked evidence of construct validity, especially when test-tasks only required test-takers to fill in the blanks, or in the case of speaking, produce isolated sentences.

2.1.2.3 The Communicative Language Testing Era

In the third phase, a shift to testing communicative language arises from research in linguistics and language teaching. Language teaching saw a move away from the teaching of de-contextualised language elements to those elements in communicative language (Shohamy 1996). Particularly influential was Hymes (1972), who introduced concepts of communicative competence and communicative performance to the field. He argued that linguistic competence, as defined by Chomsky (1965), comprising, on the one hand, the largely unconscious grammatical knowledge of an idealised speaker of a language (competence), and on the other, the manifestation of language use (performance), was too narrow a model, and proposed a distinction between language *knowledge* and *ability for use*. In Hymes's view, communicative competence is represented by *knowledge* of grammatical and other formal linguistic rules, sociolinguistic

rules, etc., and *ability for use* is comprised of a range of underlying language-relevant but not language exclusive cognitive and affective factors (including general reasoning powers, emotional states and personality factors) which are involved in performance of communicative tasks (McNamara 1996), thus pointing to the relationship and interaction between grammatical and sociolinguistic competence, and performance.

There have been a few re-formulations of the influential Hymes model such as the one proposed by Canale and Swain (1980). In this view, communicative competence is restricted to language *knowledge* only and is hypothesized to comprise linguistic, sociolinguistic, discourse and strategic competence. Their model deliberately excludes Hymes's notion of *ability for use* because they considered this to be impossibly complex to model "since no theory of human action could adequately explicate it" (McNamara 1996, p.61). Critics of the Canale and Swain model cited the exclusion of performance from their view as problematic, and called for a need to further study how the four components interacted in communicative competence (McNamara 1996).

During this era, there was a growing emphasis on language proficiency. It is during this period that the well-known U.S. 'proficiency movement' emerged and became recognized outside of the U.S. agencies, which had been its main proponents. The trend of testing communicative language, which is language that is viewed as 'authentic', 'direct' and 'communicative', through an interview with

an interlocutor, (as opposed to the practice of testing speaking in a language lab), became widespread. The Foreign Service Institute (FSI) interview, and the oral proficiency interview (OPI) in use in other U.S. agencies such as the Defence Language Institute (DLI), are examples of direct and communicative tests (Shohamy 1996).

a) General Proficiency

One of the definitions of proficiency relates to a general type of knowledge or competence in the use of a language, regardless of how, where or under what circumstances it has been acquired (Davies et al., 1999). Proficiency is now "viewed as multifaceted, with recent models specifying the nature of its component parts and their relationship to one another. There is now considerable overlap between the notion of language proficiency and communicative competence" (Davies et al., 1999, p.153). Proficiency is thus conceptualised as a global construct that transfers across contexts, tasks, and events, while proficiency tests attempt to sample the underlying competence by eliciting behaviours on tests that generalise to domains of interest. Proficiency can also be defined as the language knowledge that is needed to function in a future situation. The performance elicited in a speaking proficiency test, for example, is usually measured or judged against a set of criteria, represented in a rating scale.

b) Language Scales

As mentioned above, ratings scales are the instruments that were developed as criteria for assessing the quality of the language samples obtained from the

different types of communicative tests (Shohamy, 1996, p.147). Results of speaking tests can be expressed as numerical scores, as profiles or as verbal categories such as 'elementary' or 'advanced', etc. In order to categorise such results, a series of descriptive statements usually accompanies the score ranging from lowest to highest and presented in the form of a rating scale (Luoma 2004). These ascending series of descriptors are in some sense related to Second Language Acquisition (SLA); however, are not a direct application of SLA research since the purpose of SLA research has not been scale construction, and because of the complexity of defining speaking ability in general, it is uncertain how feasible the modelling of all learners' learning paths can be expressed (Luoma 2004). North (2000) states that "definitions of scales of language proficiency in the literature depend somewhat on the perspective of the writer and the argument they are in the process of putting forward" but adds that "despite the problems attached to trying to describe complex phenomena in a small number of words on the basis of incomplete theory, scales of language proficiency have the potential to exert a positive influence on the orientation, organisation and reporting of language learning" (North 2000, p.13).

Scales of proficiency can be used to:

- Provide a 'stereotype' with which the learner can compare his self image and roughly evaluate his position,
- Provide learner goals and descriptions of proficiency at notional levels in order to provide targets for learners, to allow the results achieved to be measured against expected outcomes, and to provide society with a pragmatic means of placing students in appropriate future learning or work

environments by referring to an individual's profile across the sub-scales of the system,

- Increase the reliability of subjectively judged ratings, especially of productive language skills, and provide a common standard and meaning for such judgments, and
- Enable comparisons between systems or populations using a common metric or yardstick. (Selected from North 2000, p.12).

Many existing scales of language proficiency are *holistic scales* in that they express an overall impression of an examinee's ability in one score (Luoma 2004). These scales are practical from a reporting perspective, and "make rating quick because there is less to read and remember than in a complex grid with many criteria" and are also "flexible in that they allow many different combinations of strengths and weaknesses within a level" (Luoma 2004, p.62). These types of scales contrast with *analytic scales*, which normally contain a number of criteria, such as grammatical accuracy, vocabulary use, fluency, cohesion, etc, each representing the examinee's profile separately. The advantages of such scales "include the guidance that they give to raters, and the rich information they provide on specific strengths and weaknesses in examinee performances" (Luoma 2004 p.68).

In the past 25 years or so, there have been developments of a number of scales of language proficiency that are related in one form or another to the Foreign Service Institute (FSI) and the Interagency Language Roundtable (ILR) scale. The ILR

and ACTFL scales are best known for this approach. Another scale derived from this approach is the NATO STANAG 6001 scale[11].

The influence of this scale in proficiency testing cannot be underscored. In Canada, both the Department of National Defence (DND) and the Directorate of Foreign Affairs and International Trade (DFAIT), use the ILR scale for foreign language testing purposes. Second language testing (English and French) of most Canadian government personnel, including the Canadian Armed Forces is carried out by the Public Service Commission (PSC) of Canada. The PSC's language scale levels were derived from the ILR level descriptors (PSC document, 1993). The STANAG is only used with foreign nationals who study English or French as a foreign language in Canada. However, as outlined in Chapter 1, the STANAG is in use in all NATO and PfP countries to report proficiency of members to NATO, as well as in many European countries to measure and certify members' proficiency in-country.

Although the ILR scale and the STANAG scale have never claimed to embody theories of SLA, its detractors have voiced loud concerns over the lack of empirical evidence to back up the claims that it represents a language progression typical of learners. Brindley (1998) points out that although these scales have been met with widespread acceptance within the language training fraternity, their theoretical foundations are questionable. He and others question the validity of

[11] The STANAG is equivalent in terms of bands of descriptors, levels and orientation to the ILR scale, or to use Lowe's (1985) terms about the ACTFL scale, it is a 'derivative' and 'commensurate' scale.

the scale as indicators of language ability on the basis that the scales fail to demonstrate that they indeed describe language development according to what is known about SLA, and therefore could be seen as not having construct validity, that is, not measuring what they intend to measure. Researchers and practitioners who develop and use such scales as the American Council on the Teaching of Foreign Languages (ACTFL) and the ILR, have defended their position and have pointed to empirical evidence for validity (Henning, 1992). Byrnes (1987) captures the essence of the scale by pointing out that the rating scale is built around a postulated hierarchy of tasks, and states that this implies that a substantial ability to perform effectively the tasks of a certain level must be observable before we can assume efforts to handle the tasks of the next higher level. The tasks in this type of framework are not work-related tasks (such as in performance tests defined below) but are what Lowe (1985) terms 'task universals' that characterise specific levels in the framework, such as the ability to describe, narrate in different time frames, hypothesize, etc., depending on the level of demonstrated ability. Lowe (1985) defines the nature of ILR proficiency, below:

> *...[P]roficiency equals achievement (functions/tasks, content, accuracy) plus functional evidence of internalized strategies for creativity*, i.e., to be proficient you must be able to use the language ... [P]roficiency must be reflected in performance: *proficiency is the global rating of general language ability over a wide range of functions and topics at any given level.* (Italics in original, Lowe 1985 p.p.15-16).

McNamara (1996) writes that the scope of this definition has been criticized by Bachman and Savignon (1986) and Bachman (1988) regarding the weight placed

on sociolinguistic and discourse-related knowledge on the one hand, and on accuracy on the other. McNamara quotes Lowe (1988) who suggests that the proficiency movement and Bachman's CLA model 'may prove incompatible. The position of accuracy within CLA remains unclear' (as cited in McNamara 1996, p.77). Interestingly, McNamara then goes on to quote Bachman and Savignon (1986) who note many commonalities between the two approaches:

> Discussions of language proficiency [within the proficiency movement] explicitly include components of language use other than grammar... We are struck more by similarities than differences [in the two approaches]. Indeed, we find substantive differences only in the related importance accorded to the different components, rather than in the components themselves (as cited in McNamara 1996, p.77).

The proficiency approach, as indicated above, can be related to componential models. In their Relative Contribution Model (RCM), Higgs and Clifford (1982) hypothesized a relationship between the levels on the scale, and the relative role played by language ability components of other models. They argue that for any given level, there is a particular mix in the degree of importance that a number of factors, or components, such as vocabulary, grammar, fluency, pronunciation, and sociolinguistic knowledge, will carry depending on the speaker's performance. These components are subsumed in other models such as Canale and Swain (1980) and Bachman and Palmer (1996) under slightly different wording.

Bachman (1990) states that a common thread that runs through much writing in language testing is the belief that a precise, empirically based definition of

language ability can provide the basis for developing a 'common metric' scale for measuring language abilities in a wide variety of contexts, at all levels, and in many different languages (p.5). He goes on to say that such a scale would always indicate the same level of ability, regardless of the skill being tested, and therefore could be used in any context and even for different languages. Bachman (1990) cites Bachman and Clark (1987) who state the advantages of a common metric as follows:

> [T]he obvious advantage of such a scale and tests developed from it is that it would provide a standard for defining and measuring language abilities that would be independent of specific languages, contexts and domains of discourse. Scores from tests based on this scale would thus be comparable across different languages and contexts. (cited in Bachman 1990, p.6).

As North asserts with regard to a common reference framework, the purpose is to "provide a metalanguage of criterion statements which people can use to roughly situate themselves and/or their learners, in response to a demand for this. It is widely recognised that the development of such a taxonomy entails a tension between theoretical models developed by applied linguists (which are incomplete) on the one hand and operational models developed by practitioners (which may be impoverished) on the other hand" (North 1998, p.242). Until an empirically derived model that supports theories of SLA is available, taxonomies such as the ILR, the STANAG, or the Common European Framework of Reference (CEFR) are useful.

c) Oral Proficiency Interviews

As mentioned above, the OPI is a very popular instrument for assessing second or foreign language speaking proficiency in many U.S. government institutions and in use in some NATO countries (e.g. Canada). Nongovernmental institutions like the Educational Testing Service (ETS) and the American Council on the Teaching of Foreign Languages (ACTFL) also use the OPI. The *ILR Handbook on Oral Interview Testing* (1982) and the more recent *OPI 2000 Tester Certification Workshop* (1999) manuals expand on the structure, elicitation techniques, and rating procedures of OPI testing. Because the two Canadian OPI samples used for the purposes of this research follow the standard OPI structure and have similar rating procedures, they are important to the study, and will be briefly outlined below. (A more detailed summary of the OPI framework is provided in Appendix B).

The OPI ratings are expressed in global or holistic terms, even though the assessment is based on the factors that contribute to the candidate's overall speaking proficiency: pronunciation, fluency, grammar, vocabulary, and sociolinguistic/cultural factor. The candidate's level of proficiency reflects the integration of all of these factors/skills in performing a variety of language functions. In the early 1980s, the OPI underwent major changes with respect to both the elicitation and rating procedures. Higgs and Clifford (1982) introduced the functional trisection[12] to the OPI system, which had major conceptual

[12] According to the OPI 2000 Tester Certification Workshop Manual (1999) it is now referred to as the Functional Quatrosection with the addition of 'Text Type'.

implications for both elicitation and ratings procedures. The functional trisection includes three components: "*Function* that refers to the types of language use task that the speaker is expected to be able to carry out at the specified level; *content* that refers to the kinds of topics or subject areas at issue in the communication, and *accuracy* that describes the degree of structural, lexical, and phonological precision with which the examinee is able to communicate" (Clark and Clifford 1988, p.38).

Although the OPI appears to be a complex and difficult testing procedure to master (see Appendix B), training in the interpretation of the scale, in language elicitation techniques and in rating procedures have demonstrated that there can be high levels of inter-rater and intra-rater reliability with such a test on a proficiency scale (Clark 1986). Reliability refers to the consistency and accuracy of measurement both from one occasion to another, and from one test to another. In multiple-choice tests or any objectively scored instrument, reliability is generally estimated by internal consistency, which determines how well the items on a test correlate with each other. Two methods are frequently used to verify reliability on subjectively scored tests such as the OPI or on writing tests with productive tasks such as essays: inter-rater reliability and intra-rater reliability. Inter-rater reliability refers to how much the testers' judgments of the individual's performance are comparable. Intra-rater reliability looks at how much the same tester consistently assesses the same examinee's performance on two different occasions.

Clark (1986) studied the comparability of proficiency interview ratings and techniques employed by three U.S. government agencies that use the ILR scale. In his reliability study, two teams of two testers from each agency tested examinees in German and in French language studies, for a total of 115. The ratings the examinees received were compared using analysis variance and chi-square, and indicated no significant statistical differences from one agency to the next in global ratings.

The Canadian Forces Language School (CFLS) evaluation grid used for rating OPIs is a marriage of a communicative theoretical model, with a general proficiency orientation. The two are not mutually exclusive in this context. First, CFLS incorporated a rating grid with rating factors and a rating protocol that differs slightly from the OPI above. The Canale and Swain (1980) model of communicative competence was incorporated into the testing framework in the late eighties, in the form of a grid or checklist which intended to guide the interviewer/rater to pay attention to an examinee's demonstration of the four components, in addition to using the descriptors in the scale. The grid was modified over the years to reflect more recent approaches to the construct of speaking such as Bachman's (1990) model and since 2003, is comprised of 5 sections: the first representing the tasks/functions to be performed at each of the 4 levels of language proficiency tested[13], a linguistic competence component (grammar, vocabulary) a sociolinguistic/strategic competence component, a

[13] The test is multi-level, however neither Level 0 nor Level 5 is tested in this framework, in Canada.

discourse competence component (coherence, cohesion) and a delivery component (fluency, pronunciation)[14]. At each of the four proficiency levels tested, and within each of the four components, the evaluation grid defines the performance expected of speakers at these levels. The OPIs conducted follow all of the phases, elicitation techniques and rating protocol of the OPI mentioned above except for two main differences: the criteria is not the ILR but the NATO STANAG 6001 scale, and probes are not presently used to assign plus levels, but to ensure that the ceiling has been established and that the testers are on the appropriate working level. (See Appendix B for more OPI information).

2.1.2.4 The Performance Testing Era

As mentioned above, the OPI is considered to be a general proficiency test, not a performance test. Again, there are varying definitions in the LT literature. The term performance testing was "used mostly in the communicative testing era as one of the features of communicative tests, along with a number of terms such as *direct, functional* and *authentic* testing" (italics in original, cited from Shohamy, 1996, p.147). Performance in this context referred to what the test-takers were expected to replicate, i.e. the type of language used in non-testing situations (Bachman 1990). In language testing, a performance test is a test "which requires candidates to perform tasks which replicate the sorts of things they are or will be required to do in particular contexts" (Davies et al., 1999, p.143). Performance testing has been most frequently used in other settings, where one would test the ability to perform certain tasks, especially in an occupational setting.

[14] The CFLS Evaluation Grid can be found in Appendix E.

Performance tests are the most common method of testing languages for specific purposes (LSP).

Other models and conceptions of language also emerged during this era. "Since performance testing consists of the interaction of linguistic skills and a specific domain it is no longer a *pure* language test, but rather depends heavily on the knowledge of the domain in which the language is exercised" (Shohamy, 1999, p.150). Bachman (1990) hypothesized a model of *communicative language ability* (CLA) comprised of three components: language competence, strategic competence and psycho-physiological mechanisms. He built upon the Canale and Swain model by expanding the role of strategic competence, which had been primarily looked at as pertaining to the ways language users employed compensatory strategies, such as circumlocution, to compensate for their lack of language knowledge. Also, Bachman's description of the functions of strategic competence in planning, assessment and execution provided a means for explaining how the various components of language competence interacted with each other (p. 104). Bachman and Palmer (1996) expanded the Bachman (1990) model, "to include for the first time an explicit modelling of the role of affective factors in language use; that is, the role of non-cognitive factors underlying performance is explicitly addressed" (McNamara, 1996, p.72). This Model of Language Use (MLU) presents refinements from the previous one whereby the problematic term 'competence' is discarded and an affective component is added. *Strategic competence* is re-conceptualised as a grouping of *Metacognitive*

strategies (Bachman and Palmer, 1996). They reject the notion of the four traditional skills of listening, speaking, reading, and writing and instead "argue that they should be seen as language use activities. Their concept of language knowledge identifies components of knowledge that are relevant to all modes of language use" (Luoma, 2004, p.101). Luoma views this model as useful for test development since "the concept of language ability names dimensions in language use, and it helps assessment developers check how well rounded their tasks and assessment criteria are in terms of them" (p.101). However she also points out their limitations regarding other knowledge types (which appear in the model but receive less emphasis), and suggests that developers may want to use this model in conjunction with other frameworks (Luoma, 2004).

Performance tests are said to be useful in situations where a clientele shares second language needs that can be identified, and translated into tasks and test design. This practice is common in the sub-field of LT that is testing languages for specific purposes (LSP), but it has not been without its fair share of critics. Researchers such as Davies (2001) have argued based on a review of LSP tests that there is no theoretical basis for LSP testing and that it has not proven itself to be more valid than a general proficiency test. Davies (2001) offers the view that LSP testing is premised on the assumption that there are distinct varieties of a language, for example 'medical English', 'legal English', 'business English', (p.133). For the purposes of this discussion, 'military English' is added to the list. Such a postulation implies that if one has sufficient knowledge of the type of

language which one encounters in these specific areas, one can participate in communications with others who use this same 'variety' of language. Davies (2001) reviews existing research on LSP tests and concludes: "that what distinguishes varieties or LSP test tasks is content rather than language, and that the justification for LSP testing is practical need and pragmatic effect" (p.134). He points to research by Jensen and Hansen (1995) on English for Academic Purposes (EAP) listening whereby findings demonstrated no evidence that "high proficiency listeners who have indicated prior study of a topic will perform better on lecture comprehension than listening alone would predict" (as cited in Davies, 2001, p.139). Research by Jennings et al. (1999) also supports the fact that allowing for choice of topics does not indicate a significant difference in performance for test-takers. Scepticism from other researchers such as McNamara (1990), Skehan (1984) and Davidson (1998) are cited in Davies as questioning the validity of the construct. Even so, Davies (2001) concludes that despite an apparent lack of evidence that would support a difference between LSP's testing of content versus general testing of content, and "since it does not predict any less well than a general proficiency test and since there are sound pragmatic reasons for its use, then the LSP test project remains worth pursuing" (p.144). This is arguably a weak endorsement from Davies, but not a dismissal altogether.

What then are the pragmatic reasons for LSP testing? From a practical end, it could be argued that it is useful to know how well a test taker's command of the

language in specific situations, performing specific language-related tasks can be demonstrated and therefore, assessed. So, what constitutes a specific purpose language test? According to Douglas (2000):

> A specific purpose language test is one in which test content and methods are derived from an analysis of a specific purpose target language use situation[15], so that test tasks and content are authentically representative of tasks in the target situation, allowing for an interaction between the test taker's language ability and specific purpose content knowledge, on the one hand, and the test tasks on the other. Such a test allows us to make inferences about a test taker's capacity to use language in the specific purpose domain (Douglas, 2000, p.19).

Douglas (2000) discusses at length the need to "first describe the target language use (TLU) situation in terms of characteristics of context and task, then specify how these characteristics will be realized in the test so as to engage the test taker in test tasks, performance on which can be interpreted as evidence of language ability with reference to the target situation" (p. 20). This requires a thorough needs analysis of the TLU situation or domain, as well as of the tasks that are required in the specific context. That is to say that the tasks present in the test must reflect tasks that one *infers* will be represented in the non-test TLU situation. This speaks to the need of replicating authentic circumstances for language use, a setting from which the performance of the test-taker will be generalised. This is viable in some contexts, but may be thorny in others, where security reasons may prevent the inclusion of developers or non-members for research purposes, such as some military contexts, for example. Certainly, settings such as universities whereby English for Academic Purposes (EAP) programmes are not entirely

[15] Bachman and Palmer (1996) use the term domain.

discipline-specific but contain many common features, lend themselves relatively well to a study of tasks and TLU domain.

It has been pointed out that despite the importance of linguistic interoperability, little NATO-wide research has been carried out into actual language used on postings, on missions and in operations (Crossey, 2005; Green and Wall, 2005). In light of the fact that there is such a wide variety of military tasks, and that a rigorous needs analysis of the type of language that is necessary to carry out the tasks for which the military are being trained has yet to be conducted, most countries such as Canada have developed training materials which are based on a framework of a general proficiency language progression through the STANAG 6001 scale. These materials are often supplemented with vocabulary commonly used in military communications, and contexts are situated within the general military domain, in one form or another. General proficiency with a military flair has been the approach of choice.

Having said that, it appears clear that with a needs analysis of the TLU domain and tasks, and a comprehensible description of the characteristic of these tasks one can begin exploring the test specifications, a blueprint of sorts, which will form the basis from which the test-taker will be assessed. But seeing as specific purpose content knowledge is an inherent part of the equation, one must first disentangle this knowledge from language use. The test designer must be able to distinguish between what constitutes language knowledge and background

knowledge in order to be fair to the test-taker. It still is necessary to understand what Douglas means by specific purpose language ability. Douglas defines it as follows:

> Specific purpose language ability results from the interaction between specific purpose background knowledge and language ability, by means of strategic competence engaged by specific purpose input in the form of test method characteristics' (Douglas 2000, p.40).

The Bachman and Palmer (1996) components of communicative ability, language knowledge and strategic competence, are central concepts in the LSP literature. Douglas (2000) stresses the importance of strategic competence since it 'serves as a mediator between the internal traits of background knowledge and the external context, controlling the interaction between them (p.28). His view seems framed within acceptable SLA models, and would appear to be a valid construct for making inferences of test taker's performance on test tasks, and being generalisable to the actual TLU situation. But it must be made explicit that it is only *those* tasks and the specific language knowledge/ability that is transferable. There is evidence to show that language ability in one discourse domain doesn't imply transference to another. Shohamy (1998) reports that various studies have examined how specific discourse features manifest themselves in testing situations. Specifically, she mentions that Douglas and Selinker (1985, 1990) investigated the effect of subject matter scores on reading comprehension scores. "Their research demonstrated how performance can vary according to choice of topic. They found that familiarity with the content domain affected performance and, as a result, the nature of the fluency of the communication. They concluded

that generalisations from one discourse domain to another on language tests may not be justified" (as cited in Shohamy 1998: p.159). The issue of transferability in the government and military context is crucial seeing as job postings of military and civilian employees are usually of a short duration, and many jobs probably require both general and specific language-related capabilities. In order to warrant transferability, the wide sampling of topical domains is necessary, and is a pre-requisite of a general proficiency test. However, as with any discourse community, it is assumed in this context that the lexical particulars will be acquired on-the-job as one acculturates to the environment, if one already has enough general language proficiency.

But the question of background knowledge remains. How much background knowledge is necessary, and why should it matter if language is what concerns us? Douglas (2005) argues that "true specific purpose tests must share a focus on an interaction between language knowledge and specific purpose background knowledge and that background knowledge, far from being a factor leading to 'construct irrelevant variance' (Messick, 1990), or error, may be an essential aspect of the construct of specific purpose language ability" (as cited in Douglas 2005, p.859). As Davies (2001) has written, "LSP testing cannot be about testing for subject specific knowledge. It must be about testing the ability to manipulate language functions appropriately in a wide variety of ways" (p.43). Douglas (2005) agrees that language testers are not in the business of assessing how much knowledge one might have about a particular field. Despite the points made, he

still opines that "in measuring language knowledge in professional, vocational or academic contexts, we may indeed decide to include specific purpose background knowledge in the construct to be measured on the grounds that it is very difficult to separate language knowledge from the content that it conveys" (Douglas 2005, p.859). Bachman and Palmer (1996) have argued that when we are testing people who have had experience in their respective fields, then we can make the assumption that some background knowledge is present, but that novices may not have this knowledge, perhaps despite their language ability, and that this could lead to a source of measurement error in testing LSP. Douglas (2005) suggests that a test of background knowledge may then be needed to sort out the two traits. Douglas (2000) proposed "a continuum of specificity based on the amount of background knowledge required to carry out test tasks, and the narrowness of interpretations which can be made on the basis of test interpretation" (p.14).

There is clearly a lack of generalisability of interpretations that can be made about a performance if the test is on the too-specific end of the specificity continuum, a situation that can be envisioned as potentially problematic in some contexts. For example, The *Eurocontrol Standard Test in English for Trainee Air Traffic Controllers* is used, as the name suggests, to assess the language of Air Traffic Control (ATC) staff. This is a performance test and the tasks presented to examinees are assumed to mimic real-life ATC language-related tasks. Teasdale (1996) describes this context as being one of the best defined LSP domains. He explains that "there exist recommended phraseology in the form of a treated

reduced code which defines the content, form and ordering of elements of utterances, as well as specifying the circumstances in which specific phrases are to be used" (p.213). Teasdale (1996) reports on the research conducted in the process of developing this test from the domain specification to the needs analyses conducted using corpus-based approaches and criterion-referenced approaches, to test design. The data gathered for the linguistic analysis were recordings of over twelve hours of uninterrupted transmissions from seven airports. Data were categorised into series that would comprise either *General Language Functions* or *Specific Language Functions.* Although the research documents the process of attempting to tie real-life tasks with test objectives, there are areas of concern, one of which is sampling of the domain. Teasdale (1996) acknowledges that since the corpus was restricted to single utterances, it is possible that "discoursal characteristics may not be properly accounted for" (p.224). This may be viewed as problematic when larger segments of speech require interpretation from a wider context. One would hope that the novice ATC has strong levels of general proficiency to be able to reconstruct the language s/he is faced with, which may not have been sampled in the tasks selected for the test. Conversely, if the larger segment of speech requires a narrower interpretation than the ATC is using, s/he may over generalise.

Task-based second language performance assessments are primarily resting on the premise that language proficiency is made up of both knowledge and ability for use. Skehan's (1998) expanded model of oral test performance, itself an

adaptation of the Kenyon-McNamara model of oral test performance (1995), re-introduces the concept of ability for use as "drawing upon dual-coding capacities and organizes the way processing is adapted to performance conditions" and adds to the model the notion that "tasks themselves are susceptible to finer-grained analysis, such that an effective distinction can be made between task characteristics and task implementation conditions" (Skehan, 1998, p. 171). In his view, the inferences that are to be made about the performance on task-based language assessments must relate to the underlying ability, and therefore must sample broadly. He states that "[t]he basis for sampling would then not be (or not simply be) an abilities model, but instead a processing framework which would provide a more robust basis for generalizing to a range of performance conditions, as well as a surer basis for establishing construct validity. This is in contrast with the approach taken by Brown, Hudson, Norris and Bonk (2002) who view the accomplishment of the task being central to task-based assessment. The following is the Long and Norris (2000) definition of task-based assessment, to which they subscribe:

> Task-based language assessment takes the task itself as the fundamental unit of analysis motivating item selection, test instrument construction, and the rating of task performance. Task-based assessment does not simply utilize the *real-world* task as a means for eliciting particular components of the language system, which are then measured or evaluated; on the contrary, the construct of interest in task-based assessment is performance of the task itself (Brown et al., 2002, p.9).

Literature in LT reflects concerns regarding the validity of task-based approaches to assessment, and especially regarding the above-mentioned view. Messick

(1994) in particular voices concerns over the inferences that are made from task-based assessments as representations of knowledge. He contrasts between task-driven performance assessments (reflecting accomplishment of behavioural tasks) and construct-driven performance assessments (reflecting competencies and skills underlying the performance). The view is that in the performance assessment of competencies "where the performance is the vehicle not the target of assessment" (p.14), one cannot ignore issues of replicability and generalisability, because "they establish boundaries on the meaning of the scores and on how consistent that meaning is likely to be" Messick, 1994, p.15). His point is that construct-relevant knowledge and skills are what should be driving performance assessments instead of domain-relevant tasks and performance, and points to two fundamental threats to construct validity in performance assessment, that of construct under-representation and construct-irrelevant variance (Messick, 2004). This view is supported by Bachman (2002) who argues that "because of the complexity and diversity of tasks in most 'real-life' domains, the evidence of content relevance and representativeness that is required to support the use of test scores for prediction is extremely difficult to provide" (Bachman, 2002, p.453)

2.1.2.5 The Alternative Assessment Era

Although proficiency is the ultimate goal of foreign language learning, "one must realize that it is not possible to reach proficiency without a series of carefully structured steps which precede it, i.e. the achievement component of knowing a language" (Shohamy, 1996, p.153). Shohamy (1996) states that achievement

testing, as is done in school taught language programmes, also provides valuable information on the language ability of speakers. She advances that in an alternative assessment framework, proficiency testing as well as achievement testing contribute to painting a fuller, more complete picture of the language knowledge a learner has acquired, seeing as "no one procedure can capture the complex phenomena of language knowledge" (p.153). She states that the tendency in recent years has been to incorporate portfolios, self-assessments, peer assessments, interviews and observations, for example, into an assessment battery, especially in school contexts where progress is important to map out. There have been concerns regarding the efficacy of this type of assessment in an institutionalised context, as well as other concerns pointing to issues of reliability and validity, and pertaining to the development, the scoring, and inferences or extrapolations made from such instruments (Kane, Crooks, and Cohen, 1999).

It must be also stated that there are other perspectives as well: language can also be conceived as being much more than underlying traits as in the cognitive view of Skehan (1998); it is viewed by some as a socially acquired and acted activity, as in a sociocultural view of language; "[t]he current formulation of sociocultural theory is activity theory" (Lantoff, 2000, as cited in Luoma, 2004, p.102). In this view, the focus is not on the individual, but on the activity. McNamara (1996) argues that the co-construction of meaning that takes place in the interactions between examinee and interlocutor has neither been conceptualized nor modelled satisfactorily in LT. In a sociocultural view, any interaction is considered to be a

joint action governed by cultural norms, therefore it is perceived that "a speaker speaks with internalised 'voices of others'", building on the view that "language is culturally mediated and learned through experiences with others in direct contact or indirectly, via reading, television and film" (Luoma, 2004, p.103). Taking a more sociocultural view of testing would imply offering more choices to examinees, such as choice of topic, or of texts, for example. In other, more traditional testing systems examinees "may be asked to simulate some real language-use situation, such as giving directions about how to drive somewhere, or treating a patient, but since they know they are in a test, this imposes another set of expectations and norms on the communication" (Luoma, 2004, p.103). She argues that to be fair to examinees, instructions about the kinds of strategies that 'earn them good scores' should be made explicit.

2.1.3 Summary

The fact that there are so many views on the approach to language testing, on the constructs that are of interest, and the means by which we can measure language knowledge, proficiency, competence, performance, ability for use and/or all of the above, points to the complexity of the endeavour on the one hand, and our lack of understanding of it on the other. Some views are complementary, while others are conflicting. It is clear, according to the literature reviewed, that there is no single approach to testing which can be said to be the best or most reliable and valid. As Luoma (2004) remarks, theoretical models provide a very abstract framework from which to build a testing system. What is important to keep in

mind is that the particular context and purpose of assessment should drive the definition of the construct, and the application of the testing system.

In sum, going back to the context of interest for this research, as was stated above, the perspective taken in most governments, as a solution to the performance and/or proficiency view is a requirement for language-general tests to provide information on a candidate's linguistic skill, as it fulfills an operational need for 'transferability'. That is, the government needs "to know that an employee's linguistic ability does not depend on a specific context or job" (Lowe 1985, p.17). This is understandable in light of the military's typical three-year posting duration. In the military's case, "speech samples have to be expanded and supplemented by speech production on non-job-specific, non-interest areas to ensure sustained creativity across a large number of subject areas in order to obtain a ratable sample of speech adequate for a language-general test" (Lowe 1985).

Having said that, it must also be acknowledged that LSP is alive and well in the Canadian military realm, particularly in foreign languages where such skills as translation, transcription and interpretation are concerned. Performance tests, which measure specific skills, for which some military members are trained, are administered once proficiency tests have confirmed a necessary threshold. However, test results are not matched up to the STANAG or ILR scales, and are not reported as a profile; they are job-specific, performance tests, and are reported

as pass/fail (CFLS Foreign Language Testing Directive, 2001). They do not provide generalisations to other contexts.

2.2 Rater Variance in Assessment

2.2.1 Introduction

As described in the previous section, the assessment of second language speaking proficiency involves appraising an elicited speech sample, usually obtained in a simulated context, and matching it against set criteria. Speaking tests require examinees to produce complex language responses, integrating not only various skills, but comprising a multitude of factors and aspects related to language competence. Bachman (1990) reiterates the notion that the purpose of language tests is to be able "to make inferences about one or more components of an individual's communicative language ability. A major concern in the design and development of language tests therefore is to minimize the effects of test method, personal attributes that are not part of language ability, and random factors on test performance. Similarly, the interpretation and use of language test scores must be appropriately tempered by our estimates of the extent to which these scores reflect factors other than the language abilities we want to measure" (p.166). The first two broad categories of test method and personal attributes mentioned by Bachman are also referred to as systematic effects in the sense that they exert the same influence upon scores from one administration to the next. The third category, the unpredictable random factors, is said to be unsystematic, as the

name would suggest, in that they may be linked to temporal conditions of emotional states, testing environment or administrative fluctuations.

Since the present research focuses on ratings, on raters and on rater/scale interaction, the following section of the literature review Chapter focuses on rater variance, and studies looking at rater background, rater training and rater perceptions will be examined in more detail.

2.2.2 Interviewer Variation

As is the common practice in the testing of speaking, prior to the assessment, an interviewer will have elicited a speech sample. Brown (2003) conducted research in interviewer variation by doing discourse analysis of two interviews involving the same candidate being tested by two different testers. She looked at the way the conversations were structured; at the techniques interviewers employ and found interesting results on the impact that the interviewer has on the sample of speech produced. She compared the two interviewers, whose differing interviewing styles affected the rating the examinees received from the eight participating raters. The first interviewer posed more open-ended questions, used follow-up questions, and scaffolded the examinee's comprehension of what was being asked by making use of conversational strategies, which sent explicit cues enabling her to participate fully in the exchange. With the second interviewer, who was not as engaged (as one would expect from an interviewer), questions were not efficiently posed, and therefore did not effectively elicit much speech

from the examinee. This interviewer did not follow patterns of an elicitation-response chain whereby the examinee's responses are followed-up on, and frequently left the examinee wondering what was expected of her. While this examinee was reported as being communicative and forthcoming with information when interviewed by the first interviewer, she was also perceived as reluctant and uncooperative when interviewed by the second. Although this is a small-scale study, these findings have important implications for considering rater bias in that the source of the bias found here was generated by the interviewers, not by the raters. As Brown points out, the issue at play in this context revolves around fairness. She also points out that "interviewer training has generally tended to be somewhat overlooked in relation to rater training, with interviewer behaviour rarely being scrutinised once initial training is completed" (p.19) and cautions test administrators to ensure that these differences in elicitation styles are not inadvertently affecting the construct by introducing varying degrees of difficulty depending on who does the interviewing.

Ross and Berwick (1992) also investigated interviewer variation by focusing on the role of *accommodation* during oral proficiency tests, and on the issue of the OPI as a sample of extended discourse. The interviewer usually initiates topics, and leads or directs the conversation in a way that is not likely to be found in 'real-life' communications. It is also known that native speakers generally use accommodative speech when speaking with non-natives, in order to facilitate communicative with 'foreigners'. Ross and Berwick (1992) therefore set out to

investigate empirically, "the way OPI interviewers accommodate to their interlocutor, how this accommodation reflects both conversational interaction and interview structure, and the uses of accommodative behaviour as a potential alternative source of rating criteria" (p.163). Fifteen audiotapes of OPIs at varying levels of proficiency were used in the study, and were analysed looking at accommodation exponents. The researchers found ten prominent ones that were deemed representative of non-interview 'foreigner talk' interactions. Findings indicated that interviewers are making use of these conversational features in order to enable them to develop an "emerging image of the interviewee's current level of proficiency" and it may be a "potentially useful metric of proficiency" (p.170). Of particular relevance in the context of this study is the relationship between accommodation and rating. What was apparent and was expected is that accommodation exponents were used more frequently with lower proficiency examinees in the 1/1+ range of proficiency on the American Council on the Teaching of Foreign Languages (ACTFL) scale. What also emerged however is that overaccomodation of some examinees also took place, especially with examinees at the 2/2+ level of proficiency, which would suggest that with some examinees more 'foreigner talk' took place than was needed. This can be perceived as a potential rating problem and in commenting on overaccomodation, Ross and Berwick caution that "certain guidelines to performance within the OPI may not be especially well learned or well understood" (p.169) and the ratings assigned may actually be a reflection of the poor quality of some of the interviews,

and interviewers. Standardization of interviewing techniques may diminish variability in accommodation practices.

Lazaraton (1996a) also used a qualitative research format, specifically conversation analysis to investigate one aspect of interviewer-candidate interaction, namely, interlocutor support, during a face-to-face oral interaction test, the Cambridge Assessment of Spoken English (CASE). As reported above, Ross and Berwick 1992 have shown that participants in OPI interviews do engage in 'conversation-like' behaviours such as accommodation, repetition, etc., which can be viewed as adding to construct validity evidence, but Lazaraton adds that "on the other hand, interviewer speech modification, unless systematic and consistent, adds an element of uncontrolled variability to the assessment process" (p.154). Her study identified at least eight types of interlocutor support used by the ten trained examiners who took part in the research, which are said to also be "supportive practices found in other forms of native speaker - native speaker and native speaker - non-native speaker interactions" (p.166). What is critical in her findings, however, is that this interlocutor support was not found to be consistent, and therefore may afford some candidates opportunities to demonstrate ability with some interviewers, opportunities not necessarily given to all. She concludes by cautioning that what is unclear at this stage, and which seems to have been relatively understudied, is whether this has any impact on ratings in that it may lead to unequal or unfair and biased ratings.

Researchers have also wondered if interviewer variance could be in part attributed to gendered differences in communicative style (O'Loughlin 2002). In testing systems where the interviewer is also the rater, it is therefore possible that gender will be a source of bias in assessment of performance. As Sunderland (1995, cited in O'Loughlin) suggests, "the behaviour of interviewers of either gender may vary according to whether they are paired with a male or a female candidate. In both cases, it is feasible that the gendered behaviour of the interviewer will influence the outcome of the test by either strengthening or undermining the candidate's performance" (p.171). In order to investigate whether or not there are any effects of gendered differences between interviewers or more importantly, if they have any effect upon test scores, O'Loughlin examined the audio-taped performances of eight female and eight male test-takers who each undertook separate practice IELTS tests (International English Language Testing System), one with a male interviewer and one with a female interviewer. A subsequent analysis of each of the 32 interviews, which had been rated by four raters (two of each gender), was performed using multi-faceted Rasch analysis. O'Loughlin focused the discourse analysis on the participants' use of three conversational features: that of overlaps, interruptions and minimal responses which appeared to be "highly 'gendered' in spoken interaction" according to research reviewed by Coates (1993, as cited by O'Loughlin p.175). Although findings indicated that overall, male candidates used fewer overlaps than female candidates, high degrees of variability in the use of overlaps within the paired ratings did not reveal a clear gendered pattern of use. There were also few interruptions found during the

analysis of the interviews, (possibly due to the fact that an oral speaking test cannot be construed as being an 'ordinary' conversation), therefore this feature did not yield any gendered pattern of use. Even though there were many more instances of the third feature, minimal responses, this once again yielded no clear pattern of gendered use. Test score analyses revealed that there were no significant differences in rating based on gender with these raters. Interestingly, O'Loughlin found that "both male and female participants indicated their ability to make supportive contributions to the interviews through their use of positive overlaps and minimal responses in particular. A collaborative style is clearly not exclusively the province of female speakers in the testing context" (p.198). O'Loughlin acknowledges that it is possible that not using the interviewer as the original rater may have masked a gendered rating effect, or that using video recordings as opposed to audio recordings might have produced different results.

Shohamy (1983) examined the stability of oral proficiency assessment by introducing variables into the test situation. To investigate both validity and reliability of the oral interaction, four separate administrations of oral tests were conducted on students of Hebrew as a foreign language in a university context, where differences in the testing occasion, the tester, the speech style and the topic were introduced as variables to see if they had any effect on the assessment result. The data indicated high levels of agreement between the two testers, and more importantly, indicated that scores on the test did not fluctuate due to test occasion or tester. Significant differences however were found when the assessment

changed speech style, from the interview style to the reporting style, on different topics, indicating that these are factors influencing scores and stated that "different speech styles may imply different aspects of oral proficiency" (p.537). These findings have enormous implications for testing systems where the testing protocol, tasks and choice of topics are not prescribed, but left up to the interviewer based on the proficiency first hypothesised, or other test formats. As Shohamy pointed out "there is a need for drawing stringent guidelines and instructions about the procedure and content of the test and the tester to assure uniformity and maximum consistency of the assessment" (p.538) since these have the potential to skew interviewer practice and influence test scores.

2.2.3 Rater/Scale Interaction

McNamara (1996) considers the many ways in which raters may be different from one another and says that 'raters may display particular patterns of harshness or leniency in relation to only one group of candidates, not others, or in relation to particular tasks, not others (p.123), a finding supported by Fox (2003). Most performance-based assessments evaluate various aspects of the communicative ability within the same task such as fluency, accuracy and cohesion. Raters may be evaluating these various aspects differently in that they may be harsher or more lenient towards one and not the other aspect. McNamara's 1990 research on how trained raters of the performance-based Occupational English Test (OET) interpreted the rating scales for which they had been trained to use when assessing speaking and writing of health professionals, demonstrated that "the raters, in

each case, were overwhelmingly influenced in their judgement of the candidates by their impressions of the candidates' grammatical accuracy" (as cited in McNamara 1996, p.216).

Systematic effects on the rating of speaking were researched by Upshur and Turner (1999), through the development of a framework of test taking and test scoring of the speaking ability of Grade 6 learners of English. They reported on incidental findings relating to the rating of the two tests, Audio-Pal and Story Retell, and found that the scoring of one of the two tasks, that of Story Retell (whereby students were asked to retell a story which is known to them and to the raters) showed that raters were unexpectedly severe or lenient. Upshur and Turner found that the raters employed differing rating strategies for rating this task in that they brought their own knowledge of the story into the mix, or the expectations they had of how the story went. This allowed them to better understand what the student was intending to communicate as opposed to only rating that which was produced by the examinee. The Audio-Pal task, on the other hand, had no pre-conceived outcome, and the raters therefore only assessed that which was being communicated. This would indicate that pre-conceived knowledge of the content by the raters or knowledge of the communicative intent of the students bore weight in the assessment, a source of bias that had not been predicted. Interestingly, Upshur and Turner (1999) also found in their study that teachers who have been involved in rating scale development rated differently than those who had not been involved, and were the most severe raters of the

group. The raters who had not been involved in scale development appeared to be more lenient, and although they have no explanation for this finding, it seems to indicate that raters tend to be more lenient when rating instructions or scale descriptors are less well understood or internalized.

It has been demonstrated by LT researchers that not only does rater background affect test results but also the test method. An interesting study by Chalhoub-Deville (1995) looks at how both test method and rater groups affected second language test scores given to college-level learners of Arabic as a foreign language, tested with three different instruments: an oral interview, a narration and a read-aloud. The raters formed three different native speaker groups, including teachers of Arabic as a foreign language living in the USA, a non-teaching native speaker group also living in the USA (having been there for at least one year), and a non-teaching native speaker group not living in the USA but in Lebanon. The purpose of the study was to derive the criteria underlying the oral ability scores attributed to the eighteen speech samples rated by all three groups to delineate the dimensions that raters considered when rating the subjects' overall second language ability. In particular, the researcher was interested in finding out the relative weights of those dimensions for each of the three groups of raters, and deemed as most appropriate in this case, the use of multidimensional scaling that would "account for individual differences in the perceptual or cognitive processes that generate the ratings" (p.20). Based on regression analyses, three dimensions emerged from the data: the first was

'grammar-pronunciation', the second was 'creativity in presenting information' and the third was 'amount of detail provided'. Not all of the dimensions were as evidently applicable to each test method, but results indicated that the students performed differently across the three tasks. More interestingly for the context of this study is Chalhoub-Deville's reported differences in the ways the three rater groups emphasized the three dimensions in assessing the students' overall ability; a way that is not consistent with other research, (such as McNamara 1990, cited above) and that can possibly be explained by the fact that other studies have predominantly looked at the rating of the English language, and not at Arabic, or in this case more precisely, Modern Standard Arabic (MSA). Her results showed "that the group of non-teaching Arabs in the USA emphasized all three dimensions in their ratings, although dimension three, 'amount of detail provided', had the most salience...that raters in the teaching group seemed to be relying most heavily on the dimension two 'creativity in presenting information', and that the group of non-teaching Arabs residing in Lebanon, however, emphasized almost solely dimension one, i.e. 'grammar-pronunciation'" (p.25). Teachers have been reported in previous research by McNamara (1996) as generally being more focused on grammatical aspects of assessment whereas non-teachers have been reported as being more concerned with communicative aspects of the language (Chalhoub-Deville 1995). This study demonstrates the degree of difference that exists when various raters' backgrounds, a range of perceptions of what is necessary to communicate effectively in diverse contexts and various tasks interact in the rating of performance assessment. Chalhoub-Deville points

to the question of who should be used as the rater criterion in second language oral assessment and if a variety of rater groups should be used in different testing situations since the choice of rater and their perceptions may influence scores.

Insight into understanding the effect of rater background as a potential source of bias in assessment of performance on a speaking test has also been contributed by Brown (1995). Her research examined the occupational and linguistic backgrounds of raters for the Japanese Language Test for Tour Guides. She compared the assessment of 51 examinees made by 33 assessors who had either experience teaching the language or had worked as Japanese tour guides, and who were native and near-native speakers of Japanese. The data was analysed using an extended Rasch model, including three facets in this case – the candidates, the raters and the assessment criteria, in order to look at an array of rater behaviour. In terms of intra-rater reliability, non-native speakers were the most consistent group. Brown explains that "it may be then that non-native speakers are less adventurous in their assessments and more constrained by the criteria provided, whereas native speakers are more likely to be influenced by an intuitive feeling which is not captured in the descriptors; they interpret the descriptors in a more nuanced way" (p.6). This finding is very interesting in the context of using non-native raters; contexts that are not frequently found in institutionalised testing systems, but are more frequent in testing systems such as the NATO context, of interest for this study. Brown also found that there were significant differences between individual raters in terms of their harshness, however minimal

differences were found regarding native vs. non-native speakers in terms of harshness or leniency, (although the native speaker group was found to be harsher), and raters with industry background were found to be harsher than those with a teaching background, (although these differences were also minimal and deemed insignificant). Therefore, there is no evidence to suggest that non-native speakers are any less suitable than native speakers, or that a teaching background is preferable, if raters are provided with acceptable training and unambiguous assessment criteria, in an occupational assessment setting. Perhaps the most interesting finding of Brown's research points to the way in which the different groups of raters perceived the assessment criteria, and the way in which they applied the scale. There were different perceptions of the various language features, and different perceptions of the tasks: teachers were harsher in evaluating grammar, vocabulary and fluency than industry raters, while the latter group marked pronunciation more harshly (p.9). Brown offers the possible explanation that teachers are accustomed to students' language and therefore may not perceive the non-native Japanese pronunciation as impeding communication as much as non-teachers, who tended to rate language more globally, not focusing on linguistic factors as much as teachers do. In terms of linguistic background, non-natives tended to be substantially harsher towards the politeness and pronunciation factor, which is explained by Brown as being the result of having learned this complex procedure. She states "non-natives, by virtue of having systematically built up their proficiency, may have different perceptions from the native speaker, the ultimate judge of the performance in the target situation"

(p.14). This is true of most foreign language contexts; however, it may be a debatable point in a context where a language has *lingua franca* or international status, such as English, which is moving to an international language (Llurda, 2004). Brown concludes by stating that "raters appear to have inbuilt perceptions of what is acceptable to them and these perceptions are formed to some extent by their previous experience" (p.13). These findings are consistent with what Upshur and Turner (1999) observed during the Story Retell task in their ESL testing context mentioned above whereby raters, despite the training received, brought in their own knowledge of the intended communication into the rating and therefore, created a previously unforeseen source of bias.

Kondo-Brown's (2002) investigated rater severity, rater-candidate interaction and rater-category interaction. Her study identified significantly biased rater-candidate interactions in the assessment of Japanese second language writing using a modified version of an existing rating scale for norm-referenced placement decisions. Although the raters were said to have demonstrated high levels of inter and intra-rater reliability, it was discovered that there were significant differences in overall severity. Some candidates and criteria were rated more harshly and leniently overall, just as in Lumley and McNamara (1993) and in Weigle (1998). Her bias analysis also showed that 'all raters had their own unique bias patterns suggesting that an individual rater's severity or leniency pattern can be very complicated and variable' (p.24). One of the noted overall patterns of bias found involved ratings of candidates in the extreme ranges, that is,

with very high and very low proficiency. Kondo-Brown suggests that the criteria at the ends of the scale may need to be more explicit and that further training of raters at these points might be needed, as well, it would seem to indicate that multiple ratings are most favourable in this context.

These findings are also in accordance with Kenyon's (2000) study of ratings of oral proficiency interviews (OPIs) at lower levels on the comparison of high stakes direct and semi-direct tests of German in a university setting. In order to study the comparability of the scores, Kenyon conducted Generalisability (G) studies of the two sets of scores to determine the amount of score variance due to "consistent differences between the students, consistent differences between the raters, and an interaction between the students and the raters" (p.91). G theory is an appropriate methodology for examining rater behaviour especially when it is important to ensure rater agreement (p.91). Although this analysis did not specifically include rater variance as a factor, Kenyon found that some of the descriptors in the lower levels of the ACTFL scale were not "sufficiently well defined to support consistent interpretations by all raters" and that "reliability increases greatly as the number of raters is increased"...therefore showing that "double ratings and arbitration appears to be crucial when greater numbers of testers and raters are involved" (p.98).

It is worth noting that it has been generally accepted in large institutionalised testing systems that rater agreement serves as evidence of reliability of assessment

and therefore that rater disagreement would be an indication of an unreliable evaluation (Henning 1996). Henning calls attention to the fact that it is common practice to resolve disagreement by bringing in a third rater to referee when agreement between two raters fails, or to average the two scores thereby coming to a 'consensus', which would be deemed more reliable than the score of any one rater. In his study of simulated performance ratings on a six-point scale by two independent raters, Henning (1996) endeavours to account for non-systematic error in performance ratings, and points out that there could potentially be situations where raters are in perfect agreement on the score, but that both could be wrong in their judgement by either rating above or below the true ability of the examinee (p.54). His study "examined the assumption that rater agreement, as reflected in the correlation coefficient or other statistical indices of covariance, is an adequate indication of score reliability, and considered whether rater discrepancy is an appropriate indication for the need for adjudication of scores"(p.60). Henning concedes that situations where the two raters misjudge true ability in the same direction and magnitude represents a small proportion of non-systematic error, and that it is not of great measurement concern. Nonetheless, despite the fact that in many cases correlation coefficients provide a needed indication of rater agreement, it is possible that rater agreement may be used as a score reliability indicator in circumstances where variance potentially should be explored, or that adjudication may not be sought where it may be needed. Again, it seems that sources of bias sometimes emerge unpredictably and

that testers and administrators should be aware of the potential that 'assumed' good practices may have as impact on test scores.

Fox (2003) takes a different approach to look at whether there is a significant source of bias in the trialling of a new version of a high-stakes, fully integrated[16] topic based language test, the CAEL (Canadian Assessment of English Language). Using an 'ecological approach' (Bronfenbrenner 1979) [17] to look at the interrelationship between raters and test-takers, Fox found that bias was evident in raters' scoring of the written task on the new science version. This source of bias would probably not have been noticed had more traditional methods of investigation been used exclusively, namely pre-identified groups such as gender, language groups, etc. It was discovered that the reading and listening texts presented in the new version of the test led the raters to perceive a 'correct answer' to the writing task that had been designed to measure argumentative ability. Instead of rating the texts produced based purely on the intended development of argumentation (the given criteria), it was found that the raters were more often than not, giving students (specifically those at key cut-offs/borderline), the 'benefit-of-the-doubt' and giving them the passing grade, when their argumentation followed the perceived 'correct answer' formula. Fox (2003) states that "the careful and systematic elicitation of the perceptions and responses of raters and test-takers, as co-participants in the testing process, helped

[16] The CAEL is a thematic test, whereby a single topic is introduced in the listening, reading and writing sections (Fox et al., 1993).
[17] From an ecological perspective, test takers and raters are regarded as interconnected and as interrelated participants in the testing process (Fox 2003, p.22).

these test developers identify key differences between versions of a test, potential sources of bias, and limitations with regard to generalization on the basis of test performance" (p.40). What is particularly interesting here, and relates to findings from Upshur and Turner (1999) above, is that traditional approaches to data analysis in LT research sometimes fail to identify bias, especially when the sources of bias are identified *a priori* instead of emerging experimentally from the data gathered.

2.2.4 Rater Training

One of the common ways to reduce variance among raters is to ensure that the raters adhere to the rating process. Rater training is regarded as one of the best ways to reduce error in judgements among raters, and aims to reduce noted patterns of overall severity or leniency, and randomness. During rater training, raters are usually required to rate a selection of texts, be they spoken or written samples, and to measure them against particular rating criteria. Some of the samples will be identified at borderline, and others will represent the full range of abilities that are characteristic of the scale or criteria used. Ideally, raters will first rate independently and then see how their colleagues rated the same samples. Over the training period, it is expected that raters' assessment will be in line with the common interpretation of the descriptors by the other members of the team (Lumley and McNamara 1993). In reviewing related research, Lumley and McNamara (1993) state that rater training can reduce, but not eliminate rater variability because "judges often sense that they have unique standards, and it is

hard for them to alter their standards" therefore "compensation for rater characteristics needs to be built into the rating process"(p.57). Rater training is said to be most useful in identifying rater self-consistency. This is said to be crucial because if raters do not demonstrate strong intra-rater reliability, it is impossible to compensate for their randomness. Using multifaceted Rasch measurements implemented through the programme FACETS, Lumley and McNamara (1993) investigated the consistency of rater characteristics over time, and what implications this may have for rater training. The test they used was the performance-based Occupational English Test (OET). Assessment in this case was either carried out 'live' by a trained rater/interlocutor, or a trained rater using a recording of the test. In this study, two rater training sessions took place, 18 months apart from one another, and then an operational test administration took place approximately 2 months after the second training session. The tapes rated were identical for two of the three rating sessions. With regard to rater training, findings indicated that there were large differences in rater behaviour from the second to the third rating, which Lumley and McNamara say "suggest that the results of training may not endure for long after a training session...providing support for the practice of holding a moderation session before each test administration, to allow raters to re-establish an internalized set of criteria for their ratings" (p.69). These findings have important implications for testing systems where raters' services are not used for long periods of time, or only used sporadically. It would indicate that raters not only need to retrain regularly, but

that it calls into question "the practice...of certifying raters and then basing judgments of candidates on single ratings by such certified raters" (p.69).

It is established that rater training is an intrinsic part of ensuring rater reliability in testing systems. McNamara (1996) has suggested that raters will vary in terms of leniency or harshness overall, may react more harshly or leniently towards some candidates and not others, to certain tasks, in how they use certain criteria, and also in how consistently they exhibit these characteristics. Thus the need to provide training in the hopes of minimizing these potential differences, or at least to try and control them, has been seen as an important part of the testing process. In order to study the effect of training on raters, Wigglesworth (1993) investigated the impact of giving specific feedback to raters regarding the results of bias analyses, which took into account their individual performances in rating, to see if it would affect subsequent ratings. Eight raters were given analyses or 'assessment maps' which aimed to identify systematic sub patterns of behaviour which may occur with respect to some aspect of the testing situation (p.309). Following an individualised feedback session, there was a second rating, and a second bias analysis was done for each of the eight raters, "in order to determine whether similar or different patterns emerged for each rater" (p.310). Wigglesworth found that in general, not only were raters responsive to the feedback they had been given but that they had incorporated this feedback into the next testing situation whereby demonstrating a reduction of bias in aspects which had previously been evident. Although Wigglesworth acknowledges that her data

set was small, there is evidence that it is beneficial to track rater characteristics and incorporate them in rater training or (re-training) sessions regularly. As Wigglesworth concludes, what is not clear is whether or not these adjustments on the part of the raters have long lasting effects, or if raters revert to previous patterns after time has lapsed.

Weigle (1998) also studied the effect of training on raters, in comparing the effect of training on new raters pre and post training in terms of impact on rater severity and consistency. Eight 'new' and 'old' raters were asked to rate 60 English language essays used for placement purposes, in an academic context. As was hoped, her analysis showed that there were changes in rater behaviour post training. Pre-training, the new raters were more severe than the old raters, a surprising finding, and after receiving training, the new raters became more like the old raters in terms of rater harshness, although it is pointed out that training did not eliminate differences in raters in terms of severity altogether. Intra-rater reliability was improved post training which is a desirable effect, but not all raters demonstrated marked improvement which may be due to the training in that it may not have focused enough on what these particular raters needed, or as Weigle explains, it may be "that certain raters may never be trained out of inconsistent rating patterns and perhaps should not be used as raters" (p.280), which is also consistent with conclusions reached by McNamara (1996). In looking at group differences, it appears to be that training reduced the extreme severity that was present in the new raters as these raters tended to apply criteria more rigidly than

did the old raters. This would suggest that training eliminated extreme ratings in the new raters, and made them more self-consistent, however, significant differences in severity still persisted overall. Rater training does not make all raters rate in the same way, but did make each more reliable. It is the practice in this sphere, as in many other contexts reviewed previously, to have essays marked by two raters and to average out the two scores unless there are great discrepancies, in which case a third rater's result would be averaged out with the closer of the previous two ratings. If the goal of the rater training and the norming sessions is *not* to eliminate rater differences altogether (a perhaps futile expectation) but to increase reliability, then as Weigle stated, it "will presumably make examinee measurement more accurate, as predictable variations in severity among raters can be modelled and compensated for mathematically" (p.281).

2.2.5 The Native Speaker-Rater

There has been much discussion in the testing literature on the use of the native speaker either as a reference in rating scales (McNamara 1996 or as a judge in rating performance (Brown 1995). As mentioned above, Brown (1995) concluded that there were no significant differences in the ratings of Japanese writing by native and non-native judges, in terms of rater harshness, but that there existed variance in the way they used the assessment criteria. In that study, test performance was judged on a scale with reference to the native speaker as criteria (as stated in Hill 1997). Hill (1997) also explored the use of native vs. non-native raters as a potential source of bias, but in the context of a test of English as an

international language in South East Asia. The purpose of this study was to investigate whether native (in this case Australian) and non-native (in this case Indonesian) speaker raters rated writing performance in English differently, and to investigate the suitability of using non-native speakers as raters in the specific context of assessing local teachers of English. Data for the study came from the extended writing task on the reading/writing sub-test of a specific purpose test designed to assess English language proficiency as relevant to classroom teachers in this context (p.276). The study used 13 non-native English lecturers and 10 experienced English-speaking raters. Both groups were trained prior to rating and "were asked not to employ the idealised native speaker as a reference point...and to think of what would constitute a good model for the purposes of teaching English in Indonesian high schools, where the majority of students would ultimately use English to communicate with other Asians" (p.280). Data were analysed using FACETS, and the facets investigated were candidate, rater, item and type, to determine whether the two groups of raters were comparable in terms of consistency, harshness, and in their use of the assessment criteria (p.282). Regarding harshness as a group, Hill found that there were greater differences among the Australians in terms of their harshness than within the Indonesian group, which is surprising considering that the latter group was inexperienced. These findings are also in accordance with Brown's (1995) study; however, contrary to Brown's study, Hill noted that overall, the native speaker group tended to be harsher. One important finding was that the Indonesian group was more reluctant to use the top of the scale than the Australians were. The descriptors at

the top clearly avoided any reference to the native speaker ideal, but despite this fact, it is possible according to Hill, "that the Indonesians may still have been unconsciously applying a native speaker standard" (p.285), which would imply a bias toward the most proficient examinees that were encountered by this group. Overall, the findings of this study, not unlike Brown's (1995) study, would indicate that the non-native raters are no less suitable than the native speakers raters.

To sum up the discussion on the use of the native speaker as criterion in rating scales, it is said that the concept of 'native speakerness' itself as a target is flawed since native speaker competence is not homogeneous (Hamilton et al. 1993, as cited in North 2000), and that both native and non-native competence vary in relation not only to the communicative task, but in many ways according to the cognitive complexity and familiarity of the task, the educational experience, social class, culture or sub-culture, etc., of the user. Second language acquisition studies show that learners are not on a path to native-speakerness, but rather on a path to a level of functional ability that they would find satisfactory, and which would serve their professional, academic or occupational purposes effectively. Seeing as non-native speakers of English currently outnumber native speakers, some have argued that it is the native speaker who will one day have to "learn the conventions of English as an International Language (EIL), in order to communicate successfully with the larger community of English language speakers (Llurda, 2004, p.320). Others do not share this view such as Davies,

who considers the native speaker "a concept which is, in spite of its fuzziness, essential" (Davies 1989 p.158, as cited by North 2000). He recognises that when discussing meaning of the phrase 'language ability' among native speakers, we consider some native speakers to be better speakers than others, and that when we refer to non-native speakers, we use the term 'language ability' in a way which always implies a comparison to native speakers in deciding an acceptable level of performance at a particular level, or standard. In this view, any judgment of accuracy, socio-linguistic appropriacy, socio-cultural discursive conventions etc, can only be made by reference to the norms of the native speaker culture, or in the case of English, to the native speaker cultures. The obvious questions that come up are Whose English? Who's cultural setting? To further complicate the issue, there are competing perspectives. On the one hand, the International English (IE) movement views that the only acceptable norms are those of native English speakers, and on the other hand, the World Englishes (WE) movement views that to impose IE on users of WEs may be discriminatory against non-native English speakers (Davies, 2003). However, it can be argued that Davies' line of reasoning regarding the need for a native speaker reference holds for the use of English as a *lingua franca*, and arguably also holds for its use in assessment: when third parties use a language as a *lingua franca*, some native speaker point of reference is still necessary to ensure mutual intelligibility.

2.2.6 Paralinguistic Features in Language Assessment

And finally, another area of interest for researchers looking into rater variance and bias involves a relatively under-researched aspect related to non-verbal communication. A study by Jenkins and Parra (2003) investigated the role of nonverbal behaviour during an oral proficiency interview in the context of assessing the speaking ability of international teaching assistants (ITA) with either Chinese or Spanish as native languages, assessed by North American judges. In that study, analysis of the videotaped interviews showed that nonverbal competence assumed an important role with regard to the assessment of the 'lower' proficiency candidates, and that those test-takers who had employed conversational cues in ways that demonstrate that they are on equal footing with the interviewers were perceived as being more 'proficient' than they actually were. These candidates were said to have framed the interview as a negotiated conversation as opposed to the other candidates who framed the conversation as an examination. This was not a factor in the assessment of more proficient candidates. Interestingly, the effect of modifying the power dominance of the interviewer into more of a peer-based conversation, in this case, biased the interviewers favourably. It appears to be the case that unwittingly, these candidates gave the test a more natural, authentic flavour, where the control of the conversation did not rest exclusively with the interlocutor. It may be that "the personalities and genders of the ITAs may have influenced their approaches and responses to the interview format in subtle and unknown ways" (p103), however, it cannot be denied that this effect or variance in rating from one candidate to the

next is problematic with regard to validity. The question to ask in this case is whether or not students should be instructed to be less passive during interviews and be taught to assert their linguistic needs, such as requesting clarification.

Pollitt and Murray (1996) also found evidence of non-linguistic 'interference' in their compared-pair assessment study. The probe was primarily intended to elicit underlying constructs of proficiency to develop a rating scale that would represent the perceptions of proficiency, as seen by raters while in the act of judging students' performance, and look at 'normal rating behaviour', not sources of bias. However, they acknowledge that the raters mentioned being influenced to some extent by the candidates' personalities, physical attractiveness, nationalities and cultural background, and some of the raters "at times, appeared to be utilizing a *synthetic* process in which a holistic image of the speaker is formed derived primarily from the individual's (rater's) preconceived, that is, preconstructed, understanding of language learners...just as in meeting a stranger at a social event, a comprehensive image of the person is evoked by a few first impression" (p.86). The authors pose a difficult, albeit necessary question in light of the paralinguistic features that accompany speech: "Should personality, manner, facial expression or the impression of friendliness be considered a valid part of oral proficiency? Or perhaps we should narrow the construct down to being simply a test of the ability to produce grammatically well formed and meaningful utterances under a real time constraint?" (p.89). This question points to a lack of formal theory to support assessment practices.

McNamara (1996) acknowledges the need for underlying theory in the construct of speaking performance to shed better light on paralinguistic influence in assessment. He calls for the need of "theoretical models which acknowledge the role of non-language specific cognitive and affective variables in language performance settings...in order for us to make sense of research on performance testing and to provide a general framework within which explicit hypotheses can be formulated about the relationship between candidate and rater behaviour and test scores" (p.166).

2.2.7 Summary

Upshur and Turner (1999) state that the language testing approach to viewing the effect that the rater has on assessment has mostly focused on the score attributed. They suggest, "the rater is not only an additional source of measurement error but, as a methods facet, may also exert systematic – although unwanted – effects upon scores" (p.87). Having said that, it must be noted that Upshur and Turner (1999) say that "there is yet no theory of method to explain how particular aspects of method have systematic effects upon discourse that are reflected in test scores" (p.91), "nor is there a developed explanation of how rater and examinee characteristics interact with one another and with discourse characteristics to yield ratings, or how tasks relate to well-functioning rating scales" (p.106). The findings of their research as well as Fox (2003) would indicate that there are a number of relationships and interactions that happen in a performance testing situation that have either been overlooked by the traditional research methods, and

that should be further addressed in order to have a greater understanding of the factors that interplay in the testing of language ability. Not only are these research endeavours crucial from a psychometric perspective, but they also represent important issues of fairness, ethical practice and embody concerns that have begun to be addressed more frequently by researchers.

The first section of the literature review highlighted the various approaches, constructs and questions related to language testing, with particular emphasis on testing speaking. Studies described in this section looked at how and where rater variance may be manifested, at the severity and leniency of raters, and at the training raters receive. Some have investigated if the effects of training are put into practice, and others have taken a closer look at the raters and their background, to name but a few of the facets explored, in order to provide future links to the findings of the research, presented in Chapter 4. These studies are relevant to the research question driving this study, namely: How comparable and consistent are ratings across NATO raters and NATO countries?

Chapter 4 will report on findings related to this question, as well as a variety of related sub-research questions. Chapter 3 will now present the methodology employed in the research.

CHAPTER THREE: METHODOLOGY

3.1 Overview

Chapter 1 of this thesis described the NATO language context, and presented issues of relevance to the context. This Chapter will detail the methodology employed in the study by describing the participants, the instruments used to collect the data, the procedures followed and the analysis phase of the enquiry. Using both qualitative and quantitative methodologies, this study primarily explored the ratings that 103 participants, assigned to two Oral Proficiency Interviews (OPI) samples, 'A' and 'B' measured against the NATO STANAG 6001 Scale of Level Descriptors, Edition 2, to see if there were significant differences in ratings among NATO raters, and across NATO countries. Data were collected by survey because of the vast geographical area covered by this research.

3.2 Participants

Since NATO, PfP and aspiring countries' military members must have Standardized Language Profiles (SLPs) using the NATO STANAG 6001 scale of language proficiency descriptors, as criteria, the research focused on English language testers who test English as a foreign language for their respective Ministries of Defence, and/or military language institutions, either full-time or part-time. Since this is primarily a rater comparability study, and participants

were not required to test the speaking skill, participants hereinafter will be referred to as raters.

Eighteen countries provided participants, as well as two NATO units, for a total of 103 participants, from 20 countries.[18] The data obtained are considered a representative sample of the raters and countries within NATO, with participants from Continental Europe, Central Europe, Eastern Europe, Northern Europe and North America, several being long-standing NATO members, others newly included, and one non-NATO country who holds observer status within BILC.

Participating rater ratios varied from one country to the next. Many of the participating countries have small testing programmes with very few raters, from 1 to 12, for example. Ratios in these smaller countries ranged from 66% to 100% participation, with more than a third of countries providing nearly 100% of available[19] raters. A couple of countries have larger testing systems, from 50 to 75 raters and participation ratio in these ranged from 15% to 40%.

Parallel to data collection among BILC member countries, a small number of Carleton University graduate students undertaking a M.A. in Linguistics and Applied Language Studies, and enrolled in a language testing course at the time of the research, were used as a control group (CG). The purpose of using a CG was

[18] The two participating NATO units hereinafter will be included in the term 'country' so as not to single them out.
[19] In some cases, raters were either on leave, on holiday or on re-assignment at the time of data collection.

to triangulate the data by investigating how teachers, who are neither trained as testers nor raters, rated the two OPIs and interacted with the scale. These findings would enable comparison of their scores with scores from the lesser-trained raters within the NATO countries. The group was comprised of a professor with extensive testing experience and 4 graduate students with minimal testing knowledge and/or very little to no practical testing experience. None of the CG participants had ever heard of the STANAG (apart from the professor) and all (except for the professor) were non-native speakers of English. Although these participants were given a 30-minute briefing on the STANAG scale *a priori*, they can be considered 'naïve' raters, since the briefing given summarized the conceptual framework of the scale, explained briefly the essence of each level, but did not train these participants to interpret the descriptors or statements or to rate the OPIs.

3.3 Instrumentation

3.3.1 Oral Proficiency Interview Samples

As described in Chapter 2 of this thesis, the Canadian OPIs at the Canadian Forces Language School (CFLS) are general proficiency, multi-level speaking tests, similar to the well-known American OPI used with the ILR scale at the Defence Language Institute (DLI). Similarities pertain to levels, tasks or functions at each level, content/context areas, and accuracy statements, although the CFLS OPI uses the STANAG instead of the ILR. Differences between the

two are mostly administrative: only one interviewer conducts the test at the CFLS, it is usually only rated by the interlocutor (as opposed to being independently rated by the 2 testers), and it also makes use of a significantly different evaluation grid. As mentioned before, the CFLS does not presently assign 'plus' levels to OPIs[20]. Probes to higher levels are used to ensure that the tester is indeed at the appropriate working level[21]. Two live OPI samples were used in this study. Sample 'A' is a level 1 sample (30 minutes long) while sample 'B' is a level 2 sample (43 minutes long). Both were deemed to be ratable samples and representative of the testing practice that takes place at the CFLS in St-Jean, Canada.

3.3.2 Questionnaires

Two separate questionnaires were designed[22]. The first aimed to gather information on the rating population in the various countries. Quantitative data pertaining to age, mother tongue, number of tests conducted, years of experience, etc., was collected by means of yes/no answers and choice answers. This questionnaire also included open-ended questions to gather qualitative data on the participants' tester training and STANAG training received, and on their views of the ease of use and application of STANAG 6001 scale. The questions related to their **tester** training and experience, as opposed to their **rater** training and experience, since this was considered the most adequate means of eliciting the

[20] In the Canadian system, pluses are viewed as being incorporated within the range of the level in question, that is, a level 1+ would be within the level 1 range, for example
[21] More information on the OPI can be found in Appendix B.
[22] See Questionnaire on Rater Data in Appendix C and Questionnaire on the Rating of the samples in Appendix D.

required information. The second questionnaire in two forms, A and B, was designed to accompany the rating of the two OPI samples. Quantitative data related to the rating (level) given and whether a plus level would have been assigned (had it been available to them) was also collected. Finally, participants were asked to explain the rating process they followed. A copy of the STANAG 6001 Edition 2 speaking descriptors was also provided on the CD raters received.

Gender variations in interviewing style, in interlocutor responses and ratings were not considered in this study since both interviewers and both examinees were male. Questions pertaining to the gender of participants were not included in the rater data questionnaire. Data pertaining to the quality of the OPI, and the testers' interviewing techniques were neither requested nor collected.

Questions in both questionnaires were worded in as neutral a manner as possible to elicit as much information from the participants without imposing a perspective, introducing too much of a rating course of action, or leading them to any particular level. It was hypothesized that if raters were not familiar with the OPI format or its rating protocols, it might be difficult for them to provide information on particular features of speech, despite the fact that most scales of language proficiency are comprised of various statements referring to similar features of speech.

The questionnaires had been piloted beforehand on an expert tester/rater that was familiar with the testing process at the Canadian Forces Language School, who had been a tester trainer for many years and was also a facilitator at the BILC sponsored Language Testing Seminar in Garmisch-Partenkirchen, Germany.

3.4 Procedure

BILC member representatives from countries attending the BILC Professional Seminar held in October 2005 in Sofia, Bulgaria, were approached to see if their testing teams would be interested in participating in a research project. Interest in the study was high and many of the representatives agreed to see if the matter could be pursued upon return to their respective countries. They were informed that in return for their participation, nations could request to know where they had placed in the coded data; but they would not be given any information regarding the other participating nations[23]. In December 2005, a follow-up letter of information detailing the purpose and the procedure for participating nations was sent via electronic mail to twenty-four countries.

During January and February 2006, packages were mailed to the twenty countries that confirmed participation. Each country identified a contact person who would receive and distribute the envelopes to the raters, as well as subsequently collect them after the two-week time frame. After the period was up, the envelopes were

[23] A coded summary report of the findings will be made available to all participating nations. The codes protect a country's anonymity.

placed into a return package and mailed back to the researcher. In some countries, raters are located at various sites therefore some contacts requested and received extra time to ensure all raters had the allocated time to complete the survey. In order to ensure that all of the participants were volunteers, each individual envelope contained a letter explaining the rationale and procedure for the research as well as stated explicitly that should someone decide not to participate, they were simply to take the CD (or cassette when requested) out of the envelope, seal the envelope, and return it to the person who had handed it to them, within the allocated two-week timeframe. The letter also stated that the identity as well as the country of origin would be kept anonymous since all data would be numerically coded for reporting[24].

Rater participation time was estimated at 2.5 hours, but many raters mentioned that it had taken them longer. A randomly chosen country number was assigned to each country package received. Afterwards, each participant was assigned a rater number.

Raters were instructed to listen to each OPI, (as many times as they wished), to fill a rating grid or assessment sheet from their institution and to make use of the rating protocol employed in their own testing system. The rating protocol was purposely left up to the raters in order not to impose an unknown rating system. The instructions to raters emphasized that they should work individually and not

[24] The Carleton University Ethics Review Committee granted approval for this study in December 2005.

consult with colleagues about the OPIs and the ratings they assigned. Although in some testing systems raters rate in pairs, or as a committee, it was anticipated that if raters within countries 'team' scored the OPIs, the data would not reflect if and/or what differences in rater perceptions existed within each country.

Participants were ultimately asked to insert into the envelope the rating grid or evaluation form that they had used, so that information regarding the various factors assessed could be analysed.

3.5 Analysis

3.5.1 Oral Proficiency Interviews

Two OPIs were selected from among the 28 samples submitted to the researcher by the CFLS Head of Evaluation. OPIs chosen had to meet the following criteria. First, interviews with candidates who were members of the countries participating in the research were excluded, in case it introduced bias into the rating. Raters might know, recognise and/or feel awkward rating a candidate from their own country or military establishment. Second, many of the OPIs submitted were beyond the recommended time limit of 30 minutes, therefore most of these were also excluded.

Since most training and testing programmes only deal with levels 1 to 3, it was hypothesized that some non-native raters might have difficulty rating higher-level

examinees. Clearly, achieving 'near-native' proficiency such as is defined by a level 4 on this scale, is an infrequent training goal.

The samples of speech were chosen not only because they represented two different levels on the NATO STANAG scale, more importantly, it was perceived that they illustrated the notion of range within each level. Sample 'A' was rated 'level 1' by the original rater at the CFLS. The rater's evaluation grid indicated that the examinee in sample 'A' was a 'strong level 1' and the grid reflected that the examinee was high up in the range in all of the rating components with one component peaking over the threshold of the next higher level of the grid. The rater's comment supported the rating assigned:

> Candidate lacks the tense control (esp. past) and fluency required at the next level. *(Original CFLS Rater, Sample 'A')*

Where this examinee was positioned in the scale was deemed of interest to the study for the following reasons: first, although within the Canadian system the sample would be considered a base level[25], it was this rating within the range which warranted exploration from rater to rater, and from country to country, for sample 'A'. Because of the high position of this examinee within the range, it was hypothesized that he may be placed in the next higher level by some raters; therefore how frequently this would occur remained to be investigated.

[25] The levels, from 0 to 5 on the STANAG scale are considered 'base' levels, as opposed to 'plus' levels. The CFLS does not presently use 'plus' levels for rating productive skills (speaking and writing) in its STANAG testing system.

Sample 'B' was rated as a level 2 by the original rater at the CFLS, and the visual profile on the evaluation grid, as well as the rater's comments indicated that he was a solid level 2 speaker, but not peaking into the level 3 range. Again, the rater's comments substantiated the rating assigned:

> This candidate can speak with great ease at the paragraph level. The accuracy at L3, however, is not evident. He can use some L3 structures (e.g. hypothesis), but the accuracy, and flexible use of more complex vocabulary and structures is not yet developed. A good level 2. *(Original CFLS Rater, Sample 'B')*

Although clearly not a borderline level 2, it was hypothesized that this level might be problematical to rate from country to country; as the levels of proficiency increase, it becomes more difficult and arguably, important, to interpret the scale in similar ways, not only from one rater to the next, but from one country to the next. Level 2 is referred to as 'Limited Working Proficiency' in the STANAG scale[26], while level 3 is considered to be 'Minimum Professional Proficiency' and these labels, as explained in Chapter 1, are somewhat misleading. Nonetheless, levels 2 and 3 could be viewed as being critical levels at NATO, since they are often the objective to meet, as a result of training courses, Force Goals, or Partnership Goals (NATO Partnership Goal (Example) PG G 0355, Language Requirements, 2004). Therefore, it was deemed of interest to see how frequently the level 2 sample B would be rated as such, and if not, the reasons raters would give for placing this examinee in another level.

[26] The STANAG 6001 scale can be found in Appendix A.

3.5.1.1 True Scores

There are two sources of evidence confirming the levels assigned to the chosen samples. First, the scores were considered the true scores, or 'correct' scores, because these SLPs were returned to the originating countries as the official profile received in Canada. Upon completion of their English language course at the CFLS, the examinees in both samples were each given a certificate which included SLP results from four proficiency STANAG-based tests (in listening comprehension, speaking, reading comprehension and writing). Second, these scores represent the mean, mode and median score of the participating raters. These levels or scores will be referred to as the benchmark levels throughout Chapter 4, and will be discussed thoroughly.

3.5.2 Rater Data Questionnaire

This section will explain the analysis of each type of question in the questionnaires in relation to the scores awarded by the raters. Data from the questionnaire gathering information on participating raters were first coded for frequencies. All data were entered in the Statistical Package for the Social Sciences (SPSS) version 13.0. The data from the control group were placed in a separate SPSS file.

Questions 1-14 were in multiple-choice or yes/no format, depending on the nature of the information being elicited. These questions referred to the first language of the participant, the education level, the years of experience as a tester, the number

of tests conducted, whether they test full-time or part-time, etc. Non-parametric tests were performed on the data, crosstabbed with ratings, to see if and how they impacted the ratings given.

Questions 15 to18 elicited open-ended answers. For question 15, participants were asked to provide information on the tester[27] training they received in terms of the type of training, the duration, the number of practice tests, etc. In order to restrict the training categories, and since training reported took a variety of forms, data were first put into four categories depending on what was reported by the participants. A number of participants stated outright that they had never been trained as testers (although most of these had experience[28]), and hence were not trained raters. Therefore the first category, **'None'**, included those who said that they had received no training, or that their only training had been during their university teacher training courses (as it is generally the case that teachers are not trained to become testers/raters *per se*).

Others mentioned that they had received some on-the-job training (OJT); either had observed other testers, had viewed and/or rated videotaped performances, or had performed some practice tests.

[27] Again, the questionnaire did not specifically ask participants to detail the training received as raters, since it was hypothesized that in most countries, testers are also the raters.
[28] Question 12 of the Rater Data questionnaire, collected information regarding years of experience as a tester.

Some participants indicated that they had attended one or more seminars; or mentioned having taken a two-week OPI training session, or attended workshops such as those organised by the British Council's Peacekeeping English Project (PEP) or had attended the BILC's Language Testing Seminar (LTS) in Garmisch-Partenkirchen, to name but a few of the training listed by participants.

Some of the participants had theoretical knowledge, having either completed a Masters Programme in Language Testing, or had attended a university course in language testing. However, it was not specifically stated if they had received tester/rater training *per se* during their programme.

Many had been trained in various combinations of the above list. Therefore, those who mentioned having received training in one of the above were coded as **'ONE'**. Those who mentioned two of the above list (i.e. OJT and the LTS) were coded as **'TWO'**. The last category, **'THREE or more'**, included the remainder of participants who mentioned three or more items from the list such as seminars and workshops, OJT, as well as tester training at university, or had attended the BILC sponsored Language Testing Seminar, and/or PEP seminars, etc.

Participants were then asked in question 16 to provide information specifically on the training received on the interpretation of the STANAG scale. Many participants indicated that some of the training listed in question 15 had focused on the STANAG, but others stated the opposite. Some participants reported that

seminars had focused on one level and skill at a time (e.g. testing level 2 writing), and others that their institute conducted frequent norming sessions, where the raters' interpretation of statements in the scale were discussed, and/or 'benchmark' interviews were listened to and analysed to promote a 'standardised' view. It is interesting to note that a number of participants indicated explicitly that apart from having 'seen' the STANAG document or done a 'thorough' reading of the descriptors, no training had been given regarding the interpretation of the scale. Again, since training received took on a variety of forms, data were collapsed into the same four categories, as in question 15.

Since these data questions were the first to be coded, it was decided that all of the data should be coded a second time to ensure that they were coded reliably and that the categories held up over time, and coding experience. Questions 15 and 16 were re-coded using the same criteria 3 months after the initial coding took place. Since key findings related to these data, it was deemed important to ensure consistent coding. Therefore, a full data set of both questions 15 and 16, as well as the coding scheme was given to a second researcher for coding. Correlation between the two researchers was $r =$. 93 for question 15, and $r =$. 89 for question 16, based on Spearman's rho, a satisfactory indication that the data were reliably coded.

In order to perform non-parametric tests using SPSS, to identify chi-square indices, and to test for significance, the four above mentioned categories were then collapsed into two broad groups: the first 'NONE' and 'ONE' (little training),

and the second 'TWO' and 'THREE or more' (considerable training). This was done for both questions 15 and 16. In tests of significance, the Alpha level was set at $p < .05$ for the Pearson chi-square tests performed.

The next two questions on the rater data questionnaire related to the STANAG scale itself. Participants were asked to state in question 17, whether or not they found the scale easy to use and apply, and to mention what aspects of its interpretation/application they found to be the most challenging. At question 18, they were asked to make any other relevant comment pertaining to the STANAG scale. These comments were gathered, compiled into lists and summarized in order to report on the most frequent entries.

3.5.3 Samples A and B Questionnaires

Analysis of questionnaire responses on the ratings given to the two samples was first done in a three-step process.

Step 1: Question 1 on each questionnaire asked participants to write their score for the respective sample. Scores given to sample A, and to sample B, were entered in SPSS as 'Score OPI Sample A' or 'Score OPI Sample B'.

Step 2: Question 14 of the questionnaire on rater data had asked participants to state whether their STANAG testing system presently uses 'plus levels', and 25% of respondents answered yes. Seeing as it was known to the researcher *a priori*

that pluses were in use in some countries, once they had completed the reporting of their rating of the samples, respondents were asked at Question 9:

> 9. If your system uses 'plus levels' explain if you think this performance would qualify for a plus rating, or not. Explain. (If you do not use pluses, write *not applicable* in the following space).

And at Question 10:

> 10. If you do not use 'plus levels', state whether you think this performance would be rated at a plus level, according to your present understanding of plus levels. Explain your reasons.

The intention was not to impose any particular procedure for using pluses, or to provide any framework for their interpretation, but to see if the participating raters considered these samples to be indicative of a plus performance, however it was defined by them. Some participants had initially given pluses in their ratings at Question 1, others indicated at Question 9 or 10 that the use of the plus would be valuable to rating the samples, and others indicated that no pluses should be assigned to the samples. Since plus levels are considered to be within the range of the levels[29], all of these second scores were then coded and entered as 'Adjusted + range A', and 'Adjusted + range B' in SPSS.

Step 3: Lastly, since it was identified that these samples had a 'true score' both by the initial rater, and by the overall mean score of the raters in the study, each of the data sets were re-coded in SPSS as either a 'correct' or 'incorrect' score for tests of significance.

---

[29] In the Canadian perspective, that is.

3.5.3.1 Ratings Comparisons

A critical step in the analysis of scores assigned by the participants involved imposing an interval scale over the ordinal scores of the ratings, in order to look at variability. Ratings of level 1, 2, 3, and 4 were entered as 1.00, 2.00, 3.00 and 4.00. Ratings of level 1 'plus', level 2 'plus', and level 3 'plus' were entered as 1.60, 2.60, and 3.60 respectively. This provided a means to investigate degrees of difference in scoring, by allowing calculations of overall rater means, overall country means and standard deviations (SD)[30].

3.5.3.2 Country-to-Country Ratings Comparisons

Investigation of the dispersion of ratings of the two OPI samples from country to country was done, as well as an exploration of the dispersion of intra-country ratings, to see how the ratings of the OPIs compared from country to country, and if there were differences in scores within the same country. This allowed to see if within-country differences were any more significant than the ones that were present rater to rater, and country-to-country. The mean of all initial ratings for each OPI as well as the standard deviation (SD) were calculated. Each country's mean was calculated for the **initial** scoring of both samples A and B. The calculation of Zed scores[31] was also done in order to place all ratings on equal footing, and to provide comparisons between countries.

---

[30] The best overall indicator of dispersion of scores is the standard deviation, defined as 'a sort of average of the differences of all scores from the mean' (Brown 1988, cited from Brown, 2002, p.131).

[31] To do this, first we see what the difference is between a country's mean and the mean of all countries, and then this difference is divided by the overall SD.

3.5.3.3 Rating Process

Question 2 of each questionnaire asked: "How did you arrive at this rating? Explain the steps you took". This aimed to discover how participants went about rating the candidates in each of the samples. Since the process raters followed was similar for both samples, the data gathered from both questionnaires were collapsed under one heading and included: the number of times the participants mentioned having listened to the samples, and whether or not their rating system used percentages, converted scores or a cumulative-type rating system. If, and how frequently raters mentioned having consulted the STANAG scale in the rating process, or a rating grid only, or a combination of both were also noted. Crosstab analyses were applied to all coded data in relation to the scores of the two samples.

3.5.3.4 Rating Factors

The next five questions in each of the questionnaires dealt with the breakdown of the OPI in terms of various rating criteria or factors, that are explicit in the assessment done for OPIs at the CFLS, and that are also part of typical oral examination tester training seminars that some of the participating raters may have been exposed to. The following section explains the analysis that was performed on each question, how the data were grouped and interpreted.

Participants were asked to provide detailed explanations to the following questions for each OPI sample rated. Questions posed were: 'At what level

would you say the examinee performed the tasks or functions very well?' 'Where or during which tasks and functions did you think this examinee did not perform well?' 'What did you pay attention to in terms of the examinee's linguistic strengths and weaknesses?' 'Which discourse and delivery features were, in your opinion, most evident in this sample of speech?'

For each OPI sample, lists of all tasks or functions performed 'well' and 'not well' that the participants attributed to the examinees were compiled, in addition to lists of all linguistic features such as grammar and vocabulary, discourse features of coherence, cohesion and text type, and delivery aspects, comprising fluency and pronunciation, which were then tabulated for frequency. These questions aimed to further breakdown the raters' observation of the examinee's performance on the OPI in order to provide insights into their perceptions of the performances. Responses were then compared to the ratings the participants had given each sample to see if the reported tasks and function performed 'well', for example, were commensurate with the levels at which they rated the samples. And lastly, participants were asked: 'When comparing the performance from the examinee to the STANAG descriptors, where do you think this examinee fits in, in terms of the overall sociolinguistic aspect?' The answers provided to this question were analysed to see not only the level given for this particular aspect, but also to see what participants would state regarding their perception of its relevance to testing English in this context, although this was not made explicit in the formulation of the question.

The last question regarding the rating of the samples pertained to the STANAG scale itself. Question 8 asked participants to provide a glimpse into their interpretation of the scale by forcing them to 'pull it out' (if they hadn't already), and taking a closer look at the wording. It asked: "Are there any of the STANAG descriptors which you feel best exemplify this examinee's performance? State them." In order to code these responses, the STANAG speaking descriptors were themselves coded. Each descriptor or sentence was given a number. Each statement provided by the participants was then matched up to these numbers, which were entered in SPSS. Two separate databases in SPSS were set up to process the two sets of answers (from questionnaires A & B), and entries were analysed for frequency.

A similar analysis was done with the CG data, comparing both the ratings assigned to the two samples and the rating process followed by these participants, with the other participants in the survey. Only the most salient of these findings will be reported.

All data gathered from the questionnaires were tabulated and results are reported in Chapter 4.

CHAPTER FOUR: RESULTS

4.1 Introduction

Chapter 3 explained the methodology employed in collecting, coding and analysing the data provided by the 103 raters who volunteered to rate the two OPI English language samples from the CFLS. Chapter 4 presents the results of the research, reporting on the main findings of data analysed for both sample 'A' and sample 'B'. The overarching research question was: How comparable and consistent are ratings across NATO raters and NATO countries?

Findings are organised and are presented in relation to three specific questions, by:

FINDINGS 1:	Comparing the ratings given by participants, and investigating if there were significant differences between raters and between countries;
FINDINGS 2:	Comparing the ratings awarded by participants with varying degrees of tester training and STANAG scale training, comparing related testing experience, and language background; and
FINDINGS 3:	Comparing the rating processes followed by raters, and reporting on the comments made regarding the scale.

4.2 Findings 1: Comparing Ratings

4.2.1 Rater-to Rater Comparisons

How did ratings of the same oral proficiency interviews compare from rater to rater? Did the use of plus levels increase rater agreement?

Sample 'A': As indicated in Chapter 3, sample A had been rated as a level 1 by the original CFLS rater. Sample A was rated by all 103 participants, with the majority (60), rating within the level 1 range, with 46 participants rating it at level 1 and 14 at level 1+, or 58.3% of participants (as highlighted below), while 40.7 % rated in the level 2 range, and 1% in the level 3 range. (See Table 1, below).

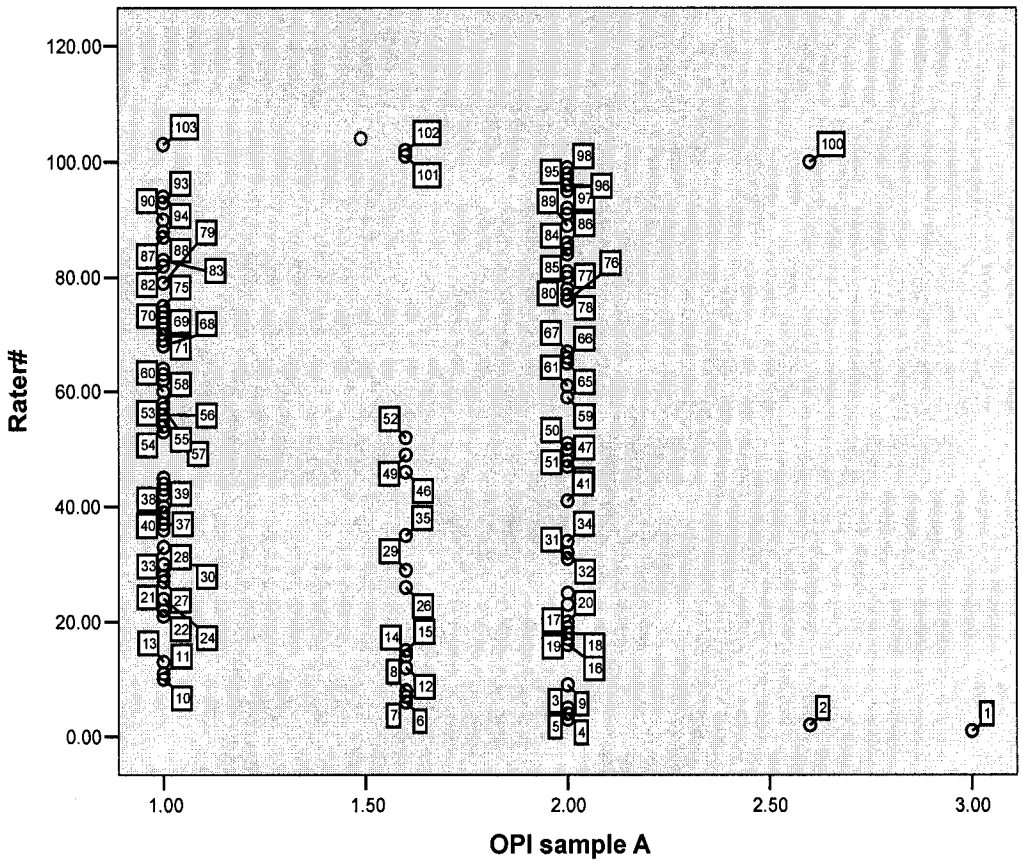
Table 1: Scores OPI Sample A

Levels	Number	Percent
1	46	44.7
1+	14	13.6
2	40	38.8
2+	2	1.9
3	1	1.0
Total	103	100.0

The overall mean was 1.52, the median was 1.60[32] and the mode was 1.00. Since sample A was originally rated as a 'strong level 1', it was expected that a number of participating raters might indicate that a plus level would be appropriate for this candidate, and that some might award level 2. Figure 1 below illustrates where all of the raters are positioned on a scatter plot. The solitary dot represents the mean score of all raters.

[32] The code 1.60 was entered in SPSS to represent ratings of level 1+.

Figure 1: Scatter plot of all Ratings for OPI Sample A



Raters were asked at Question 9 and 10 to state if they thought the performance qualified for a plus rating, (or not), according to their present understanding of plus levels. Table 2 below, shows the shift of ratings.

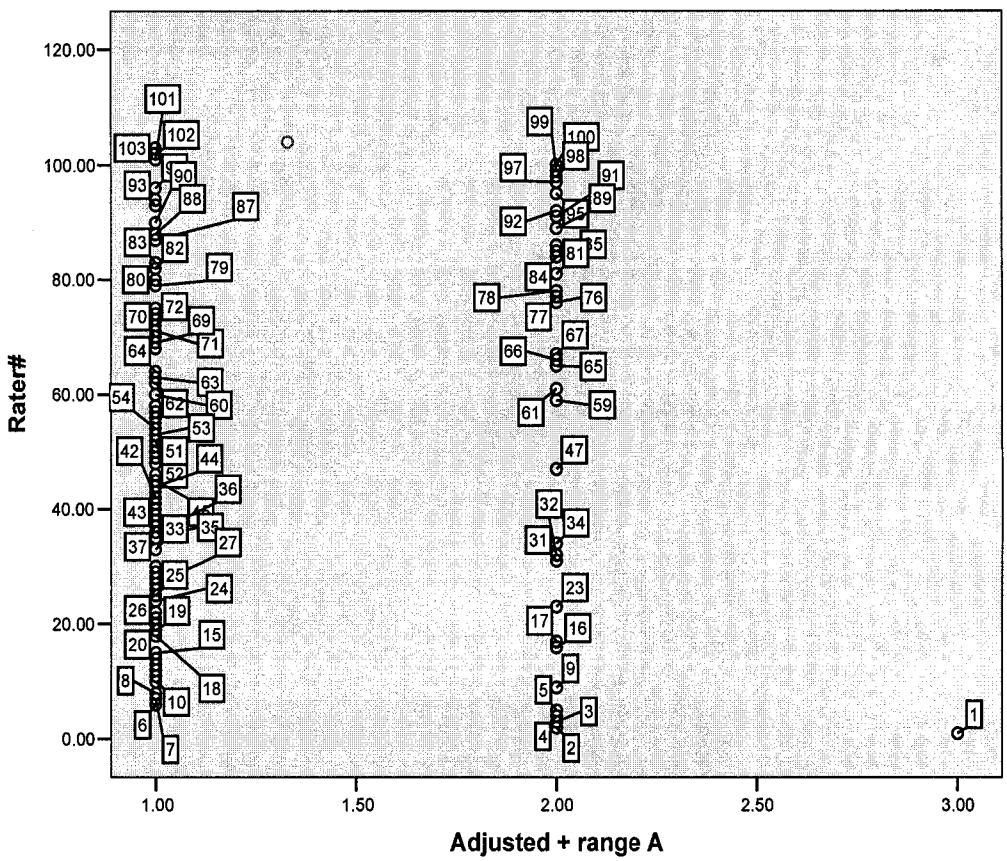
Table 2: Adjusted Scores, Sample A

Levels	Number	Percent
Within Level 1 range	70	68.0
Within Level 2 range	32	31.1
Within Level 3 range	1	1.0
Total	103	100.0

As noted above in Table 2, the numbers changed considerably when the scores were adjusted to reflect the addition of plus levels, with 68 % scoring sample A in

the level 1 range, 31 % in the level 2 range, and 1% in level 3 range. What is most interesting and relevant to note is that 10 raters or 26% of raters who had given a level 2 to sample 'A' initially, if given the opportunity to give a plus, would go down to a level 1+. Therefore with these adjustments, the mean became 1.33, and both the median and the mode were 1.00. Given the true score, the addition of pluses indicated that raters were brought closer to the mean. Figure 2 below illustrates where all of the raters were now positioned in a scatter plot, with the mean, represented by a solitary dot.

Figure 2: Scatter plot of all Adjusted Ratings for OPI Sample A



It appears that raters who perceived the candidate to be close to the next higher level placed him in the next level when pluses were not used. Plus levels offered these raters a scoring possibility closer to the observations made about this candidate's performance, as will be evident in rater comments, provided below. The data also demonstrated that there were raters who initially rated sample A at level 2 who then gave him a level 2+. It can be argued that although the results were now being 'boosted' with the addition of the plus ratings, the scores remained within the range of the 'base' level, and that this was not as 'problematic' as ratings on two different levels.

The responses from the participants must also be viewed in a broader context to see the reasons they gave for their ratings. The qualitative data gathered on the questionnaires allowed for a glimpse into the raters' perceptions and provided the link between the ratings given and the reasons that influenced the ratings. The following examples from the raters' questionnaires illustrate the wide array of opinions regarding the examinee in sample A's proficiency, as do the answers provided to the question asking if they thought the performance qualified for a plus rating (or not), according to their understanding of plus levels. The rating process employed by raters to reach their decision will be reported in Findings 3.

Starting with the lowest ratings within the mean, and moving up the scale with examples from raters who were away from the mean, different 'types' of raters emerged, with the three most frequently occurring: the 'evidence-based' rater

(who supported the rating with observations from the examinee's performance on the test), the 'intuitive' rater (who either did not justify the rating, or did so unconvincingly), and the 'extra-contextual' rater, (who introduced other issues into the rating that are generally considered to be beyond the rater's concerns).

For example, the following participant rated sample A at **level 1**, and according to the comment provided below, s/he gave an 'evidence-based' rating. *Rater 37* had this to say regarding assigning a 'plus' level:

> No, I don't think that Candidate A's performance would be rated at a plus level, although I readily admit that my understanding of plus levels is vague. We don't use them in our system, so I haven't had to concern myself with them. If by a plus level we mean that the candidate peaks up into the next higher level but does not sustain that performance at that level, then I would be hard-pressed to find evidence that Candidate A performed at level 1+. Overall, I feel that he is a level 1 speaker – an eager speaker, perhaps, but still a level 1. *(Rater 37)*

Although *Rater 37* stated having only a vague notion of the use of pluses, the response offered is based on what s/he was able to observe regarding the 'evidence' gathered by the interlocutor.

Rater 37 contrasted with the following rater, who could be called an 'extra-contextual' rater, in the sense that concerns that are outside of the examinee's performance permeated the rating process. *Rater 20* rated sample 'A' at **level 2**, and when asked if a 'plus' rating would be useful in rating this sample, answered:

> Yes! I would be happy to give him a 1+. Since we do not use 'plus levels' I am afraid that rating him as a clear 1 would disadvantage him and, for this reason, I would rather give him a very low 2. *(Rater 20)*

There sometimes appeared to be external factors intervening in the judgments some of the raters made, and as in the case of *Rater 20,* can be termed 'extra-contextual'. The rater's concern for the consequences the result may have on the candidate appears central to the rating assigned. The comment regarding rating him as a 'clear 1' indicated that this is potentially how the rater viewed the performance. Perhaps he was perceived as better than a 'clear' level 1, possibly meaning better than a threshold level 1. Another rater who rated sample A at **level 2** had this to say:

> Based on the system I am used to, I would have rated the speaker in sample A as a "1+" based on his unsystematic control of the irregular past tense. Even if the candidate demonstrates level 2 in all aspects of the OPI, he/she cannot receive a rating of level 2 if he/she does not have systematic control (60% or higher) of the irregular past tense. *(Rater 48)*

The comment regarding the unsystematic use of the past tense was mentioned by the raters seen above, (e.g. the original sample A rater, Rater 37) as the reason this examinee did not qualify for a level 2, and was rated at level 1. *Rater 48* is an 'evidence-based' rater, but s/he placed a lot of emphasis on the grammatical aspect of the performance. Also, this sample certainly could have been rated in the same manner as is done within her/his testing system. However, if it were

perceived that this OPI was different from the OPI in *Rater 48*'s testing system; perhaps it could explain why s/he did not initially use the plus levels[33].

The following rater initially rated sample A at **level 1+**. This comment typifies the 'evidence-based' justifications given for rating this sample at level 1+.

> I think this candidate is not level 1 according to the STANAG 6001 norm, but, in all fields, something is missing to him to reach SLP 2. There are frequent inaccuracies in grammar; the vocabulary is mainly level 1, very simple. The candidate's ability to understand, on the other hand, was at a good level (I mean 2). Those are the reasons for my decision to rate him SLP 1+. *(Rater 14)*

Some of the raters who perceived the examinee in sample A as being above the threshold of level 1 but as 'missing' some of the requirements for the next higher level, made use of the plus levels, even though they might not be used in their own systems. *Rater 18,* who rated sample A at **level 2**, represents another stance, but also an 'extra-contextual' rating:

> If this is rated 1+ it would be unfair to the examinee, because in my opinion, he is closer to level 2 than to level 1. *(Rater 18)*

It became clear that there were varying definitions as to what constitutes a plus level; e.g. how close to the next higher level does the performance have to be? For *Rater 18* and other participants in this research, the issue of fairness was a

[33] At the time of data collection the STANAG 'plus level descriptors' had not formally been approved by the BILC, therefore were not included into the participating raters' envelopes.

recurring theme. Again, another 'extra-contextual' rater who at question 10 lowered the score to level 1+ from the initial **level 2** rating offered this outlook:

> Because the range is large without plus levels, students and/or candidates who undergo months of language training would no doubt be discouraged if the profile they left with was exactly the same they came in with. Plus levels could address issues generated by such situations. *(Rater 41)*

It would appear that some of the raters who had originally rated sample A at level 2 for reasons other than 'language' lowered the score from a level 2 to a 1+ when pluses became an option. Having said that, many of the participants who rated this sample at **level 2**, were certain he was well within that range, yet most did not give him a plus level, increasing to level 2+, because they perceived his language to be too low, such as this rater:

> I don't think he could be rated as ST-2+ because his command of English is not as firm as it should be due to his pronunciation and grammar problems/mistakes. *(Rater 66)*

And finally, one 'intuitive' rater who initially rated sample A at **level 2** opined the following on granting a 2+:

> He is a plus level. He was able to answer all questions and did not need repetition. *(Rater 4)*

This rater did not comment on the level of accuracy or on the quality of the 'answers' to the questions by the sample A examinee, and did not convincingly explain the use of the pluses, nor give justification. However, it must be reiterated that no specific definition was given to the participants regarding the

interpretation or framework of plus levels; therefore raters' answers provided insight into reasons for the ratings. The examples also highlighted some of the concerns that raters had such as the impact ratings have on students, and issues of fairness, which in large testing systems are sometimes considered to be outside the mandate given to testers/raters, but that are important to the practice, and should not be ignored.

Sample 'B': As discussed in Chapter 3, sample B was rated as a level 2 by the CFLS rater. A total of 96 participants out of the 103 raters scored sample B. Some said that they ran out of time, but most did not give an indication as to why they did not rate it. As with the previous OPI sample, there was diversity of scores, but with ratings now spanning more levels. (See Table 3 below):

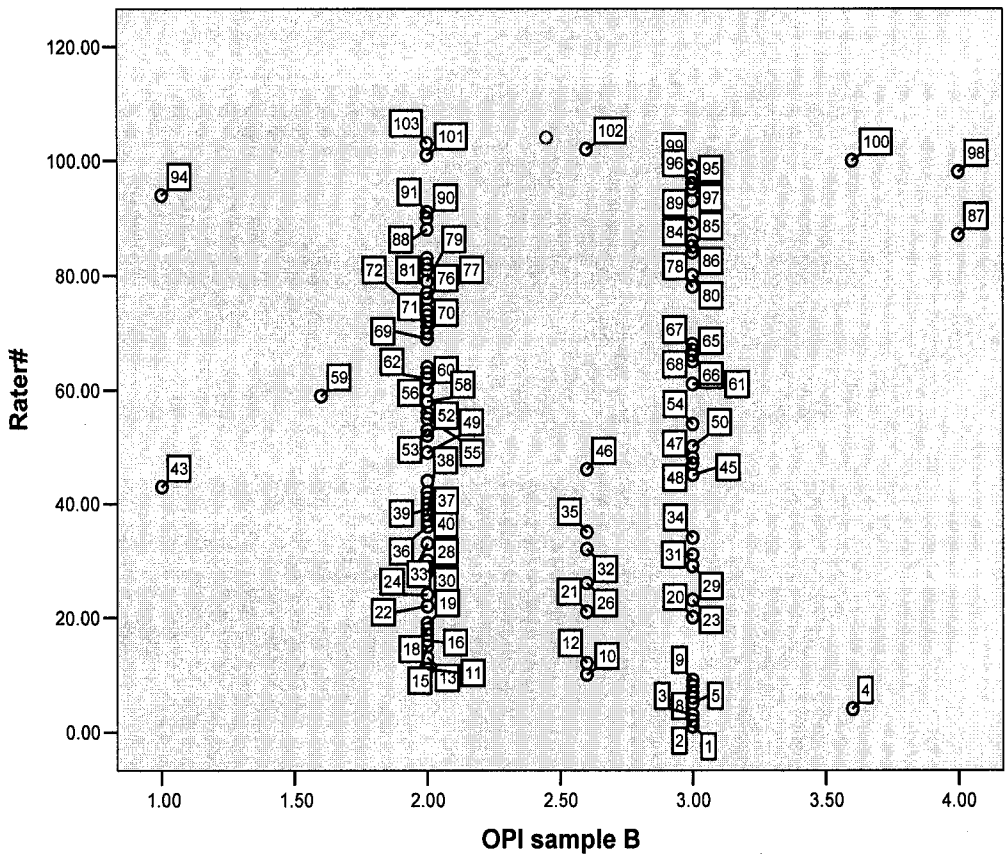
Table 3: Scores OPI Sample B

Levels	Number	Percent
1	2	1.9
1+	1	1.0
2	47	45.6
2+	8	7.8
3	34	33.0
3+	2	1.9
4	2	1.9
Total	96	93.2
Missing	7	6.8
Total	103	100.0

OPI sample B was rated as a 'solid' or mid-level 2 by the original rater. He was not considered to be 'peaking' into the next higher level, contrary to sample A, therefore it was not expected that this sample would receive as many initial plus ratings. 3 outlier raters (or 3.1%) indicated that this sample was in the level 1 range, a total of 55 (or 57.3%) raters placed this sample in the level 2 range

(which is highlighted above), while a total of 36 raters placed sample B within the level 3 range (37.3%). Finally, another 2 outlier raters (2.1%) perceived this examinee to be in the level 4 range. Figure 3 below illustrates all of the participants' ratings in a scatter plot, with the overall mean represented by a solitary dot.

Figure 3: Scatter plot of all ratings for OPI sample B



Both the median and the mode were 2.00, while the mean was 2.45 on sample B. The main question of interest at this stage was whether the inclusion of plus levels, (as indicated by those who would give sample B a plus), would pull the rating away from the mean, or bring it closer to the mean. Table 4 below, demonstrates

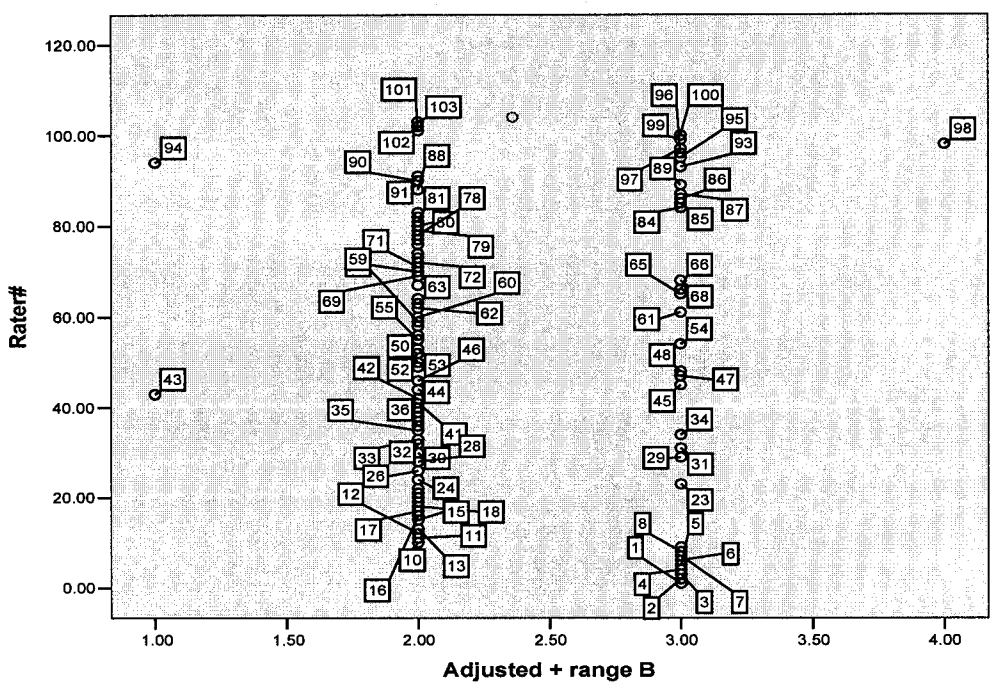
that the percentage of raters rating in the level 2 range only increased from 57.3% to 59.2%, while the ratings in the level 3 range decreased from 37.5% to 31.1%.

Table 4: Adjusted Scores, Sample B

Levels	Number	Percent
Within Level 1 range	2	1.9
Within Level 2 range	61	59.2
Within Level 3 range	32	31.1
Within Level 3 range	1	1.0
Missing	7	6.8
Total	103	100.0

As expected, the ratings awarded remained relatively stable when the notion of plus was introduced for sample B. Though a small change toward the mean, now 2.30 overall, there are almost twice as many raters within the level 2 range than within the level 3 range. The median and mode remained at 2.00. Figure 4 below illustrates the ratings of all participants for sample B when the pluses have been added, with the mean represented by a solitary dot.

Figure 4: Scatter plot of all ratings for adjusted OPI sample



The contrast with Figure 3 is visible, as there were fewer outlier scores in Figure 4. There were 5 raters who initially rated sample B at level 3 who lowered to level 2+, and therefore came within the mean, as well as 1 rater who had rated at level 1+, and moved to level 2.

Raters had differing views on the use of pluses for sample B, with perspectives in relation to the initial score they awarded. For example, awarding the plus, as done by *Rater 94* below, was in relation to the level 1 score initially given. Again, the scope of raters' views provided by the examples of comments, informed on the reasons given for using plus levels with sample B.

Starting with the lowest ratings, one participant who initially rated this OPI sample as a **level 1** due to 'poor grammar' and 'awful' pronunciation wrote:

> Because of good vocab and fluency, 1+. *(Rater 94)*

Rater 94 provided an 'evidence-based' rating, albeit lacking elaboration and with a rating far from the mean. The following two raters, who scored him at **level 2,** provided detailed 'evidence-based' explanations of their view on why this sample should not be given a plus:

> This candidate's performance cannot be rated as 2+. Grammatical/structural control is inadequate and does not rise above (even occasionally) into the upper level. Mispronunciation detracts from the delivery and can be problematic. No evidence of well-controlled but extended discourse. (The interviewer's prompts should not have been necessary). No clear evidence of the use of even some complex structures that might raise the performance to

> the + level. Finally, there is no evidence that the performance occasionally rises and crosses into level 3 (which would indicate a +). *(Rater 36)*

And succinctly put by *Rater 71*:

> No I don't [think it's a plus]. Although, as previously stated, candidate can converse quite extensively, his linguistic skills do not allow him to discuss level three topics using level three structures and vocabulary. *(Rater 71)*

Some of the raters, as mentioned before, raised the score from **level 2** to level 2+, such as can be seen in the next two excerpts:

> Level 2+, yes, it could be rated as a L2+ because the candidate passed 2 or 3 of the probes given. Only his accuracy failed in highly complex structures, as well as some of the pronunciation errors might have distorted the meaning. *(Rater 58)*

> Yes I would award 2+. If you have a look at the attached sheet, you can see that there was substantial evidence of level 3, but the majority of evidence pertained to level 2. The examinee handled functions and vocabulary well. The control of grammatical structures (tenses, plural) was not sustained, even though there were level 3 structures used. *(Rater 53)*

Here again, comments from the two 'evidence-based' raters above indicated that a common understanding of plus levels might enable raters to make comparable judgments. According to the original CFLS tester's evaluation grid, only two probes at level 2 were given to this examinee. In an OPI framework, this would indicate that both of these probes were considered to be unsuccessful by the tester.

It is interesting to see the comments from raters who initially gave sample B **level 3** and then lowered the score to level 2+, due to the wide diversity of reasons given. For example, 'extra-contextual' *Rater 20* below, offered the same perspective for sample B as s/he did for sample A:

> Yes, I think it would be rated at 2+. Since we do not have plus levels, I would rather give the candidate a 'low' 3. *(Rater 20)*

Rater 20 had appealed to the issue of fairness in not wanting to 'disadvantage' the candidate in sample A by awarding a lower score, however, 'low' scores are represented as the same number, and have the same score outcome in the STANAG framework (e.g. a low 3 is a 3). In 'evidence-based' ratings, only the language produced by the examinee is assessed, whereas it seems that in 'extra-contextual' ratings such as this one, the perspective is not about the proficiency the candidate *has*, but rather it becomes about what the rater has *given,* for reasons peripheral to language ability. *Rater 50* provided 'evidence-based' reasons for lowering the score from **level 3** to level 2+:

> This candidate might receive a 2+ based on his wide range of concrete vocab and an admirable amount of abstract language, but perhaps not a 3 due to his pronunciation/accent and stress problems. *(Rater 50)*

Rater 50 seemed somewhat reluctant about staying with the original level 3 rating and considered using the pluses. The following excerpt is another example of the uncertainty and doubt that some raters expressed about their level 3 ratings by the time they got to the question on using a plus level with this sample. *Rater 80* below explained the reasons for doing so, which could be considered outside of

the rating behaviour normally desired from raters, and provided an indication of motives behind some of the ratings:

> Yes, I think it would be appropriate here. In fact, I would tend to give this candidate a 3- (minus) based on his grammar, vocabulary and pronunciation mistakes. If this were not possible, then I would downgrade to a 2+, to indicate his ability, to encourage a further course at level 3. *(Rater 80)*

Concerns that may be viewed as related to achievement on a course, or diagnostic in approach, seemed to be influencing *Rater 80*'s assessment. This rater reported that 'downgrading' to a 2+ would be done in order to 'indicate his ability', which is clearly the primary mandate given to testers/raters.

The majority of raters who placed sample B in **level 3** would not raise his score to level 3+, as indicated by the following rater, but *Rater 89* provided a comment with an interesting twist:

> To rank 3+, the candidate would need much better command of structures. I might put him at 3- (minus) if I listened to the interview again and noted his mistakes. *(Rater 89)*

This comment expresses that this 'intuitive' rater was aware that s/he may have 'accepted' a lot of linguistic inaccuracies that are usually not acceptable at level 3 on the STANAG scale, but s/he did not provide a 'thorough' and convincing rating.

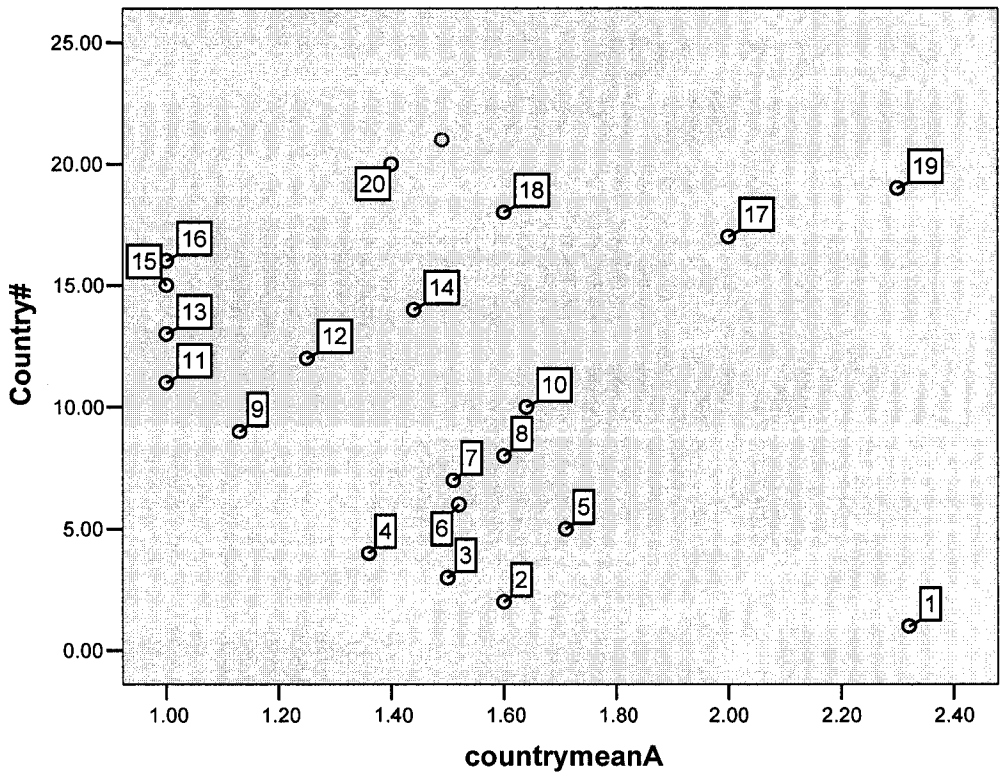
To summarise, the analysis of these data in relation to the first two research questions indicated that a majority of raters placed sample A within the level 1

range and sample B within the level 2 range in their initial ratings, and ratings were in closer agreement when pluses were introduced into the equation, regardless of how pluses were interpreted or applied by the raters. Also, the majority of raters provided evidence-based ratings, a very desirable finding. Nonetheless, there were raters who assigned ratings that were not based on language assessment but that brought external factors into the rating. The rating process participating raters followed will be presented in the section on Findings 3, related to the rating process.

4.2.2 Country-to-Country Comparisons

How did the ratings of the OPIs compare from country to country? Were there differences in scores within the same country? And were these differences more meaningful and/or significant than the ones that were present rater to rater? In this next section, results of analyses of the dispersion of ratings of the two OPI samples from country to country will be presented, in comparison with findings regarding intra-country ratings, as per the methodology explained in Chapter 3. Figure 5 below, represents all participating countries' means, for the initial scoring of sample A. The overall country mean is once again represented by a solitary dot.

Figure 5: All Countries' Means for Sample A



As can be seen in Figure 5 above, and in Figure 6 representing the means of all countries' initial scores on sample B (below), the means are dispersed throughout the diagram.

Figure 6: All Countries' Means for Sample B[34]

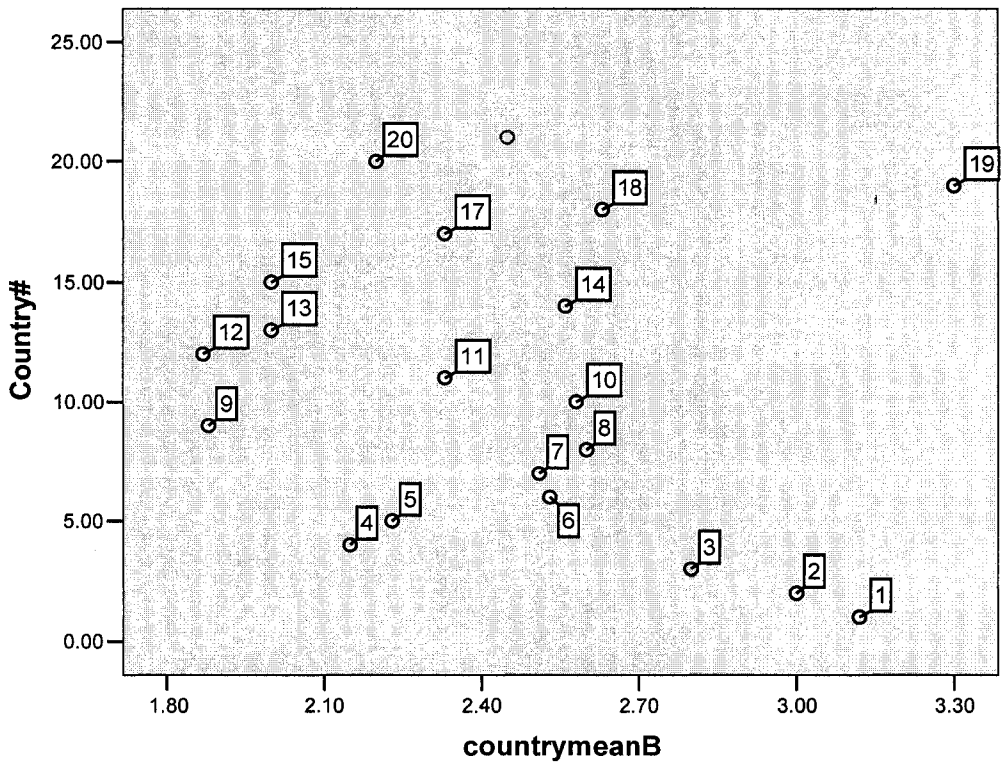


Table 5 below indicates the mean and the standard deviation of each country. Countries that are indicated by an **asterisk** (in Table 5) have a within-country or intra-country SD larger than the overall country SD. For sample A, the overall country SD was .39, and there were 10 out of 17 countries that displayed a larger number. For sample B, the overall country SD was .41, and 8 out of 17 countries were above. In other words, this indicated that scores were more dispersed within those countries than they were overall. The **bolded** SDs indicate countries where there were greater variations on both samples.

[34] Sample B was not rated by country 16, therefore is omitted from the diagram.

Table 5: Comparison of Means and Standard Deviations

	Mean OPI A	Stand. Dev. OPI A	Mean OPI B	Stand. Dev. OPI B
All Countries Together	1.49	.39	2.45	.41
Country Results				
1	2.32	.46*	3.12	.27
2	1.60	.00	3.00	.00
3	1.50	.71*	2.80	.28
4	1.36	.33	2.15	.30
5	1.71	.49*	2.23	.41
6	1.52	**.50***	2.53	**.50***
7	1.51	**.50***	2.51	**.50***
8	1.60	X**	2.60	X**
9	1.13	.35	1.88	.35
10	1.64	**.41***	2.58	**.49***
11	1.00	.00	2.33	.58*
12	1.25	.50*	1.87	.23
13	1.00	X**	2.00	X**
14	1.44	**.53***	2.56	**.53***
15	1.00	.00	2.00	.00
16	1.00	X**	X***	X***
17	2.00	.00	2.33	.58*
18	1.60	**.50***	2.63	**.76***
19	2.30	**.42***	3.30	**.42***
20	1.40	.35	2.20	.35

**The SD could not be calculated for these countries due to small numbers of raters.
***Country 16 did not rate sample B.

In Table 6 below, Zed scores provided another way to compare the data. For example, Country 1, who had a mean score of 2.32 on sample A (in Table 5), which is a difference of 0.83 from the overall mean, had a Zed score of 2.13. With Country 2, who had an overall country mean of 1.60, and a difference of 0.11 from the mean of all countries together, the Zed score was .28 and provided a 'correct' rating, as defined in Chapter 3.

Table 6: Differences between Country Means and Countries

	Mean OPI A	Stand. Dev. OPI A	Mean OPI B	Stand. Dev. OPI B
All Countries Together	1.49	.39	2.45	.41
Countries		ZED SCORES		ZED SCORES
1	0.83	2.13	0.67	1.83
2	0.11	0.28	0.55	1.34
3	0.01	0.03	0.35	0.14
4	-0.13	-0.33	-0.30	-0.73
5	0.22	0.70	-0.22	-0.54
6	0.03	0.08	0.08	0.19
7	0.02	0.05	0.06	0.15
8	0.11	0.28	0.15	0.36
9	-0.36	-0.92	-0.57	-1.39
10	-0.09	-0.23	0.13	0.32
11	-0.49	-1.25	-0.12	-0.29
12	-0.24	-0.62	-0.53	-1.29
13	-0.49	-1.25	-0.45	-1.10
14	-0.05	-0.13	0.11	0.27
15	-0.49	-1.25	-0.45	-1.10
16*	-0.49	-1.25	X	X
17	0.51	1.30	-0.12	-0.29
18	0.11	0.28	0.18	0.44
19	0.81	2.08	0.85	2.07
20	-0.09	-0.23	-0.25	-0.61

*Country 16 did not rate sample B.

Figures 7 and 8 below offer another view of the dispersion of the scores, with each country number placed on the X axis.

Figure 7: Distribution of Country Zed score units for Sample A.

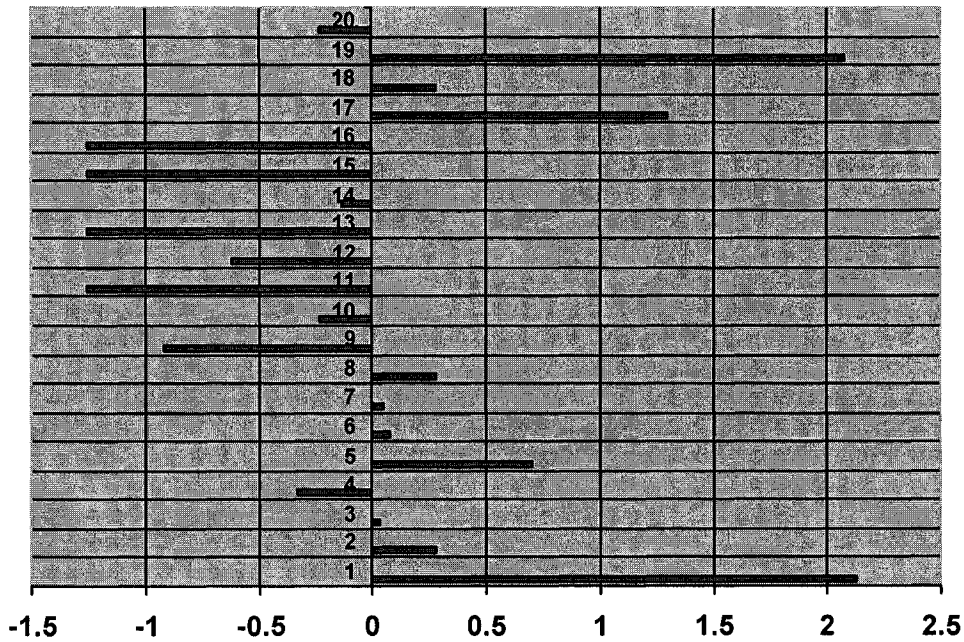
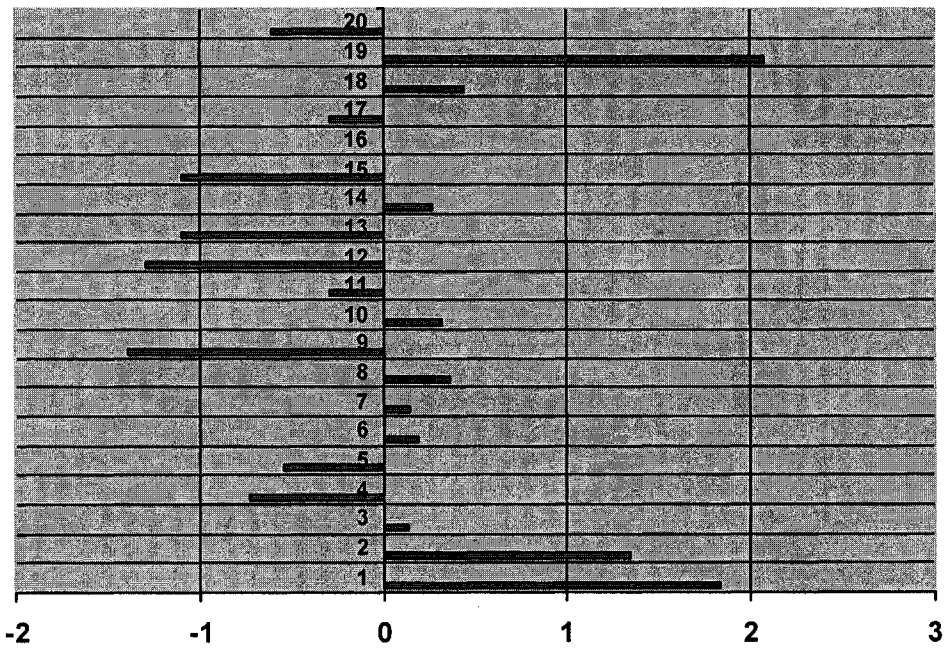


Figure 7 above illustrates and reiterates from Table 6, that there were 2 countries more than 2 SD away above the mean, while 1 country was one SD above the mean, and 4 countries were more than 1 SD below the mean.

Figure 8: Distribution of Country Zed score units for Sample B[35].



For sample B in Figure 8 above, only 1 country was more than 2 SD above the mean, and 2 countries were more than one SD above the mean. Also, 4 countries were more than one SD below the mean.

Having examined the dispersion of ratings from country to country, one question remained to be answered: Were the intra-country ratings more dispersed than the overall ratings? In order to discover if this was the case, or not, comparisons between country means and overall ratings were examined. Table 7 below compares the country means to the overall rater means.

[35] Country 16 did not rate sample B.

Table 7: Differences between Rater Means and Countries

	Mean OPI A	Stand. Dev. OPI A	Mean OPI B	Stand. Dev. OPI B
All Raters Together	1.52	.51	2.45	.58
Countries		ZED SCORES		ZED SCORES
1	0.80	1.57	0.67	1.16
2	0.08	0.16	0.55	0.95
3	-0.02	-0.04	0.35	0.60
4	-0.18	-0.35	-0.30	-0.52
5	0.19	0.37	-0.22	-0.38
6	0.0	0.0	0.08	0.14
7	-0.01	-0.02	0.06	0.10
8	0.08	0.16	0.15	0.26
9	-0.39	-0.76	-0.57	-0.98
10	0.12	0.24	0.13	0.22
11	-0.52	-1.02	-0.12	-0.21
12	-0.27	-0.53	-0.53	-0.91
13	-0.52	-1.02	-0.45	-0.78
14	-0.08	-0.16	0.11	0.19
15	-0.52	-1.02	-0.45	-0.78
16*	-0.52	-1.02	X	X
17	0.48	0.94	-0.12	-0.21
18	0.08	0.16	0.18	0.31
19	0.78	1.53	0.85	1.47
20	-0.12	-0.23	-0.25	-0.43

*Country 16 did not rate sample B.

Table 7 above facilitates the investigation of the differences in scores within each country, with the differences in scores between countries taken together, and all raters taken together. In other words, it displays whether any one country's ratings were more dispersed within, than they were in comparison to other raters. In Figure 9 below, only one country, Country 6, had perfect intra-country agreement, and was directly on the mean (represented by zero) with all raters for sample A.

Figure 9: Sample A, Zed Scores of Comparison between Country means and Rater Means

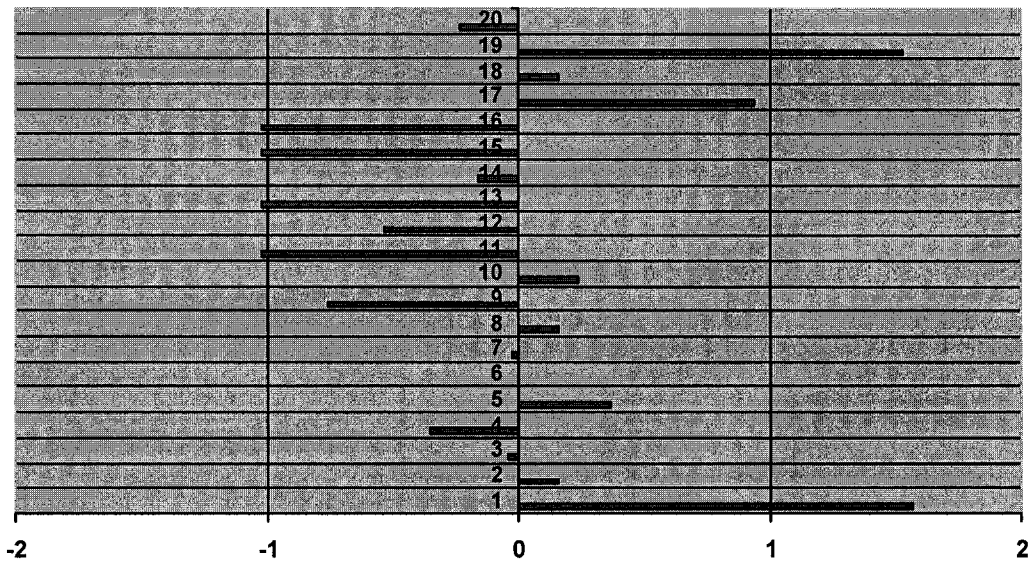
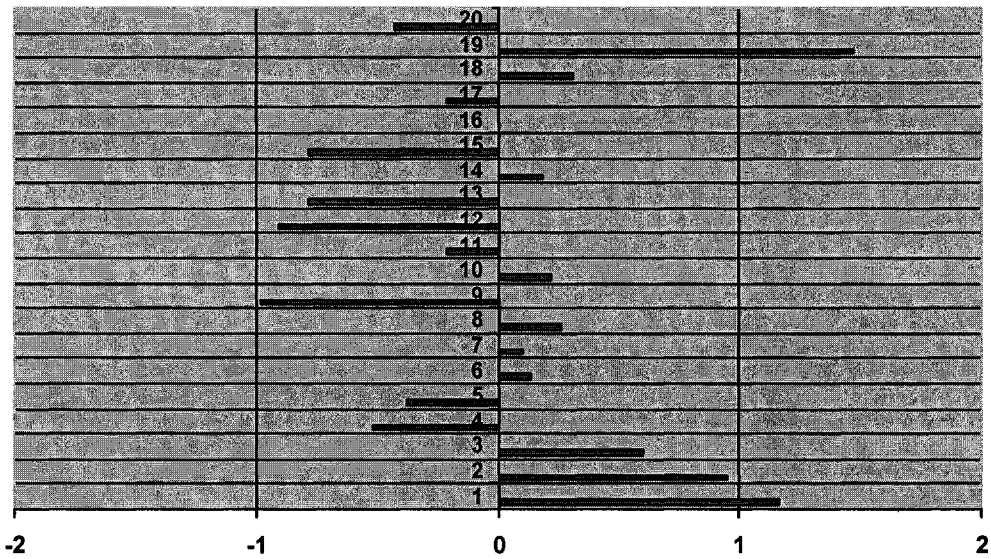


Figure 9 above illustrates that when country means with the raters' mean were compared, there was less dispersion. For example, only 2 countries were more than 1 SD above the mean, with 4 countries slightly 1 SD below the mean. Many countries were very near the mean, either above or below, which is a desirable finding. Nonetheless, the analysis suggested that in some countries, the overall dispersion of scores within countries was greater than when compared to rater scores overall, indicating less agreement within those countries, than across raters.

Figure 10: Sample B, Zed Scores of Comparison between Country means and Rater Means[36]



In Figure 10 above, again there was less dispersion than when comparing country means for sample B. This time, only 2 countries are more than 1 SD above, with none below the mean. This is an indication that within countries, there are likely to be some raters who 'skew' the mean for the group. Figure 2 and Figure 4 illustrated all raters' ratings dispersed all over the diagrams, and it can now safely be assumed that those raters represented different countries (or were from different countries). Therefore within many countries, it is possible that some of the raters are either less experienced, less well trained, or for whatever reason, did not score in agreement with the other raters from their countries.

[36] Country 16 did not rate sample B.

To summarize the findings related to country-to-country comparisons, it appeared that in some countries, there were greater differences in ratings within a country than when their country mean was compared to all raters. Ideally, no one should be more than 1 SD above or below the mean. One must also be cautious in viewing the findings, in this case, with regard to being below the mean. It could be argued that it is 'preferable' to be below the mean than above, since in sample B's case, the outlier scores of level 4 were not omitted.

And finally, of both interest and concern were analyses of correlations between ratings of OPI samples A and B, conducted to estimate the degree of relationship between the two sets of scores, or to see the degree to which both ratings were related. A Spearman rank-order correlation coefficient or $\rho = .57$, and a Pearson product-moment correlation coefficient or $r =. 55$ indicated that there were no statistical correlations between the two sets of data. A higher correlation between the two sets would have indicated that scoring was consistently low, or high. The result of this analysis is somewhat problematic, since one (a country, a testing system, etc.) cannot account for measurement error if some raters are very strict with one rating, but very lenient with another, for example. 'Random' ratings cannot be compensated for within a system, and this becomes a source of error, as mentioned in the Chapter 2 literature review.

4.3 Findings 2: Comparing Raters

Chapter 3 described the coding and analysis done to the data pertaining to the tester[37] training participating raters reported having received, and the training received on the interpretation of the STANAG scale. This section reports on findings related to the impact training had on score reliability.

4.3.1 Rater Training

The type of training participants reported having received was described in Chapter 3, and a total of 98 out of the 103 participating raters provided related information. Table 8 below provides the numbers and percentages, and demonstrates that the majority of participants in this study (63.3%) had received what can be interpreted as considerable training.

Table 8: Tester/Rater Training Received by Participants

	Number	Percent	Valid Percent
NONE	8	7.8	8.2
ONE	28	27.2	28.6
TWO	31	30.1	31.6
THREE or more	31	30.1	31.6
Missing	5	4.9	0
Total	103	100.0	100.0

These four categories were split into two broad groups, the first being, 'NONE and ONE' (little training) and the second, 'TWO and THREE or more' (considerable training), based on the criteria defined in Chapter 3. The data were crosstabbed with the ratings that the participants assigned to sample A and to

[37] Participants were asked to provide information on the **tester training** received, in terms of duration of training, number of practice tests, etc. They were not asked about **'rater'** training since in most countries, the testers are also the raters.

sample B (viewed as either a 'correct' or 'incorrect' rating, based on the mean, median and mode and the initial ratings).

Chi-square analysis of sample A ($p<.147$) did not indicate that tester training made a significant difference in scoring sample A 'correctly': ($\chi^2 = 2.10$, df =1). Nonetheless, a considerable number of participants (21/36) with 'little training' rated sample A at level 1, and more than twice as many participants with 'considerable training' (45/62) rated sample A as a level 1. Training possibly enabled these raters to rate 'correctly', however, the findings also shows that a near equivalent number of considerably trained (n=17) vs. little-trained participants (n=15) rated sample A 'incorrectly'. In other words, while it was expected that trained raters would score 'correctly', the results showed that just as many of the trained raters scored sample A 'incorrectly'.

The data, however, presented different findings when sample B was analysed. Chi-square analysis indicated that there was a significant relationship between these variables ($p<.004$), with ($\chi^2 = 11.13$, df =2). Here, 44/62 of the 'considerably' trained raters (or 76% of this group) scored sample B as a level 2 whereas only 14/36 raters (or 24% of the 'little' trained group) rated 'correctly'. Also, 20/36 of the 'little' trained raters scored 'incorrectly', whereas only 14/62 of the 'considerably' trained raters rated 'incorrectly'. (See Table 9 below).

Table 9: Score B and Tester Training Crosstabulation

		Summary of tester training		Total
		Little training	Considerable training	
Score B correct?	Yes	14	44	58
	No	20	14	34
	Missing	2	4	6
Total		36	62	98

Tester/rater training increased the likelihood of rating sample B accurately. Lower level examinees, such as in sample A, generally present fewer rating challenges, since errors, mistakes and linguistic failures (or breakdown as in OPI terms), are quite obvious even to the un-trained rater. This however, was apparently not the case for a higher-level examinee such as in sample B, whose performance required more subtle assessment. Little-trained raters may not possess the skills necessary to discern the criteria inherent in the rating process, a finding that is consistent with Henning (1996).

4.3.2 STANAG Training

In this section, findings regarding the differences in ratings between raters who received varying degrees of STANAG training are presented. 95 of 103 participants answered this question in the rater data questionnaire, and described the type of STANAG training they had received. Table 10 shows that a high percentage (60%) of participants reported having received little training or no training at all on the interpretation of the scale, highlighted below.

Table 10: STANAG Training Received by Participants

	Number	Percent	Valid Percent
NONE	15	14.6	15.8
ONE	42	40.8	44.2
TWO	22	21.4	23.2
THREE or more	16	15.5	16.8
Missing	8	7.8	0
Total	103	100.0	100.0

The trend this time was reversed compared to the tester/rater training received, in that 60% of participants indicated that they had little training 'NONE and ONE' (little training), and 40% of participants' responses falling into the group 'TWO and THREE or more' (considerable training).

The same phenomena as was reported for sample A crosstabbed with tester training, was evident when non-parametric tests were performed with STANAG training. As was seen before, scoring sample A accurately was not that challenging for testers who received no STANAG training or little STANAG training. Chi-square analysis for these variables was ($p<.092$) with ($\chi^2 = 2.84$, df = 1), which was not significant, indicating that STANAG training was not critical for rating sample A 'correctly', however the findings indicate that only 9 raters (or 28.1% of the considerably trained group) scored sample A 'incorrectly' as opposed to 23 raters (71.9% of little-trained participants). In other words, very few participants who received 'considerable' STANAG training, as defined in Chapter 3, scored sample A 'incorrectly', or outside of the mean.

Once again, results indicated a strong relationship with rating sample B, where the Chi-square value was significant at ($p<.028$) with ($\chi^2 = 7.17$, df = 2). Table 11 below, shows that similar numbers of participants of each group rated accurately, but that only 8/38 raters (or 25% those who received 'considerable' STANAG training) rated 'incorrectly', as opposed to 24/57 raters (or 75%) of the other group.

Table 11: Score B and STANAG Training Crosstabulation

		Summary of STANAG training		Total
		Little training	Considerable training	
Score B correct?	Yes	28	29	57
	No	24	8	32
	Missing	5	1	6
Total		57	38	95

Based on these results, it appears that tester/rater training, and training in the interpretation of the scale were significant factors in rating sample B 'correctly'.

4.3.3 Experience

Did experienced raters score more reliably than inexperienced ones? Are they scoring as reliably as trained raters, or can years of experience compensate for lack of training? A total of 101 participants indicated the number of years' experience they had as raters, in the rater data questionnaire. Table 12 below, reports the numbers, and highlights the fact that the majority (50%) of the participants in this study had 5 years or more experience.

Table 12: Participants' Years of Experience

		Numbers	Valid Percent
Valid	0 to 1 year	15	14.9
	2 to 3 years	20	19.8
	4 to 5 years	16	15.8
	5 years +	50	49.5
	Total	101	100.0
Missing	99.00	2	
Total		103	

To process years of experience with both sample A and sample B, two main comparisons were made. First, crosstabs were performed investigating how 'new' raters' scores compared to the scores given by more experienced raters. For sample A, as seen in Table 13 below, the threshold was set at 1 year of experience or less.

Table 13: Score A & 1 Y Experience Crosstab

Count		Experience		Total
		1 yr exp. or less	2 yrs exp. or more	
Scored OPI A within mean?	Yes	13	57	70
	No	2	29	31
Total		15	86	101

With Chi-square at ($p<.114$) the results did not approach significance, however, Table 13 shows that only 2 participants (13.3% of those who have 1 year experience or less), scored sample A 'incorrectly', whereas 29 participants (or 38.5% with 2 years of experience or more), did. Although few experienced and inexperienced participants rated sample A 'incorrectly' overall, it would appear that inexperienced raters did very well with sample A, much in the same way that little-trained raters and little STANAG trained raters had. Having experience is

clearly not a critical factor for rating sample A 'correctly'. The result was similar when the degree of experience was increased to 4 or more years, and crosstabbed again with sample A. Although, a Chi-square value of ($p<.09$), again, was not significant, it is noteworthy that only 7 (or 20% of the raters who have 4 years of experience or more) rated 'incorrectly', as opposed to the 24 raters (36.4%) who did. These findings, again, were not surprising because of the level of proficiency of the examinee in sample A, that is, a relatively easy sample to rate accurately, one can surmise, since it is at the lower end of the scale.

Once again though, the same cannot be said for the ratings associated with sample B. In Table 14 below, the ratings or scores given to sample B were crosstabbed with '1 Y or Less experience', as was done for sample A, above. With a Chi-square ($p<.06$) and ($\chi^2 = 5.56$, df = 2), the results are approaching significance. Highlighted are the scores from 2 raters (or 13.3% with 1 year testing experience or less) who scored 'incorrectly', while 33 raters (or 38.3% with 2 years of experience or more), scored 'correctly'. These findings demonstrate that experience did not provide for 'correct' ratings (or closer to the mean) with sample B.

Table 14: Score B & 1 Y Experience Crosstab

Count		Experience: 1 yr exp. or less	Experience: 2 yrs exp. or more	Total
Scored OPI B within mean?	Yes	13	47	70
	No	2	33	31
Total		15	86	101

The analysis was repeated using the data from participants with 4 years of experience or more. A Chi-square at ($p<.025$) with ($\chi^2 = 7.36$, df = 2), indicated that **there was a statistically significant relationship between the raters' experience and the score they gave to sample B. However, this was a largely surprising result.** (See Table 15, below).

Table 15: Score B & 4 Y Experience Crosstab

Count		Experience		Total
		3 yrs exp. or less	4 yrs exp. or more	
Scored OPI B within mean?	Yes	26	34	60
	No	6	29	35
Missing		3	3	6
Total		35	66	101

These results indicated that having 4 years or more testing experience, or what can be considered as a lot of experience, did not 'help' these participants rate the two OPIs more accurately (or closer to the mean). Quite the reverse seemed to be occurring. Many more raters with considerable experience, 29/66 of them, rated outside of the mean (44%), as opposed to the 6/35 raters (17%) from the other group.

In order to investigate this finding further, more non-parametric tests investigating the relationship between tester/rater training and years of experience, and STANAG training and years of experience, were performed. With a Chi-square ($p<.038$) and ($\chi^2 = 8.44$, df = 3), the result was significant. Table 16 below,

identifies who the raters were with regard to their experience and training histories.

Table 16: Years experience & tester training Crosstab

		Summary of tester training		Total
		Little training	Considerable training	
Years experience	0 to 1 year	10	5	15
	2 to 3 years	4	14	18
	4 to 5 years	4	12	16
	5 years +	17	31	48
Total		35	62	97

Table 16 above highlights that the 'considerably' trained testers have more experience, with 31 (or 32% of the total), which is a desirable finding, however, when the crosstab was done with training on the scale, there was a significant result with a Chi-square ($p<.03$) and ($\chi^2 = 8.97$, df = 3). The largest group without much STANAG training is the one with the most experience, with 29 raters with 5+ years of experience (or 31%) having received 'little' training, as presented in Table 17 below.

Table 17: Years experience & STANAG training Crosstab

		Summary of STANAG training		Total
		Little training	Considerable training	
Years experience	0 to 1 year	11	2	13
	2 to 3 years	11	8	19
	4 to 5 years	5	11	16
	5 years +	29	17	46
Total		56	38	94

What can be assumed from these findings is that 'newer' testers are being given more training in the interpretation of the STANAG scale. Perhaps more importantly, it would appear to be essential to receive tester/rater training, and specifically scale training, and not to count on years of experience, in order to achieve a reliable training system.

4.3.4 Part-time vs. Full-time

At question 9 on the rater data questionnaire, participants were asked if they conducted OPIs[38] part-time, or full-time. Nearly all of the participants answered this question (101 out of 103 participants) and with a high response rate, the data was considered representative of the participating population. 33% of respondents (n= 34) indicated that they tested full-time, and 65% (n= 67) said that they tested part-time. Are there differences in ratings between participants who test part-time versus full-time?

As a last endeavour to understand how experience influenced these ratings, and to see if there was a relationship between testing full-time and rating 'correctly', more non-parametric tests were conducted. Chi-square analysis at ($p<.003$) indicated that a significant relationship between testing (therefore rating) full-time, and scoring 'correctly', with ($\chi^2 = 8.63$, df = 1). In Table 18 below, we see that only 4 (or 12% of the full-time testers) rated sample A 'incorrectly', whereas 27 of the part-time testers did (or 40% of that group). The large majority of full-time

[38] Some participants indicated, as an aside, that they conducted oral interviews, or speaking tests, not OPIs.

testers, 30 (or 88%) rated 'correctly', and the majority of part-time testers 40 (or 60% of that group) scored 'incorrectly'.

Table 18: Conducts OPIs full-time and Rating A Crosstab

		Is rating A correct?		
		Yes	No	Total
Conducts OPIs full-time	Yes	30	4	34
	No	40	27	67
Total		70	31	101

The same trend was apparent with sample B. With a Chi-square ($p<.007$) and ($\chi^2 = 9.88$, df = 2), the same relationship is present with the vast majority of full-time testers 26 (or 77% of the group), rating sample B 'correctly', and only 5 raters (or 15% of the same group) having tested 'incorrectly'. (See Table 19 below).

Table 19: Conducts OPIs full-time and Rating B Crosstab

		Is rating B correct?			
		Yes	No	Missing	Total
Conducts OPIs full-time	Yes	26	5	3	34
	No	33	31	3	67
Total		59	36	6	101

Also, at question 11 in the rater data questionnaire, raters were asked to report on the number of OPIs conducted per year. The numbers are reported in Table 20 below. The results were also crosstabbed with both samples A and B to see if there was any relationship between the number of tests conducted per year and the ratings that were assigned to each sample. Although it is of interest to see that the largest group (29%) of participating raters estimated that they conduct between 101 and 300 OPIs (or any type of oral tests) per year, the result of the non-

parametric tests indicated that there was no statistical significance between the two.

Table 20: Estimate of # of OPIs conducted per year

		Number	Percent	Valid Percent
Valid	1-20 OPIs	15	14.6	15.5
	21-50 OPIs	21	20.4	21.6
	51-100 OPIs	28	27.2	28.9
	101-300 OPIs	22	21.4	22.7
	301+ OPIs	11	10.7	11.3
	Total	97	94.2	100.0
Missing	99.00	6	5.8	
Total		103	100.0	

In sum, these findings indicate that although many of the participants in this study have considerable experience as testers (therefore as raters), it is the full-time, 'considerably' trained participants who rated 'correctly', and especially those who reported having received training on the STANAG scale.

4.3.5 Native vs. Non-native Speaker Raters

This section reports on the investigation of the raters' background to see if they impacted the ratings assigned to the two OPI samples. More specifically, research questions asked if there were differences in ratings between native and non-native speakers of English.

Rater data question 2 asked participating raters to state their mother tongue and 101 of the 103 raters responded to this question. In order to preserve anonymity of participants (who potentially could be identified by first language), data were

then re-entered in SPSS as English mother tongue (yes or no), to investigate if there were differences in ratings between native speaker, and non-native speaker participants.

Table 21: Is English mother tongue?

		Numbers	Percent
English?	Yes	39	37.9
	No	62	60.2
	Missing	2	1.9
	Total	103	100.0

As highlighted above, the majority of participants in this research (60%) were non-native speakers of English. Non-parametric tests were conducted to see if having English as a first language, was a significant factor in rating the samples 'correctly'. Results of crosstabs with sample A indicated no significance. Results with sample B, however, were approaching significance with a Chi-square ($p<.066$) and ($\chi^2 = 8.80$, df = 4). Although not a statistically strong relationship, findings indicated that a nearly equivalent number of native and non-native speakers of English rated OPI sample B 'incorrectly' and that **many non-native speakers of English scored it 'correctly'**, as shown in Table 22, below.

Table 22: Score B and English mother tongue Crosstab

		Is English mother tongue?			Total
		Yes	No	Missing	
Score B correct?	Yes	19	40	1	60
	No	17	19	0	36
	Missing	3	3	1	7
Total		39	62	2	103

This finding, as stated above was not statistically strong; however, it was consistent with findings previously reported in Chapter 4, such as the factor of experience in rating 'correctly'. The fact of being a native speaker, of having a lot of experience, cannot be considered adequate conditions for assuming that ratings will be accurate or 'correct', as previously defined in Chapter 3. It could be that non-native speakers of English are receiving or have received more training than the native speakers. It is also possible that native speakers interact with the scale more intuitively than non-native speakers; however these are merely speculative considerations, since the data collected did not allow for a deeper analysis of this finding.

The rater data questionnaire also asked all participants to provide their 'self-assessed' level of English proficiency based on the STANAG scale, by circling one of the levels, and 96 of the 103 participants did so. It is interesting to note that although 'plus' levels were not offered with this question in the questionnaire, 10 participants drew the pluses beside the level they had circled. In Table 23 below, we note that the majority of participants placed themselves in level 5, for a total of 36% of participants who answered this question. Incidentally, as seen in Table 21 above, 38% of participants in this study reported being native-speakers of English.

Table 23: Participant Self-assessed English Proficiency

		Frequency	Percent
Levels	Level 2	1	1.0
	Level 2+	1	1.0
	Level 3	18	17.5
	Level 3+	8	7.8
	Level 4	21	20.4
	Level 4+	1	1.0
	Level 5	37	35.9
	Don't know	9	8.7
	Total	96	93.2
Missing	99.00	7	6.8
Total		103	100.0

The next largest group (25%) placed in the level 3 range, 21% in the level 4 range, while only 2 participants estimated their own proficiency level to be in the level 2 range. One of the raters, *Rater 59*, had self-assessed his/her proficiency at level 3, and stated on questionnaire B that s/he could not understand sample B well enough to feel comfortable rating him due to his 'strong' accent and frequent mispronunciation, although s/he provided a rating. This factor could possibly account for some of the missing ratings on sample B.

4.3.6 'Old' NATO vs. 'New' NATO Country Raters

This section presents more findings related to raters' background, and the impact, if any, to the rating of the two OPI samples. More specifically this time, research questions asked if there were differences in ratings between participants from 'Old' and 'New' NATO countries.

Data for each participant were entered in SPSS regarding their country of origin, and then, whether they were from a 'new' NATO country or from an 'older'

NATO country. Participants that fit in neither group were excluded, and are represented by the designation 'other' as indicated below in Table 24.

Table 24: New NATO Country?

		Frequency	Percent
New NATO?	Yes	37	35.9
	No	55	53.4
	Total	92	89.3
Others		11	10.7
Total		103	100.0

We see above that 53% of are from 'older' NATO countries, and 36% of participants are from 'newer' NATO countries. In order to investigate whether or not there were differences in ratings that can be attributed to 'old' NATO or 'new' NATO countries, (as has been anecdotally reported in Chapter 1), further non-parametric tests were conducted to see how each group had rated the OPIs. Crosstabs of 'new' NATO with ratings for sample A did not yield any statistical significance and the patterns of distribution of 'correct' and 'incorrect' ratings, (or close and far from the mean), were similar for each group. Once again, the same could not be said for sample B. The largest group of raters was from 'old' NATO countries, but as Table 25 below shows, equivalent numbers of raters from both groups rated sample B 'correctly', while only 6 raters from the 'new' NATO group of raters rated 'incorrectly'.

Table 25: New NATO and Ratings B Crosstab

		Ratings B 'Correct'?			
		Yes	No	Other or Missing	Total
New NATO?	Yes	27	6	4	37
	No	27	26	2	55
Total		54	32	6	92

With a Chi-square ($p<.007$) and ($\chi^2 = 10.03$, df = 2), the results are significant. Since this study suggests that training, both as a tester/rater and in the interpretation of the scale were the most important factors in rating 'correctly', then a hypothesis would be that raters in 'new' NATO countries, have received more training than raters from older more established NATO countries. This was explored further, and crosstabs were performed to see if tester/rater training and country of origin had a statistically significant relationship. In Table 26 below, we see that the participating raters from 'new' NATO countries were the ones who reported having received the most tester training. With a Chi-square ($p<.006$) and ($\chi^2 = 7.68$, df = 1), these results were indicative of that relationship.

Table 26: Country and Tester Training Crosstab

		Summary of Tester Training		Total
		Little training	Considerable training	
New NATO?	Yes	6	30	36
	No	23	28	51
Total		29	58	87

Only 6 raters from the 'new' NATO group (or 17% of the total new NATO participating population) reported having received what was coded as 'little

training', and 30 raters (or 83% of the same group) fell into the 'considerably trained' group. The same tests were conducted with the amount of STANAG training that the participants reported having received, and none of the tests approached significance. In sum, the hypothesis advanced above was demonstrated as being true; testers in 'new' NATO countries appeared to be the 'better' trained group overall.

4.4 Findings 3: Rating Processes and the Scale

The findings in this section of Chapter 4 will describe the findings regarding the rating process reported by participants, and present their evaluation of various linguistic, discursive, and sociolinguistic components, for each of the OPI samples rated. This section will also provide a glimpse into the raters' interactions with the NATO STANAG scale, seeing as raters were asked to list the STANAG descriptors they felt best described each of the examinees. Finally, it will provide a summary of the participants' comments regarding the NATO STANAG scale and its user-friendliness.

4.4.1 Rating Processes

In order to understand how they went about rating the candidates in each of the samples, both questionnaires, the one on sample A and on sample B, posed this question to participants: 'How did you arrive at this rating? Explain the steps you took.' The rating process followed was explained by 81% of participants (n=83), with 19% either having left this blank or having explained their ratings without providing insights into the steps taken to arrive at the ratings, therefore those entries were omitted. Seeing as a rater's process was comparable for both samples, the data were collapsed into one main heading with various sub-components, each crosstabbed for significance with ratings. Only the most interesting will be reported below.

For example, only 14 participants mentioned listening twice or more to the samples. Also, 12 participating raters indicated that some form of score 'manipulation' took place within their systems, such as converting the score (based on criteria not supplied), or converting percentages (cumulated from the components on their grids), into levels, or adding up various ticks from components and arriving at a numerical score then converted to the scale. These various rating methods were crosstabbed with scores of both A and B and none were found to be significant.

In the rating process, only 36% of participants (n=37) mentioned consulting the STANAG scale during the rating process. Did consulting the STANAG enhance scoring 'correctly'? Crosstabs with the 'correct' score of sample A indicated significance, with Chi-square ($p<.046$) and ($\chi^2 = 3.98$, df = 1). Table 27 below demonstrates the distribution of scores:

Table 27: Mentioned consulting STANAG with Score A Crosstab

		Is rating A correct?		
		Yes	No	Total
Mentioned consulting STANAG?	Yes	30	7	37
	No	41	25	66
Total		71	32	103

As can be ascertained from Table 27, 7 raters (or 19% of those who mentioned consulting the STANAG in their rating of Sample A) scored 'incorrectly' as opposed to 25 raters (or 38% of the raters who rated 'correctly'). Crosstabs for sample B were not significant, which is an interesting, albeit contradictory finding, since sample B was more challenging to rate. Lastly, 34% of participants (n=35)

mentioned consulting their own institute or school's rating grid (or evaluation criteria) during the process; however this was not found to be significant in the tests conducted with scores assigned to either sample.

In order to represent the diversity of approaches taken by the raters, the next section will provide excerpts from the raters' questionnaires that are characteristic of the samples collected. In addition to reporting on the rating process, included are the answers provided to two questions that aimed to gather more fined-tuned and specific responses from the participating raters. Question 3 asked: "At what level would you say the examinee performed the tasks or functions very well? Please provide detailed explanations". Question 4 asked: "Where or during which tasks and functions did you think this examinee did **not** perform well? Please provide detailed explanations". As explained in Chapter 3, these questions were designed to elicit open-ended responses, and were formulated as broadly as possible so as to avoid suggesting any particular level on the scale to the raters. This allowed for an analysis of the rating factors that had been considered by them, and how they judged the performances of the examinees in both samples. Again, we can contrast between the 'evidenced-based', the 'intuitive' and the 'extra-contextual' raters, as viewed through their reported rating processes.

The following rater provided a full description of the steps taken to arrive at the sample A rating:

1. I listened to the CD (the examinee's performance) carefully.

2. While listening, I was taking notes on his performance. I jotted down sentences that either contained basic errors or were quite good/correct.
3. I went over my notes to get a clear picture of the examinee's overall performance.
4. I compared my notes (the examinee's performance) against the STANAG level 1 and level 2 descriptors.
5. Finally, I decided to rate the candidate at 1+. *(Rater 26)*

Another typical rating process:

> I listened to the tape and filled in the observation grid, listing the tasks assigned by the interlocutor. Most of them were for level 2 (narration past, present, future, description, role-play). I tried to assess the working level, which on most of the tasks performed seemed to be level 2, sometimes below, sometimes slightly above this level, but generally level 2. However, I was often hindered by the quality of the sound sometimes missing phrases and sentences (on the tape). I was consulting the Interpretation of Language Proficiency Levels. *(Rater 18)*

And yet another, this one for sample B:

> I compared the candidate's performance with the STANAG criteria (levels 2 and 3) and decided that he did not meet the requirements for level 3 with regard to flexibility and the use of structural devices. Errors were frequent not only in low frequency structures, but in some high frequency areas as well. *(Rater 90)*

These raters as shown above proceeded systematically in reviewing the STANAG criteria for each level, and then assigning an 'evidenced-based' rating.

Rater 1 below who rated sample A as a **level 3**, therefore was one of the outliers in the data, was in stark contrast to the raters above, and gave this 'intuitive' rating:

> I would say that just about every single sentence in the Interpretation of the LEVEL 2 speaking could be applied to this

> man. And because of that I would say that he is literally at the top of level 2. He is on the verge of level 3 literally. So I would automatically up him to a low 3. *(Rater 1)*

Rater 1 provided no evaluation of linguistic features, and seemed to lack knowledge of concepts of threshold, of range, or of sustained performance, and rated without any of the 'usual' considerations that trained raters look for. Further in the questionnaire, in response to the question on the use of pluses, *Rater 1* added:

> I wouldn't give him a 2 plus but I would give him a 3 minus. I have to admit that I am basing that decision on the fact that by demonstrating he is a high 2 in every single aspect of the description of a level 2, I would give him a sort of vote of confidence that in any job abroad he might have a hard time at first but I think he could handle really working in the language. *(Rater 1)*

Rater 1 now provided an 'extra-contextual' rating. Instead of conducting an objective assessment, based on scale-related criteria, *Rater 1* now appeared to be motivated by concerns beyond the test.

This participant had the following process for rating:

> First I tried to listen to him carefully without thinking of the STANAG. Then I tried to concentrate on things the candidate can do and what he cannot do. I analysed the different tasks and topics, wrote them down on my scoring sheet and I tried to find out his working levels and to give him a final score. *(Rater 101)*

Rater 101 did not explain why or how not thinking of the STANAG would assist in rating, and that approach certainly was not typical of the processes explained

by the participants. Nonetheless, *Rater 101* rated 'correctly' as previously defined.

The following rater rated sample B as **level 3** and gave the following answer to the question asking raters "At what level would you say the examinee performed the tasks or functions very well? Please provide detailed explanations."

> First I thought of a level 2 student. But when he came to discuss sophisticated topics like 'NAFTA' related problems and international terrorism, he performed very well and was able to find satisfying solutions to the problems. *(Rater 86)*

Next in the questionnaire, at question 4: "Where or during which tasks and functions did you think this examinee did not perform well? Please provide detailed explanations", the same participant answered:

> There were no tasks he had problems with. *(Rater 86)*

It is interesting to note that the 'intuitive' rater above seemed to perceive the introduction of topics as an indication of the level the examinee had, regardless of the examinee's language-related performance on those topics, however, the reply to question 4 indicated that performing the tasks, or attempting the tasks was a sufficient condition for success, for that rater.

Rater 95, who did not mention consulting the STANAG to assist in the rating, mentioned that after getting to questions pertaining to specific statements from the STANAG, read the STANAG and then changed the score to a lower rating. *Rater*

95 initially rated sample A at **level 3** and then L3 was marked with an X. The participant wrote: "I got to question 7 and re-read the STANAG document and now I think '2' is more appropriate". In the next section where information on the rating process and the steps following to reach the rating, this same participants added:

> Level 3 is the basic level needed for officers in (my country). I think the candidate could perform the tasks required of him. He could easily be bulldozed by native speakers in a meeting, but would hold his own with non-native speakers. He makes mistakes that very rarely distort meaning and are rarely disturbing *(Rater 95)*.

This rater defended the original rating of **level 3,** although it appeared that it was felt he was at level 2 when s/he consulted the STANAG. Again, it was interesting to note the 'extra-contextual' raters' views and concerns. These possibly reflected the institutional view, or the type of guidance and approaches to testing (and rating) that are in practice in some countries or institutions.

Comments made by raters were in accordance with the level they perceived the examinee to have. Regarding sample A, this rater wrote:

> For a level 3, the exam should have been 'deeper' and more difficult topics should have been introduced & more difficult vocabulary used. *(Rater 98)*.

Rater 98 saw the examinee in sample A as having a level 3 or more specifically, as being 'given' **a level 3 test**. Most raters felt that the examinee in sample A did very well on level 1 tasks, managed some level 2 tasks, but had difficulty with much of level 2 language, as defined by the STANAG. What the raters had to say

about the examinees' linguistic strengths and weaknesses and other factors investigated, is reported below.

Question 3 of both questionnaires asked: "At what level would you say the examinee performed the tasks or functions very well? Please provide detailed explanations".

Sample A: A total of 73 of 103 participants responded with mention of tasks and functions in relation to these questions, and matched them up with the level it is expected these tasks belong, as well as the examinee's level of linguistic 'comfort'. 26 participants responded by describing sample A's linguistic features such as grammar, lexis, or his cohesion and communicative features, but did not specify tasks or functions, or any level of language which could be matched up with the scale.

The most frequently mentioned task that was reported as done 'well' by the examinee was **simple conversation**, which respondents described in a variety of ways from short conversation, as in the STANAG, or casual conversation on simple topics, basic everyday conversation about simple topics, etc (24 respondents mentioned this). Next are the many **descriptions** that took place during the interview, such as of his hotel room, his trip, etc. (with 20 respondents). The third task that was mentioned as being the most 'well performed' by the

candidate in sample A was the **role-play** (with 12 respondents), although many of these respondents added that it had not been 'very' well performed.

Question 4 asked: "Where or during which tasks and functions did you think this examinee did **not** perform well? Please provide detailed explanations".

The most frequently mentioned task, that was judged **not** to have been well performed by sample A, was clearly **narration in the past**, with 48 responses. Next came both **future narration** and **asking questions** with 26 responses each, followed by **instructions** with 23 responses.

Out of the total number of raters who rated sample A in the level 2 range prior to the adjusted score using plus level 'speculation' (40 raters), 28 indicated that they found that the candidate's performance (in sample A) on the narration tasks, was poorly done. This was interesting, since at level 2, the STANAG states that speakers' "basic grammatical relations are typically controlled" whereas at level 1, "Time concepts are vague, and may use only one tense". Many raters indicated that this speaker could not use tenses with any control. This was evident to the 28 raters mentioned above, who despite this observation, initially gave him a level 2.

As mentioned previously, the question was posed in such a way as to avoid leading the participants to any level; however, it is clear that the term 'performed very well' is relative to the level of perceived proficiency. For example if the examinee in sample A was perceived as having level 1, respondents may have

said that he performed well on level 1 tasks, and name those where he was especially good. They would then indicate that he did not perform as well on some level 2 tasks. For example, *Rater 20*'s answer to question 3 on sample A:

> The examinee performed well the tasks about narration, past and future i.e. his plans for his immediate future after he's back from Canada. Obviously he's doing well at the concrete topics, esp. those touching upon his present activities. *(Rater 20)*

Rater 20 whom we saw as an 'extra-contextual' rater in excerpts on rating with the plus level, in response to the following question 4, remarked:

> The examinee definitely experiences difficulties with past narration; the past forms are almost non-existent. Sometimes they appear, but obviously this grammar aspect is the weakest point that prevents him from getting a sustained level 2. *(Rater 20)*

Rater 20 assigned level 2 to sample A, but later indicated that level 1+ would be a more accurate rating, if pluses were in use in his/her rating system.

And finally, in response to question 4, asking which tasks and functions this examinee did not perform well *Rater 56* said:

> It is not reflected in our rating scale. *(Rater 56)*

However, by the time the same respondent had reached the same question in the questionnaire on sample B, s/he listed tasks performed well, and not well performed by that examinee. It is possible that this participant came to understand the question during the course of answering the survey. For raters not

familiar with this type of oral test, answering the questions was undoubtedly challenging.

This rater's answer to question 3 for sample A contradicted the statements made at question 4 further below:

> In our system, he would be a solid STA 2. He could perform all tasks in an adequate way, but after he reached his ceiling he became insecure about his knowledge. He was consistently at STA 2. *(Rater 67)*

> His speech was fragmented throughout the OPI, I think he could perform adequately in all tasks until he reached his ceiling then he broke down. He had problems asking more complicated questions or giving detailed descriptions. *(Rater 67)*

Giving detailed descriptions of people, places and things is expected at level 2 using concrete language and speaking in complete but simple paragraphs. This examinee was reported by the majority of participants as having difficulty achieving this fully, especially due to grammatical inaccuracies manifested in lack of tense control.

Participants were asked to state what they thought of this examinee's sociolinguistic competence, at question 7: "When comparing the performance from the examinee to the STANAG descriptors, where do you think this examinee fits in, in terms of the overall sociolinguistic aspect?" A total of 28 respondents assessed him at level 1, and 2 respondents at level 1+. The majority, (39 respondents), placed his sociolinguistic competence in the level 2 range, and 2

participants placed him at level 3 sociolinguistically. Of the 39 who placed him in the level 2 range, 14 of them rated the sample overall in the level 1 range despite this perceived 'peak' into the next higher level (as highlighted below, in Table 28). It would indicate that he was perceived as having the ability to respond appropriately to the situation, according to the STANAG descriptors relating to this particular aspect: "Can interact with native speakers not used to speaking with non-natives, although natives may have to adjust to some limitations" and "However, the individual generally speaks in a way that is appropriate to the situation, although command of the spoken language is not always firm". What is also interesting to note regarding the responses to this question, and illustrated in Table 28 below, is that 9 respondents indicated that they did not understand this question, 5 indicated that this aspect was irrelevant or not applicable at this level, and 3 respondents wrote that the STANAG has no sociolinguistic aspect. 11 respondents left this question blank, while the remainder wrote comments, which were not pertinent to the question posed.

Table 28: Question 7 sample A: Overall sociolinguistic (SL) aspect?"

L1	L1+	**L2**	L3	Left blank	Don't Understand Q	Irrelevant at this Level	STANAG has no SL aspect	**Of 39 who gave L1 rating overall, gave L2 in SL**
28	2	**39**	2	11	9	5	3	**14**

Although it can be argued that at this level, speakers do not control formal and informal register shifts, and that they are not necessarily placed in role-play situations whereby this is specifically assessed, the fact that the sociolinguistic

aspect was deemed irrelevant, or inexistent may be due to the context in which some of the participating raters live and work. This rater's comment would seem to present this view:

> We don't have a descriptor that we'd call 'sociolinguistic'. The reactions of this candidate to the interlocutor were not always appropriate. For example, when he was told to ask about the examiner's job and asked about his children instead. *(Rater 89)*

The use of ratings grids exclusively, and not referring back to the full STANAG, could also potentially be a factor creating rater variance in the NATO language system. It is conceivable that when English is used as a foreign language in an international setting, and used with non-native speaker teachers, testers/raters and students, concepts of 'appropriacy' to the situation and 'understandability' to the native speaker become irrelevant and are not perceived as criteria to be assessed. However, as *Rater 89* pointed out, the examinee's responses were not always in accordance with the question posed.

Sample B: 60 of 103 participants answered question 3 on sample B, (asking about the level where the examinee had performed the tasks or functions well), by mentioning tasks and functions performed well. 28 participants answered this question but did not mention any tasks *per se*, while 15 participants left it blank. At question 4, requesting mention of the tasks considered **not** well performed, 56 raters provided tasks or functions, while 28 again responded with other information not related directly to the question. 19 participants left this question blank.

20 respondents felt that the examinee in sample B had performed the task **description** best, with **narration in the past** as a second choice with 16 responses, and tied for third place are **narration in the present**, and the **hypothetical** probe for 14 responses. On question 4 when asked during which tasks he had not performed very well, 27 respondents answered the **opinion** question on smoking, 19 mentioned the **hypothetical** terrorism question and 18 said it was the **abstract discussion** of NAFTA. However, as opposed to the responses to sample A, for this sample, the number of participants who gave sample B level 3 and mentioned that abstract or level 3 tasks were not performed well was only 9. Sample B was treated differently with regards to tasks. Most respondents who gave him level 3 judged that he had performed well on most tasks at level 3 despite difficulties with linguistic or discursive aspects. The excerpts below offer a view of comments provided by the participants. Regarding sample B, this rater *(Rater 67)* wrote:

> I would give this candidate STA 3, but in OPI terms a 2+. *(Rater 67)*

This indicates that there is a perception that STANAG allows for more lenient ratings (in of itself as a scale) than this perceived 'stricter' OPI test (also based on the STANAG!), however, it could also be that these raters are indicating that in 'their' system it would be rated higher, but seeing as this is not, they are offering both options. This same rater answered question 4 for sample B, as follows:

> He is not good at abstract language and expressing abstract ideas. His vocab is poor for that. *(Rater 67)*

This rater noticed sample B's limited capacity to discuss abstract topics due in part to vocabulary, and structure. It is not clear if the rater is also aware that abstract language and the ability to express abstract ideas with the appropriate level of language is a 'hallmark' of level 3. The final comment from this rater places the rating in perspective:

> It is very difficult or rather impossible to fit a multi-level performance with a single level system. *(Rater 67)*

This speaks to two issues; the first and most obvious is that this participant is following the researcher's instructions to rate the sample as one would, using their rating protocol in their country. Since in his/her institution STANAG-based tests are single level tests, there is the added challenge for this rater to place these samples into single levels since the tasks and functions are usually operating on a working level and a probed level. This rater was aware of OPI terminology such as probes, ceiling, breakdown, however these terms at times were applied in an unsystematic fashion since the single level system may not be concerned with multi-level testing mechanisms such as probing and establishing the floor, the ceiling and the breakdowns, etc.

There was consistency to the remarks made by raters who rated sample B as having level 3. Many mentioned higher-level language tasks, but the appearance of the topic seemed to indicate to them that the examinee was at that level (See Appendix B for more OPI information). In terms of accuracy at responding to the level required by STANAG, *rater 67* above obviously did not perceive a

discrepancy between the complexity of the topic and the expected performance at this level, and the examinee's responses. Also, some variations regarding rating processes can be viewed as a factor influencing judgment of the samples, as exemplified by the following rater:

> As we don't record interviews or take notes, my assessment might have been slightly less lenient than it would have been in our usual exam situation. Note-taking means I concentrate more on errors. *(Rater 89)*

Rater *(Rater 98)* explained his/her rating process for sample B:

> Comparing the student's performance to STANAG proficiency levels, he would seem to be a level 3/4 student. He is fluent, reacts very well to questions and uses a higher level of language and structures compared to student in sample A. *(Rater 98)*

Comparisons between the two examinees were being made, even though norm referencing is not a typical practice with criterion-referenced scales. A little further *Rater 98* pondered the rating:

> Is this a level 4 candidate? The longer the test went on, the better his performance and the complexity of his language skills improved. Could argue, give opinions and introduce own experience – so yes, he is a level 4 candidate. *(Rater 98)*

And when asked at question 4: Where or during which tasks and functions did you think this examinee did **not** perform well? Please provide detailed explanations, *Rater 98* answered:

> Talking about daily routines, especially his wife. *(Rater 98)*

It is interesting to note that 'talking about daily routines' is a level 2 task. As a last comment, *Rater 98* wrote:

> Is he good enough for a 5 or a 4+? No, some basic grammar mistakes prevent this, and pronunciation is not that of a level 5, but excellent level of language and communication for a non-native speaker. *(Rater 98)*

The comments did not seem to pertain to criteria of accuracy found in the NATO STANAG scale, but instead represented an intuitive and impressionistic rating. According to the STANAG, grammar mistakes in more complex language forms are acceptable at level 2. At level 3 on the STANAG scale, grammar is fully controlled except for more complex structures, etc. The reference to sample B displaying an excellent level of language and communication for a non-native speaker denotes that this rater reinterpreted the scale in a way arguably unintended by the original construct of the scale.

As had been done with sample A, participating raters were again asked what they thought of the examinee in sample B's sociolinguistic competence. As illustrated in Table 29 below, 77 raters answered this question.

Table 29: Question 7 sample B Overall sociolinguistic (SL) aspect?"

L1	L2	L3	L4	Left blank	Don't Understand Q	N/A to Sample B	STANAG has no SL aspect	**Of 32 who gave L2 rating overall, gave L3 in SL**
3	29	**32**	2	11	3	5	3	**10**

The majority, of participants (32) felt his sociolinguistic competence was in the level 3 range while 29 felt it was at a level 2, and 2 at a level 4. 3 raters placed him in level 1, 1 because of grammar, and the other 2 because of mispronunciation, even though these aspects are not linked to sociolinguistic ability. 3 participants reiterated that they did not understand the question, and 5 mentioned that this aspect or factor was not applicable to this sample, without stating specific reasons for their viewpoint. 10 of the raters, who placed him at a level 3 sociolinguistically, rated him as a level 2 overall. The remainder of participants' responses were not related to the sociolinguistic competence, or did not place him in a clear level. As with previous findings for sample A, it indicated that although some of the factors may be perceived as having 'peaks' into the next higher level, like the sociolinguistic ability of this candidate, for these raters, the performance of the examinee in sample B did not warrant a plus level.

To conclude with sample B, it is worthwhile mentioning the number of paralinguistic features that participants reported as contributing to the overall proficiency of the examinee in sample B: 5 raters mentioned his personality, 4 mentioned his sense of humour, 3 comments were made regarding his level of confidence and 3 about his relaxed demeanour, 2 comments were made regarding his energy, and 2 about his attitude, as well as 1 comment regarding his pride in his job and status. These comments may or may not have influenced the ratings

that the raters assigned to sample B, but they certainly were viewed relevant enough to be included in the assessment made.

4.4.2 Rater/Scale Interactions

At question 8 on each of the questionnaires, raters were asked: "Are there any of the STANAG descriptors which you feel best exemplify this examinee's performance? State them." A total of 81 of 103 raters (79%) provided answers to this question for sample A. The most frequently stated STANAG statements or descriptors for **sample A**, are provided as follows:

SAMPLE A:

1.11 *Speech is often characterised by hesitations, erratic word order, frequent pauses, straining and groping for words (except for routine expressions), ineffective reformulation, and self-corrections.* **(30%)**

1.4 *Can typically satisfy simple, predictable, personal and accommodation needs; meet minimum courtesy, introduction, and identification requirements; exchange greetings; elicit and provide predictable, skeletal biographical information; communicate about simple routine tasks in the workplace; ask for goods, services, and assistance; request information and clarification; express satisfaction, dissatisfaction, and confirmation.* **(28%)**

1.7 *Seldom speaks with natural fluency, and cannot produce continuous discourse, except with rehearsed material.* **(27%)**

1.8 *Nonetheless, can speak at the sentence level and may produce strings of two or more simple, short sentences joined by common linking words.* **(26%)**

1.10 *Time concepts are vague. May often use only one tense or tend to avoid certain structures.* **(25%)**

The most frequently mentioned descriptors shown above are all from the level 1 STANAG descriptors. And again, at **sample B** the same question was posed. A total of 79 of 96 raters (77%) provided answers to this question. The most frequently stated STANAG statements or descriptors are provided below.

SAMPLE B:

2.3 *Can confidently handle most normal, casual conversations on concrete topics such as job procedures, family, personal background and interests, travel, current events.* **(33%)**

2.10 *However, the individual generally speaks in a way that is appropriate to the situation, although command of the spoken language is not always firm.* **(32%)**

2.6 *Can combine and link sentences into paragraph-length discourse.* **(30%)**

2.2 *In these situations the speaker can describe people, places, and things; narrate current, past, and future activities in complete, but simple paragraphs; state facts; compare and contrast; give straightforward instructions and directions; ask and answer predictable questions.* **(28%)**

3.1 *Able to participate effectively in most formal and informal conversations on practical, social, and professional topics.* **(28%)**

2.9 *Errors in pronunciation, vocabulary, and grammar may sometimes distort meaning.* **(25%)**

Statement 3.1 above is a level 3 descriptor from the STANAG scale. This was considered by 28% of raters as one of the statements best describing this examinee's proficiency. Theses findings are considered relevant in providing an

insider's view of what raters considered to be STANAG-based qualifiers for the performances on each sample.

4.4.3 Rater Views on the STANAG

In the rater data questionnaire, participants were also asked to state whether or not they found the STANAG scale easy to use and apply. Only 64 participants answered either yes or no to this question, with 52 answering yes, and 12 answering no, to the question. However, this cannot be construed as a completely positive endorsement, because as the following two tables demonstrate, more comments regarding STANAG's challenging aspects were collected (59), than positive ones (22). In Table 30 below, the positive comments have been tabulated for ease of reading, and the aspects most frequently referred to have been highlighted.

Table 30: Positive aspects of STANAG

		Frequency	Percent	Valid Percent
Topics	Its clarity of terms	4	3.9	18.2
	Clear-cut requirements	3	2.9	13.6
	Not overly detailed	2	1.9	9.1
	Gets easier with time & practice	7	6.8	31.8
	Just follow descriptors	2	1.9	9.1
	User-friendly	4	3.9	18.2
	Total	22	21.4	100.0
Missing	99.00	81	78.6	
Total		103	100.0	

The majority of those who responded with positive comments felt that the STANAG scale became easier to use with time and practice, which is likely to be

the case with any rating scale. Table 31 below highlights the points made by the participants regarding some of the challenges they have encountered with the STANAG scale.

Table 31: Challenging aspects of STANAG

		Frequency	Percent	Valid Percent
Topics	Size of bands	4	3.9	6.8
	Vagueness of terms	13	12.6	22.0
	Borderline cases are difficult	10	9.7	16.9
	No plus levels	4	3.9	6.8
	Native speaker criteria	2	1.9	3.4
	Difficult to interpret	2	1.9	3.4
	Vague language requires rater to use general impressions	2	1.9	3.4
	Subjectivity, human factor	4	3.9	6.8
	Too much variance in interpretation	4	3.9	6.8
	Criteria for 3 too high	2	1.9	3.4
	Not detailed enough	4	3.9	6.8
	No clear theoretical basis	4	3.9	6.8
	Difficult to evaluate higher levels	4	3.9	6.8
	Total	59	57.3	100.0
Missing	99.00	44	42.7	
Total		103	100.0	

Finally participants were asked to write any other comment pertaining to the STANAG scale, and a total of 38 participants commented, with the two most frequent ones being the need for STANAG to have plus levels, with 8 entries or 21% of respondents, and the need for the STANAG to offer more amplification, a sort of more detailed guide to understanding what is meant by some of the statements, such as 'typically controlled' and 'high frequency utterances' with 7 entries or 18% of respondents.

4.5 Control Group Results

As mentioned in Chapter 3, graduate students enrolled in a language testing course, as well as the professor, were used as a control group (CG). Since none of these participants had ever heard of the STANAG (apart from the professor), they were first given a 30-minute briefing on the STANAG scale. The briefing summarized the concepts of the scale and of each level, but did not train them to interpret the descriptors or statements or how to rate the OPIs. They were informed that 'plus' levels could be used and a brief explanation of what these represented was provided[39]. The purpose of using a CG was to see how close or far they would be from the overall rater mean, and to compare the CG ratings to the ratings assigned by lesser-trained participants.

Sample A: Regarding their rating process, CG members mentioned that they had paid close attention to the STANAG scale, reading the descriptors of each lower and higher level carefully more than once. 2 CG raters placed sample A in level 1 while 3 placed him in level 2. The mean was 1.72; both the median and mode were 2.00. When at the end of the questionnaire they were asked if pluses would be helpful in rating this sample, 1 rater who rated level 1 would not assign a plus on account that 'basic needs were not performed well in his tasks (role play)'; 1 rater stayed with the original rating of level 1+, one rater lowered from a level 2 to a level 1+ rating and 2 level 2 ratings remained at level 2, therefore shifting the median and mode to 1.00 and with a new mean at 1.40. It was interesting to see that participants in the CG did not make use of the pluses to increase their ratings,

[39] The explanation was based on the draft BILC STANAG 'plus levels' document.

and that one lowered the score from a level 2 to level 1+. When the overall participating raters' mean of 1.52, and adjusted mean of 1.33, were compared with the CG mean of 1.72 and adjusted mean of 1.40 for sample A, there are only small differences between them, with the overall CG mean being higher. Rating sample A accurately was not so difficult for this group, despite not being acquainted with the scale, and having never rated an OPI before.

Sample B: Of the CG, 1 rater placed him at level 2, 1 rater at 2+, and 3 raters at level 3. The mean was 2.72; the median and the mode were 3.00, compared with the participants' sample B mean of 2.45, median and mode of 2.00. None of the raters made use of the plus levels with sample B at the end of the questionnaire. CG members generally found this sample more difficult to rate. One CG participant provided additional information on her rating process:

> As an untrained rater in the STANAG scale, but an experienced holistic/analytic rater on other tests, I was surprised that I needed the entire speech sample before I decided where to place this candidate – but I was convinced he was a solid 2 by the 31 minute mark. The rest was just confirmation of this judgment... *(CG Rater 5)*

This rater's comment was interesting because it demonstrated that each part of the OPI must have been adding and contributing to the sample. Only once she felt that enough of a sample had been heard, was she 'convinced' of the rating. This may also be a good argument to keep the OPI time at around the 30-minute recommended time.

The CG's results on sample B were similar to the overall results, with some at level 2 and some at level 3; however, the majority of CG members rated him at level 3. Much like the raters in the study who rated sample B at level 3, the CG members commented on his ability to discuss issues, but not on his language accuracy while doing so. As was presented in the overall findings, training, or lack thereof, was an issue for the CG and rating sample B correctly.

The raters in the CG 'behaved' much like the lesser-trained raters from the main study, but since they had no vested interest in rating outcome, were all 'evidenced-based' raters.

A review of the findings will be presented in Chapter 5, as well as a discussion of their implication for the NATO language testing context.

CHAPTER FIVE: CONCLUSION

5.1 Review of Findings

This study has provided a first look at how raters evaluated the same two samples of speech in the NATO context, and compared the ratings not only from rater-to-rater, but also across a representative sample of countries, through quantitative and qualitative lenses. Findings from the data analysis reported in Chapter 4, revealed that there were differences in ratings for both OPI samples. The following summary highlights the most important findings, which will be further discussed.

Sample A was initially rated by 58% of participants as a **level 1**, while 41 % rated in the **level 2** range, and 1% in the **level 3** range. Scores adjusted with plus levels became 68% in the **level 1 range,** 31% in the **level 2 range,** with 1% in the **level 3 range.**

Sample B was initially rated by 3% of raters in the **level 1** range, 57% raters placed this sample in the **level 2** range while a total of 37% raters placed sample B within the **level 3** range, with a last 2% who perceived this examinee to be in the **level 4** range. Scores adjusted with plus levels became 2% in the **level 1 range,** 59% in the **level 2 range,** with 31% in the **level 3 range.**

It was clear from the findings that the addition of plus levels improved inter-rater reliability for both sample A and sample B, and especially with sample A. The proficiency of the examinee in sample A was considered to be high in the level 1 range, therefore it was expected that plus levels would provide the raters in this study a scoring possibility that was closer to the observations they reported about the candidate's performance.

Differences across all raters were not as pronounced as they were across all countries, but it became clear that country means in some cases were affected by discrepant ratings within a country. That is, in some countries, there was larger dispersion of scores within country, than across countries. So indeed, it can be stated that although inter-country ratings differed, they differed no more than intra-country ratings overall.

The next group of research questions aimed to see what differences relating to training and experience existed among raters, and to see how differences may impact the ratings that were assigned. Training was the first factor explored, both training as tester (and therefore as rater) and training in the interpretation of the scale. Tester training was a key factor in rating 'correctly', and although both groups provided similar ratings for sample A, very few of the lesser-trained group of participants rated sample B 'correctly'. STANAG training was also important, and a surprising number of participants reported never having received training in the interpretation of the scale.

The findings of the study indicated that the number of years of experience as tester, and the number of tests conducted per year, did not enable testers to score more accurately. It was clear however, that participating raters who test full-time performed considerably better than part-time testers.

Comparisons were drawn between the native speaker and the non-native speaker participants, and no major differences were found, although it was surprising to see that rating sample B presented more challenges to the native-speakers of English. This was further investigated, and one of the most striking differences appeared to be the fact that non-native speakers reported having received more training than the native speaker group. Comparisons between 'old' NATO countries and 'new' NATO countries yielded similar findings. The newer NATO country participants rated more accurately than the participants from older NATO countries, a factor which can once again be attributed to training received. It became clear that participants from the new NATO countries reported having received considerably more tester/rater training, as well as scale training.

The study also investigated the rating processes that participants reported having followed to arrive at their ratings. The steps taken to arrive at the ratings varied from rater to rater slightly, with some using converted scores, or averaged scores. None of these were deemed significant in impacting the results, although the raters who indicated that they had consulted the STANAG scale to assist in the rating, rated sample A 'correctly'. The excerpts of the participants'

questionnaires allowed for a view into the reasons behind the ratings assigned. Findings showed that although most ratings were 'evidence-based', some were 'intuitive' and others were 'extra-contextual' as defined in Chapter 4. In some cases, raters brought external concerns into the assessment process.

Finally, comments from participants supported the addition of plus levels into the framework to help bridge the gap between levels of proficiency. Participants also reported that adding supplementary explanations would be useful in further clarifying some descriptors in the scale.

5.2 Implications

What are we testing, and why? Chapter 2 outlined that one of the features most attractive and salient for government purposes is the flexibility of the general proficiency orientation to accommodate the evaluation of the communicative competence military members have, in any given language, without being concerned with how one has acquired the language. While the focus of the present study was on English, from 5 to 25 foreign languages are presently taught in many military institutions, and in most cases, these languages are also tested with the STANAG scale[40]. If the purpose of reporting language ability in governments, in the military and in NATO is to ensure interoperability, mutual intelligibility and generalisable skill, then the literature surveyed supported a general proficiency approach.

[40] See the BILC Website (in references) 'National Reports' section for further information.

In some institutions, it appears that the full STANAG scale is not in use. Participants from these systems, when answering the rater data questionnaire, pointed out that they had received no tester training and no STANAG scale interpretation training. Instead, they indicated that test observations were used as the method of acculturation into the testing role for teachers. Testers in these systems rely on the testing grids elaborated by their schools or institutes, based on the STANAG scale. This is the criterion against which a candidate's proficiency is assessed. The many variations in evaluation grids pointed to the possibility that the original construct of the scale was being reinterpreted in ways that may influence the perception of proficiency levels, as outlined in STANAG 6001. Perceiving rating grids as the only criteria or as the final criteria, or not referring to the full STANAG scale (and/or never having seen the full STANAG), could be viewed as problematic, since it may be introducing variance into the testing framework. Not understanding what is meant by tasks or functions, not being concerned with accuracy statements, or not perceiving the STANAG scale as containing statements referring to sociolinguistic aspects of communication, are all issues which could be resolved with training.

As presented in the literature review Chapter, Lumley and McNamara (1993) affirmed that rater training could reduce, but not eliminate rater variability because it is very difficult to alter raters' standards and perceptions. They state that the results of training may not last for long after a training session, therefore justifying the need for frequent moderation sessions, preferably before each test

administration. Some countries have a separate testing system whereby testers are not involved in teaching activities, but focus exclusively on test development and testing activities. It would appear to be the case, according to the findings of this study, that setting up a separate testing system, where testers focus on testing duties exclusively, may have a positive impact on ratings. In systems where teachers are asked to test and rate the speaking ability, it becomes crucial that they retrain regularly, especially in testing systems where raters' services are not used for long periods of time, or only used sporadically. Raters need to be monitored, and should be given regular feedback on their ratings. As Wigglesworth (1993) concluded from her study on the effect of providing regular feedback to raters, while comments do get put into practice, it is not clear whether or not the adjustments made by the raters post-feedback have long lasting effects, or if raters revert to previous patterns after time has lapsed. Weigle's (1998) research has led her to conclude that certain raters perhaps should not be used at all as raters, a conclusion which is also supported by McNamara (1996).

The findings in this study concur with the findings of Brown's (2003) research on non-native speakers as raters, namely that nothing would suggest that non-native speakers are any less suitable than native speakers, if raters are provided with acceptable training and unambiguous assessment criteria. The newer NATO countries have mostly non-native speaker testers/raters, and in this study, they reported having received the most training. The findings showed that the native speaker group was the most experienced group of participants, however, findings

demonstrated that the native speaker group had not received as much training, as the non-native speaker group.

It was clear from the information provided by the participants that there are a number of differing types of oral interviews being given in the different countries. Test method is a known source of variance in testing (Bachman 1990, 1995, 2002; Shohamy 1983) and although every country's speech samples collected are rated against the same scale, there are obviously many variables in the NATO framework.

Also, in some organisations, speaking tests are not recorded and are never rated individually, that is, committee ratings form the norm. As reviewed in Chapter 2, Henning (1996) comments that situations where the two raters misjudge true ability is infrequent, however, individual ratings provide additional 'insurance' that raters have perceived the performance in the same way. If there is no record of the interview, one assumes that if a student wishes to contest a score received, another objective look at the speaking test cannot be made.

The topic of how plus levels should be interpreted and used is a question that warrants discussion. A number of comments were collected pertaining to the width of the levels in the STANAG scale, and can be regarded as speaking to two issues, score reliability and fairness. Some raters indicated that if examinees are at the higher end of the level or band, they may feel compelled to edge them into

the next higher level (even if the examinee does not meet the criteria for the next level) if they perceive it to be 'unfair' that two examinees with such differing levels of proficiency, (e.g. a threshold level 1 and a high level 1) should receive the same score. In most systems where plus levels are presently in use, it appears that these are not considered as separate levels, but are seen as proficiency which does not meet all of the criteria of the next higher level. Regardless of how pluses are presently understood, in this study, they helped to bring ratings closer to the mean. A definition of the framework of the plus levels, as well as descriptors have been provided by a BILC committee and accepted by the BILC Steering Committee in 2006. This should facilitate a common understanding, but as reported in Chapter 4, training in their interpretation and their application into the various existing testing frameworks, from country to country may prove indispensable.

5.3 Limitations of the Study and Future Research

The data represented a large number of participants and many NATO countries; therefore it may be argued that the findings adequately represent the rating population within NATO. However, because of the great diversity of speaking tests in use among the various countries, one cannot infer that a 'correct' rating (or a rating within the mean with these samples) indicates that the same candidate would have achieved this rating with another type of speaking test, somewhere else. It was reported by some participants that rating this particular type of

speaking test was at times challenging. Some countries use single level tests, where only one of the STANAG proficiency levels is tested at a time, such as during a STANAG level 1 test, a STANAG level 2 test, etc. As a result, some of the participating raters were unsure how to rate language when it is elicited through a multi-level test, where proficiency levels are not assumed *a priori.* Furthermore, it cannot be supposed that because the participants rated the OPIs in this study in a particular way, that they would rate similarly in their own testing systems. Raters were instructed to use their own country's rating protocol to rate the samples so that no new protocol, unknown to them, would be introduced. Although there was no viable alternative, this, in of itself may have introduced more complexity than needed. It was evident that participants rated to the best of their ability, and used whatever means were at their disposal. One can only state that these participants, regardless of how their own systems operate, rated these oral proficiency interviews.

Conducting a survey was a valid method to collect data seeing as there was a large geographical area that needed to be covered, however, this method also prohibited further investigation and confirmation of participants' views and comments. Had it been possible to follow-up the survey with participant interviews, or verbal reports, it would have enriched the study's findings. Some interpretation of participants' comments was necessary at times, and although this was carefully conducted in order to avoid introducing personal bias into the study, it always remains a possibility. In such cases, there is always the chance that

what participants wrote down did not reflect what they actually thought, or that the researcher became a subjective interpreter of the intended message. Surveys unfortunately do not provide the means for further verification.

Lastly, participants were not required to test the speaking skill, only to rate the samples provided. Having participants actually test the same candidate may yield different findings.

The data in this study was looked at from many different angles, with various types of analyses, but unfortunately, there were still some facets left unexplored.

All raters were asked to insert their rating grids into the envelope provided. Unfortunately, time (and space!) constraints did not permit a full analysis of these grids, or the inclusion of observations collected by the researcher. As 50% of the raters in the study provided their rating grids, with rating grids from 15 separate countries included, a comparative analysis of these instruments in relation to the STANAG scale could be done in the future. This may allow for a closer analysis of the potentially modified constructs imbedded in the rating scale, and explain some of the discrepant views and interpretations of levels.

This study only scratched the surface of language testing practices among NATO countries. All four skills tested within the NATO framework should be the focus of future research. How do ratings compare in other tests? How do tasks in

writing tests compare country to country, and what are the rating criteria? How are texts chosen for reading and listening comprehension tests, what sub-skills are being tested, what are the methods employed? All of these issues could be explored empirically. Findings of such research could be made available to inform and benefit practitioners and stakeholders.

5.4 Concluding Remarks

The findings of this study and the issues discussed above, pose considerable challenges for international benchmarking, but they are far from insurmountable. So, are we all on the same page? Generally speaking, it appears to be the case. When a majority of raters across different countries rate language samples consistently, it is an indication that these raters share the same interpretation. It also presupposes that the common metric is a useful criterion on which to base evaluations. However, it has been clearly demonstrated that nothing can compensate for a lack of investment in rater training. Rating scales do not stand alone, and the need to norm to a common perception of the standard, is supported by the research literature in language testing. This study has provided some evidence that more training is needed within individual countries to ensure testers/raters in each testing system share a consistent and reliable scale interpretation. This is arguably the necessary first step in bringing countries to a common perception of the levels, and in achieving international benchmarking.

REFERENCES

Alderson, C.J. (1993). Judgements in language testing. In D. Douglas and C. Chapelle (eds), *A new decade of language testing research: Selected papers from the 1990 language testing colloquium* (pp.46-57). Virginia World Composition Services, Inc.

Alderson, C.J. and Wall, D. (1993) Does Washback exist? *Applied Linguistics* 14 (2), 115-129.

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

Bachman, L.F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing, 17 (1),* 1-42.

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

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

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

Berwick, R. & S. Ross (1996). Cross-cultural pragmatics in oral proficiency assessment. In M. Milahovic and N. Saville (eds.), *Performance testing, cognition and assessment.* Cambridge: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

BILC Website http://www.dlielc.org/bilc/Constitution2004.doc

Brindley, G. (1998). "Describing Language Development? Rating Scales and SLA" in Bachman, L.F. & Cohen, A.D. (eds.) *Interfaces Between Second Language Acquisition and Language Testing Research.* Cambridge: Cambridge University Press.

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

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

Brown, J. D. and Rodgers, T.S. (2002). *Doing Second Language Research.* Oxford University Press.

Brown, J. D, T. Hudson, J. Norris and W. Bonk (2002) *An Investigation of Second Language Task-Based Performance Assessments.* University of Hawaii Press.

Byrnes, H. (1987). Second Language Acquisition: Insights from a Proficiency Orientation. In Byrnes, H. & M. Canale (Eds.), *Guidelines, Implementations and Concepts.* Lincolnwood, Illinois, National Textbook Company.

Canadian Forces Language School (2001). CFLS Foreign Language Testing Directive. Internal Document.

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

Canale, M. & M. Swain (1980) Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing, *Applied Linguistics*, 11, 1-47.

Canale, M. & M. Swain (1981). A theoretical framework for communicative competence. In A. Palmer, P. Groot and G. Trosper (Eds.) *The Construct Validation of Tests of Communicative Competence.* TESOL, 31-36.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12,* 16-33.

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In Bachman, L.F. & A. D. Cohen (Eds.), *Interfaces Between Second Language Acquisition and Language Testing Research* (pp.32-70). Cambridge: Cambridge University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge (Mass): The MIT Press.

Clapham, C. (1993). Is ESP Testing Justified? In D. Douglas and C. Chapelle (Eds), *A new decade of language testing research: Selected papers from the 1990 language testing colloquium* (pp.257-271). Virginia World Composition Services, Inc.

Clark, J. L. D. (1986). *A study of the comparability of speaking interview ratings across three government language training agencies.* Washington, D.C: Center for Applied Linguistics. (ERIC Document Reproduction Service No. ED 267 645.

Clark, J. L. & R. T. Clifford (1988). The FSI/ILR/ACTFL Proficiency scales and testing techniques: Development, current status, and needed research. *Studies in Second Language Acquisition, 10,* 129-147.

Clifford, R. T. (1981). Convergent and discriminant validation of integrated and unitary language skills: the need for a research model. In A. Palmer, P. Groot and G. Trosper (Eds.) *The Construct Validation of Tests of Communicative Competence.* TESOL, 62-70.

Crossey, M. (2005). Improving linguistic interoperability. *NATO Review.* Retrieved from www.nato.int/docu/review/2005/issue2/english/art4.html June 30, 2005.

Dandonoli, P. (1987). ACTFL's Current Research in Proficiency Testing. In Byrnes, H. & M. Canale (Eds.), *Guidelines, Implementations and Concepts* (pp.75-96). Lincolnwood, Illinois, National Textbook Company.

Davies, A. (2001). The logic of testing Languages for Specific Purposes. *Language Testing, 18,* 133-147.

Davies, A. (2003). *The Native Speaker: Myth and Reality.* Multilingual Matters.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

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

Douglas, D. (2005). *Testing Languages for Specific Purposes.* In E. Hinkel, (Ed.), *Handbook of research in second language teaching and learning* (pp. 857-868), Mahway, NJ: Lawrence Erlbaum.

Douglas, D. & L. Selinker (1992). Analysing Oral Proficiency Test Performance in General and Specific Purpose Contexts. *Systems,* 20(3), 317-328.

Douglas, D. & L. Selinker (1993). Performance on a General versus a Field-Specific Test of Speaking Proficiency by International Teaching Assistants. In D. Douglas and C. Chapelle (Eds), *A new decade of language testing research: Selected papers from the 1990 language testing colloquium* (pp.235-245). Virginia World Composition Services, Inc.

Fox, J. (2003). From Products to Process: An Ecological Approach to Bias Detection. *International Journal of Testing,* Vol.3 (1), 21-47.

Fox, J., T. Pychyl and B. Zumbo. (1993). Psychometric properties of the CAEL Assessment, I: an overview of development, format, and scoring procedures. In Fox (ed.), *Carleton Papers in Applied Language Studies,* Volume X. Ottawa: Centre for Applied Language Studies, Carleton University.

Green, R. & D. Wall. (2005) Language testing in the military: problems, politics and progress. *Language Testing, 22(3),* 379-398.
Henning, G. (1992). The ACTFL oral proficiency interview: Validity evidence. *System, 20,* 365-372.

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

Herzog, M. (n.d.). An overview of the history of the ILR language proficiency skill level descriptions and scale. *Interagency Language Roundtable: History of the ILR Scale.* Retrieved from www.govtilr.org/ILRscale_hist.html, October 14, 2005.

Hill, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio and S. Luoma (Eds.), *Current Developments and Alternatives in Language Assessment-Proceedings of LTRC 96,* 275-290.

Higgs, T.V. and R.T. Clifford (1982). The push toward communication. In Higgs, T. V. (Ed.) *Curriculum, Competence, and the Foreign Language Teacher.* (pp. 57-79) The ACTFL Foreign Language Education Series, Vol. 13. Skokie, Ill.: National Textbook Company.

Hymes, D. H. (1972). On communicative competence. In Pride, J. B. and J. Holmes, (eds) *Sociolinguistics: selected readings.* Penguin, Harmondsworth, Middlesex, 269-293.

Interagency Language Round Table. (1985). *ILR skill level descriptions.* Washington, DC.: IRT.

ILR Handbook on Oral Interview Testing (1982).

Jenkins, S. and I. Parra (2003). Multiple Layers of Meaning in an Oral Proficiency Test: The Complementary Roles of Nonverbal, Paralinguistic, and Verbal Behaviors in Assessment Decisions. *The Modern Language Journal,* 87, i, 890-107.

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

Kane, M., T. Crooks, & A. Cohen. (1999) Validating measures of performance. *Educational Measurement: Issues and Practice,* 5-12.

Kenyon, D. (2000). The Rating of Direct and Semi-Direct Oral Proficiency Interviews: Comparing Performance at Lower Proficiency Levels. *The Modern Language Journal*, 84, i, 85-101.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing,* 19(1), 3-31.

Lazaraton, A. (1996a). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing,* 13(2), 151-172.

Llurda, E. (2004). Non-native-speaker teachers and English as an International Language. *International Journal of Applied Linguistics,* 14(3) 315-323.

Lowe, P. Jr. (1980). Structure of the oral interview and content validity. In A. Palmer, P. Groot and G. Trosper (Eds.) *The Construct Validation of Tests of Communicative Competence.* TESOL, 71-80.

Lowe, P. Jr. (1985). The ILR Proficiency Scale as a Synthesizing Research Principle: The View from the Mountain. In C. J. James (Ed.) *Foreign Language Proficiency in the Classroom and Beyond* (pp. 9-53). Lincolnwood, Illinois, National textbook Company.

Lowe, P. (1988). The unassimilated history. In P. Lowe & C. Stansfield (eds.), *Second language proficiency assessment: Current issues* (pp.11-51). Englewood Cliffs, NJ: Prentice Hall/Regents.

Lumley, T. (1998). Perceptions of language-trained raters and occupational experts in a test of occupational English language proficiency. *English for Specific Purposes, Vol.17, No.4,* 347-367.

Lumley, T. & T.F. McNamara (1993). Rater characteristics and rater bias: Implications for training. Cambridge, England: 15th Language Testing Research Colloquium, (ERIC Document Reproduction Service No. ED 365 091).

Luoma, S. (2004) Assessing Speaking. Cambridge: Cambridge University Press.

NATO FORCE PROPOSAL (2004). Example PGG03551.

NATO STANAG 6001, Ed.2, (2003), Retrieved www.dlielc.org/bilc/Sta_Edit2, November 4, 2005.

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

McNamara, T. (2000). *Language testing.* Oxford, U.K.: Oxford University Press.

McNamara, T., K. Hill, and L. May (2002). Discourse and Assessment. *Annual review of Applied linguistics, 22,* 221-242.

Messick, S. (1990) Validity of Test Interpretation and Use. Educational Testing Service. Princeton, New Jersey.

Messick, S. (1994) The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher,* 23, (2),13-23.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang Publishing Inc.

North, B. & G. Schneider. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15,* 217-263.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

OPI 2000 Tester Certification Workshop (1999).

Partnership for Peace (PfP) Planning and Review Process (PARP) Presentation, Bureau for International Language Co-ordination (BILC) Sofia, 2005.

Public Service Commission of Canada, (1993). *Assessing for Competence: Determining the Linguistic Profile for Bilingual Positions.* Government of Canada, 1-19.

Pollitt, A., & N. Murray. (1996). What raters really pay attention to. In M. Milahovic and N. Saville (eds.), *Performance testing, cognition and assessment.* Cambridge: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Ross, S. & R. Berwick. (1992). The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*(2), 159-176.

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

Shohamy, E. (1980). Inter-rater and intra-rater reliability of the oral interview and concurrent validity with cloze procedure. In A. Palmer, P. Groot and G. Trosper (Eds.) *The Construct Validation of Tests of Communicative Competence.* TESOL, 94-105.

Shohamy, E. (1983). The stability of oral proficiency assessment on the oral interview testing procedures. *Language Learning, 33*(4), 527-540.

Shohamy , E. (1996). Language testing: Matching assessment procedures with language knowledge, in M. Birenbaum & F. Dochy (eds.), *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge,* Kluwer Academic Publishers, Boston, MA, 143-159.

Shohamy, E. (1998). How can language testing and SLA benefit from each other? The case of discourse. In Bachman, L.F. & Cohen, A.D. (Eds.), *Interfaces Between Second Language Acquisition and Language Testing Research* (pp.156-176). Cambridge: Cambridge University Press.

Shohamy, E. (2000). The relationship between language testing and second language acquisition, revisited. *System, 28,* 541-553.

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

Spolsky (1973). What does it mean to know a language? Or, how do you get someone to perform his competence? In J. Oller and J. Richards (eds.), *Focus on the learner: pragmatic perspectives for the language teacher.* Rowley, MA: Newbury House, pp. 164-176.

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

Teasdale, A. (1996). Content validity in tests for well-defined LSP domains: an approach to defining what is to be tested. In M. Milahovic and N. Saville (Eds.), *Performance testing, cognition and assessment.* (pp. 211-230). Cambridge: University of Cambridge Local Examinations Syndicate/Cambridge University Press.

Upshur, J. A., & C. Turner. (1995). Constructing rating scales for second language tests. *ELT journal, 49,* 3-12.

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

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15(2),* 263-287.

Wesche, M. B. (1992) Performance testing for work-related second language assessment. In E. Shohamy and R. Walton (eds) *Language Assessment for Feedback: Testing and Other Strategies.* Dubuque, IA: Kendall/Hunt Publishing Company.

Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing*, 10(3), 305-336.

Wilds, C. P. (1975). The oral interview test. In: R. Jones & B. Spolsky (Eds.), *Testing language proficiency* (pp. 29-44). Arlington, VA: Center for Applied Linguistics.

Appendix A to Julie Dubeau's M.A. Thesis

STANAG 6001
(Edition 2)

NATO STANDARDIZATION AGREEMENT
(STANAG)

LANGUAGE PROFICIENCY LEVELS

Related Documents: None

Annex A: Table of Language Proficiency Levels

AIM

1. The aim of this agreement is to provide NATO Forces with a table describing language proficiency levels.

AGREEMENT

2. Participating nations agree to adopt the table of language proficiency levels for the purpose of :

a. Meeting language requirements for international staff appointments.

b. Comparing national standards through a standardised table.

c. Recording and reporting, in international correspondence, measures of language proficiency (if necessary by conversion from national standards).

GENERAL

3. The descriptions at Annex A give detailed definitions of the proficiency levels in the commonly recognised language skills: oral proficiency (listening and speaking) and written proficiency (reading and writing).

PROFICIENCY LEVELS

4. The proficiency skills are broken down into six levels coded 0 through 5. In general terms, skills may be defined as follows:

Level 0		No practical proficiency
Level 1	-	Elementary
Level 2	-	Fair (Limited working)
Level 3	-	Good (Minimum professional)
Level 4	-	Very good (Full professional)
Level 5	-	Excellent (Native/bilingual)

LANGUAGE PROFICIENCY PROFILE

5. Language proficiency will be recorded with a profile of 4 digits indicating the specific skills in the following order:

Skill A (US : L[41])	Listening
Skill B (US : S)	Speaking
Skill C (US : R)	Reading
Skill D (US : W)	Writing

6. This number of 4 digits will be preceded by the code letters SLP (PLS in French) which is to indicate that the profile shown is the Standardised (S) Language (L) Profile (P). (Example: SLP 3321 means level 3 in listening, level 3 in speaking, level 2 in reading and level 1 in writing).

IMPLEMENTATION OF THE AGREEMENT

7. This STANAG will be considered implemented when a nation has issued the necessary orders/instructions to adopt the table and to put into effect the procedures detailed in this agreement.

[41] The Code letters (US: ...) is for the use of the United States.

ANNEX A to
STANAG 6001
(Edition 2)

TABLE OF LANGUAGE PROFICIENCY LEVELS

ORAL PROFICIENCY SKILL

<table>
<tr><th>Level</th><th colspan="2">A (US:L) LISTENING</th><th>B (US:S) SPEAKING</th></tr>
<tr><td>0 -</td><td colspan="2">No practical proficiency</td><td>No practical proficiency</td></tr>
<tr><td>1 -</td><td colspan="3">Elementary

Vocabulary: Adequate for routine courtesy and minimum practical needs related to travelling, obtaining food and lodging, giving simple directions, asking for assistance.</td></tr>
<tr><td></td><td colspan="2">Listening comprehension: Adequate for very simple short sentences in face-to-face situations. May require much repetition and a slow rate of speech. Fails in situations where there is noise or other interference.</td><td>Grammar and pronunciation: Errors are frequent and may often cause misunderstanding.

Fluency: Adequate for memorised courtesy expressions and common utterances. Otherwise lacking.</td></tr>
<tr><td>2 -</td><td colspan="3">Fair (Limited Working)

Vocabulary: Adequate for simple social and routine job needs as giving instructions and discussing projects within very familiar subject-matter fields. Word-meanings often unknown, but quickly learned.</td></tr>
</table>

Level	A (US:L) LISTENING	B (US:S) SPEAKING
	Listening comprehension: Dependable in face-to-face communication within well-known subject-matter fields and in common social contexts. Sometimes requires rewording or slowing of conversational speed. Incomplete in the presence of noise or other interference. Seldom adequate to follow a conversation between two native speakers.	Grammar and pronunciation: Meaning is accurately expressed in simple sentences. Circumlocution often needed to avoid complex grammar. Foreign-sounding pronunciation very noticeable but usually does not interfere with intelligibility. Fluency: Often impaired by hesitation and groping for words.

<table>
<tr><th>Level</th><th>A (US:L) LISTENING</th><th>B (US:S) SPEAKING</th></tr>
<tr><td>3 -</td><td colspan="2">Good (Minimum Professional)

Vocabulary: Adequate for all practical and social conversations and for professional discussions in a known field.</td></tr>
<tr><td></td><td>Listening comprehension: Adequate to follow radio broadcasts, speech conversations between two educated native speakers in the standard language. Details and regional or dialectic forms may be missed, but general meaning is correctly interpreted.</td><td>Grammar and pronunciation: Full range of basic structures well understood, and complex structures used. Mistakes sometimes occur, but meaning accurately conveyed. Pronunciation recognisably foreign but never interferes with intelligibility.

Fluency: Rarely impaired by hesitations. Flow of speech is maintained by circumlocution when necessary. There is no groping for words.</td></tr>
<tr><td>4 -</td><td colspan="2">Very Good (Full Professional)

Vocabulary: Broad, precise, and appropriate to the subject and the occasion.</td></tr>
<tr><td></td><td>Listening comprehension: Adequate for all educated standard speech in any situation. Undisturbed by noise or interference in moderate amount. May occasionally have difficulty with colloquial or regional dialect variations.</td><td>Grammar and pronunciation: Errors seldom occur, and do not interfere with accurate expression of meaning. Non-native speaker pronunciation does not interfere with intelligibility.

Fluency: Similar to native fluency in known subject fields. Easy for a native speaker to listen to.</td></tr>
<tr><td></td><td colspan="2">NOTE: This level reflects extensive experience using the language in an environment where it is the primary means of communication.</td></tr>
</table>

Level	A (US:L) LISTENING	B (US:S) SPEAKING
5 -	Excellent (Native/bilingual) In all criteria of language proficiency, completely equal to a native speaker of the language. This level of proficiency is not achieved by training, and cannot normally be attained except by natives who have been educated through the secondary level in indigenous schools.	

WRITTEN PROFICIENCY SKILL

Level	C (US:R) READING COMPREHENSION	D (US:W) WRITING
0 -	No practical proficiency	No practical proficiency
1 -	Elementary Adequate for street signs, public directions, names on buildings, and elementary lesson material. In languages written by alphabet or syllabify, adequate to spell out unknown words and approximate their pronunciation in order to ask a native speaker the meaning.	Has sufficient control of the writing systems to meet limited practical needs. Can produce all symbols in an alphabetic or syllabic writing system. Can write numbers and dates, his own name and nationality, addresses, etc. Otherwise ability to write is limited to simple lists of common items or a few short sentences. Spelling may be erratic.
2 -	Fair (Limited working) Adequate for intermediate lesson material and simple colloquial texts such as children's books. Requires extensive use of dictionary to read short news items. Written material seldom fully understood without translation.	Can draft routine social correspondence and meet limited professional needs. Is familiar with the mechanics of the writing system, except in character systems where ability is limited to a small stock of high-frequency items. Makes frequent errors in spelling, style and writing conventions. Able to write simple notes and draft routine social and limited office messages. Material normally requires editing by a more highly proficient writer.

Level	C (US:R) READING COMPREHENSION	D (US:W) WRITING
3 -	Good (Minimum professional) Adequate for standard text materials and most technical material in a known professional field; with moderate use of dictionary, adequate for most news items about social, political, economic, and military matters. Information is obtained from written material without translation.	Can draft official correspondence and reports in a special field. Control of structure, spelling, and vocabulary is adequate to convey his message accurately, but style may be quite foreign. All formal writing needs to be edited by an educated native.
4 -	Very Good (Full professional) Adequate to read easily and with minimal use of dictionaries, styles of the language occurring in books, magazines and newspapers written for an audience educated to the level of a high school graduate. Adequate to read technical and abstract material in known professional fields.	Can draft all levels of prose pertinent to professional needs. Control of structure, vocabulary, and spelling is broad and precise; sense of style is nearly native. Errors are rare and do not interfere with understanding. Nevertheless, drafts or official correspondence and documents need to be edited by an educated native.
	NOTE: This level reflects extensive experience using the language in an environment where it is the primary means of communication.	
5 -	Excellent (Native/bilingual) In all criteria of language proficiency, completely equal to a native speaker of the language. This level of proficiency is not achieved by training, and cannot normally be attained except by natives who have been educated through the secondary level in indigenous schools.	

Appendix 1 Interpretation of Language Proficiency Levels

Appendix 1 to
Annex A to
STANAG 6001
(Edition 2)

INTERPRETATION OF THE LANGUAGE PROFICIENCY LEVELS

LEVEL 0 (NO PROFICIENCY)

LISTENING COMPREHENSION

1. No practical understanding of the spoken language. Understanding is limited to occasional isolated words. No ability to comprehend communication.

SPEAKING

2. Unable to function in the spoken language Oral production is limited to occasional isolated words such as greetings or basic courtesy formulae. Has no communicative ability.

READING COMPREHENSION

3. No practical ability to read the language. Consistently misunderstands or cannot comprehend the written language at all.

WRITING

4. No functional writing ability.

LEVEL 1 (ELEMENTARY)

LISTENING COMPREHENSION

5. Can understand common familiar phrases and short simple sentences about everyday needs related to personal and survival areas such as minimum courtesy, travel, and workplace requirements when the communication situation is clear and supported by context. Can understand concrete utterances, simple questions and answers, and very simple conversations. Topics include basic needs such as meals, lodging, transportation, time, simple directions and instructions. Even native speakers used to speaking with non-natives must speak slowly and repeat or reword frequently. There are many misunderstandings of both the main idea and supporting facts. Can only understand spoken language from the media or among native speakers if content is completely unambiguous and predictable.

SPEAKING

6. Able to maintain simple face-to-face communication in typical everyday situations. Can create with the language by combining and recombining familiar, learned elements of speech. Can begin, maintain, and close short conversations by asking and answering short simple questions. Can typically satisfy simple, predictable, personal and accommodation needs; meet minimum courtesy, introduction, and identification requirements; exchange greetings; elicit and provide predictable, skeletal biographical information; communicate about simple routine tasks in the workplace; ask for goods, services, and assistance; request information and clarification; express satisfaction, dissatisfaction, and confirmation. Topics include basic needs such as ordering meals, obtaining lodging and transportation, shopping. Native speakers used to speaking with non-natives must often strain, request repetition, and use real-world knowledge to understand this speaker. Seldom speaks with natural fluency, and cannot produce continuous discourse, except with rehearsed material. Nonetheless, can speak at the sentence level and may produce strings of two or more simple, short sentences joined by common linking words. Frequent errors in pronunciation, vocabulary, and grammar often distort meaning. Time concepts are vague. May often use only one tense or tend to avoid certain structures. Speech is often characterised by hesitations, erratic word order, frequent pauses, straining and groping for words (except for routine expressions), ineffective reformulation, and self-corrections.

READING COMPREHENSION

7. Can read very simple connected written material, such as unambiguous texts that are directly related to everyday survival or workplace situations. Texts may include short notes; announcements; highly predictable descriptions of people, places, or things; brief explanations of geography, government, and currency systems simplified for non-natives; short sets of instructions and directions (application forms, maps, menus, directories, brochures, and simple schedules). Understands the basic meaning of simple texts containing high frequency structural patterns and vocabulary, including shared international terms and cognates (when applicable). Can find some specific details through careful or selective reading. Can often guess the meaning of unfamiliar words from simple context. May be able to identify major topics in some higher level texts. However, may misunderstand even some simple texts.

WRITING

8. Can write to meet immediate personal needs. Examples include lists, short notes, post cards, short personal letters, phone messages, and invitations as well as filling out forms and applications. Writing tends to be a loose collection of sentences (or fragments) on a given topic, with little evidence of conscious organization. Can convey basic intention by writing short, simple sentences, often joined by common linking words. However, errors in spelling, vocabulary, grammar, and punctuation are frequent. Can be understood by native readers used to non-natives' attempts to write.

LEVEL 2 (LIMITED WORKING)

LISTENING COMPREHENSION

9. Sufficient comprehension to understand conversations on everyday social and routine job-related topics. Can reliably understand face-to-face speech in a standard dialect, delivered at a normal rate with some repetition and rewording, by a native speaker not used to speaking with non-natives. Can understand a wide variety of concrete topics, such as personal and family news, public matters of personal and general interest, and routine work matters presented through descriptions of persons, places, and things; and narration about current, past, and future events. Shows ability to follow essential points of discussion or speech on topics in his/her special professional field. May not recognise different stylistic levels, but recognises cohesive devices and organising signals for more complex speech. Can follow discourse at the paragraph level even when there is considerable factual detail. Only occasionally understands words and phrases of statements made in unfavorable conditions (for example, through loudspeakers outdoors or in a highly emotional situation). Can usually only comprehend the general meaning of spoken language from the media or among native speakers in situations requiring understanding of specialised or sophisticated language. Understands factual content. Able to understand facts but not subtleties of language surrounding the facts.

SPEAKING

10. Able to communicate in everyday social and routine workplace situations. In these situations the speaker can describe people, places, and things; narrate current, past, and future activities in complete, but simple paragraphs; state facts; compare and contrast; give straightforward instructions and directions; ask and answer predictable questions. Can confidently handle most normal, casual conversations on concrete topics such as job procedures, family, personal background and interests, travel, current events. Can often elaborate in common daily communicative situations, such as personal and accommodation-related interactions; for example, can give complicated, detailed, and extensive directions and make non-routine changes in travel and other arrangements. Can interact with native speakers not used to speaking with non-natives, although natives may have to adjust to some limitations. Can combine and link sentences into paragraph-length discourse. Simple structures and basic grammatical relations are typically controlled, while more complex structures are used inaccurately or avoided. Vocabulary use is appropriate for high-frequency utterances but unusual or imprecise at other times. Errors in pronunciation, vocabulary, and grammar may sometimes distort meaning. However, the individual generally speaks in a way that is

appropriate to the situation, although command of the spoken language is not always firm.

READING COMPREHENSION

11. Sufficient comprehension to read simple authentic written material on familiar subjects. Can read straightforward, concrete, factual texts, which may include descriptions of persons, places, and things; and narration about current, past, and future events. Contexts include news items describing frequently recurring events, simple biographical information, social notices, routine business letters, and simple technical material intended for the general reader. Can read uncomplicated but authentic prose on familiar subjects that are normally presented in a predictable sequence that aids the reader in understanding. Can locate and understand the main ideas and details in material written for the general reader and can answer factual questions about such texts. Cannot draw inferences directly from the text or understand the subtleties of language surrounding factual material. Can readily understand prose that is predominately constructed in high frequency sentence patterns. While active vocabulary may not be broad, the reader can use contextual and real-world cues to understand texts. May be slow in performing this task, and may misunderstand some information. May be able to summarise, sort, and locate specific information in higher level texts concerning his/her special professional field, but not consistently or reliably.

WRITING

12. Can write simple personal and routine workplace correspondence and related documents, such as memoranda, brief reports, and private letters, on everyday topics. Can state facts; give instructions; describe people, places, and things; can narrate current, past, and future activities in complete, but simple paragraphs. Can combine and link sentences into connected prose; paragraphs contrast with and connect to other paragraphs in reports and correspondence. Ideas may be roughly organised according to major points or straightforward sequencing of events. However, relationship of ideas may not always be clear, and transitions may be awkward. Prose can be understood by a native not used to reading material written by non-natives. Simple, high frequency grammatical structures are typically controlled, while more complex structures are used inaccurately or avoided. Vocabulary use is appropriate for high frequency topics, with some circumlocutions. Errors in grammar, vocabulary, spelling, and punctuation may sometimes distort meaning. However, the individual writes in a way that is generally appropriate for the occasion, although command of the written language is not always firm.

LEVEL 3 (MINIMUM PROFESSIONAL)

LISTENING COMPREHENSION

13. Able to understand most formal and informal speech on practical, social, and professional topics, including particular interests and special fields of competence. Demonstrates, through spoken interaction, the ability to effectively understand face-to-face speech delivered with normal speed and clarity in a standard dialect. Demonstrates clear understanding of language used at interactive meetings, briefings, and other forms of extended discourse, including unfamiliar subjects and situations. Can follow accurately the essentials of conversations among educated native speakers, lectures on general subjects and special fields of competence, reasonably clear telephone calls, and media broadcasts. Can readily understand language that includes such functions as hypothesising, supporting opinion, stating and defending policy, argumentation, objections, and various types of elaboration. Demonstrates understanding of abstract concepts in discussion of complex topics (which may include economics, culture, science, technology) as well as his/her professional field. Understands both explicit and implicit information in a spoken text. Can generally distinguish between different stylistic levels and often recognises humor, emotional overtones, and subtleties of speech. Rarely has to request repetition, paraphrase, or explanation. However, may not understand native speakers if they speak very rapidly or use slang, regionalisms, or dialect.

SPEAKING

14. Able to participate effectively in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests and special fields of competence with considerable ease. Can use the language to perform such common professional tasks as answering objections, clarifying points, justifying decisions, responding to challenges, supporting opinion, stating and defending policy. Can demonstrate language competence when conducting meetings, delivering briefings or other extended and elaborate monologues, hypothesising, and dealing with unfamiliar subjects and situations. Can reliably elicit information and informed opinion from native speakers. Can convey abstract concepts in discussions of such topics as economics, culture, science, technology, philosophy as well as his/her professional field. Produces extended discourse and conveys meaning correctly and effectively. Use of structural devices is flexible and elaborate. Speaks readily and in a way that is appropriate to the situation. Without searching for words or phrases, can use the language clearly and relatively naturally to elaborate on concepts freely and make ideas easily understandable to native speakers. May not fully understand some cultural references,

proverbs, and allusions, as well as implications of nuances and idioms, but can easily repair the conversation. Pronunciation may be obviously foreign. Errors may occur in low frequency or highly complex structures characteristic of a formal style of speech. However, occasional errors in pronunciation, grammar, or vocabulary are not serious enough to distort meaning, and rarely disturb the native speaker.

READING COMPREHENSION

15. Able to read with almost complete comprehension a variety of authentic written material on general and professional subjects, including unfamiliar subject matter. Demonstrates the ability to learn through reading. Comprehension is not dependent on subject matter. Contexts include news, informational and editorial items in major periodicals intended for educated native readers, personal and professional correspondence, reports, and material in special fields of competence. Can readily understand such language functions as hypothesising, supporting opinion, argumentation, clarification, and various forms of elaboration. Demonstrates understanding of abstract concepts in texts on complex topics (which may include economics, culture, science, technology), as well as his/her professional field. Almost always able to interpret material correctly, to relate ideas, and to "read between the lines," or understand implicit information. Can generally distinguish between different stylistic levels and often recognises humor, emotional overtones, and subtleties of written language. Misreading is rare. Can get the gist of higher level, sophisticated texts, but may be unable to detect all nuances. Cannot always thoroughly comprehend texts that have an unusually complex structure, low frequency idioms, or a high degree of cultural knowledge embedded in the language. Reading speed may be somewhat slower than that of a native reader.

WRITING

16. Can write effective formal and informal correspondence and documents on practical, social, and professional topics. Can write about special fields of competence with considerable ease. Can use the written language for essay-length argumentation, analysis, hypothesis, and extensive explanation, narration, and description. Can convey abstract concepts when writing about complex topics (which may include economics, culture, science, and technology) as well as his/her professional field. Although techniques used to organise extended texts may seem somewhat foreign to native readers, the correct meaning is conveyed. The relationship and development of ideas are clear, and major points are coherently ordered to fit the purpose of the text. Transitions are usually successful. Control of structure, vocabulary, spelling, and

punctuation is adequate to convey the message accurately. Errors are occasional, do not interfere with comprehension, and rarely disturb the native reader. While writing style may be non-native, it is appropriate for the occasion. When it is necessary for a document to meet full native expectations, some editing will be required.

LEVEL 4 (FULL PROFESSIONAL)

LISTENING COMPREHENSION

17. Understands all forms and styles of speech used for professional purposes, including language used in representation of official policies or points of view, in lectures, and in negotiations. Understands highly sophisticated language including most matters of interest to well-educated native speakers even on unfamiliar general or professional-specialist topics. Understands language specifically tailored for various types of audiences, including that intended for persuasion, representation, and counselling. Can easily adjust to shifts of subject matter and tone. Can readily follow unpredictable turns of thought in both formal and informal speech on any subject matter directed to the general listener. Understands utterances from a wide spectrum of complex language and readily recognises nuances of meaning and stylistic levels as well as irony and humor. Demonstrates understanding of highly abstract concepts in discussions of complex topics (which may include economics, culture, science, and technology) as well as his/her professional field. Readily understands utterances made in the media and in conversations among native speakers both globally and in detail; generally comprehends regionalisms and dialects.

SPEAKING

18. Uses the language with great precision, accuracy, and fluency for all professional purposes including the representation of an official policy or point of view. Can perform highly sophisticated language tasks, involving most matters of interest to well-educated native speakers, even in unfamiliar general or professional-specialist situations. Can readily tailor his/her use of the language to communicate effectively with all types of audiences. Demonstrates the language skills needed to counsel or persuade others. Can set the tone of both professional and non-professional verbal exchanges with a wide variety of native speakers. Can easily shift subject matter and tone and adjust to such shifts initiated by other speakers. Communicates very effectively with native speakers in situations such as conferences, negotiations, lectures, presentations, briefings, and debates on matters of disagreement. Can elaborate on abstract concepts and advocate a position at length in these circumstances. Topics may come from such areas as economics, culture, science, and technology, as well as from his/her professional field. Organises discourse well, conveys meaning effectively, and uses stylistically appropriate discourse features. Can express nuances and make culturally appropriate references. Speaks effortlessly and smoothly, with a firm grasp of various levels of style, but would seldom be perceived

as a native speaker. Nevertheless, any shortcomings, such as non-native pronunciation, do not interfere with intelligibility.

READING

19. Demonstrates strong competence in reading all styles and forms of the written language used for professional purposes, including texts from unfamiliar general and professional-specialist areas. Contexts include newspapers, magazines, and professional literature written for the well-educated reader and may contain topics from such areas as economics, culture, science, and technology, as well as from the reader's own field. Can readily follow unpredictable turns of thought on any subject matter addressed to the general reader. Shows both global and detailed understanding of texts including highly abstract concepts. Can understand almost all cultural references and can relate a specific text to other written materials within the culture. Demonstrates a firm grasp of stylistic nuances, irony, and humor. Reading speed is similar to that of a native reader. Can read reasonably legible handwriting without difficulty.

WRITING

20. Can write the language precisely and accurately for all professional purposes including the representation of an official policy or point of view. Can prepare highly effective written communication in a variety of prose styles, even in unfamiliar general or professional-specialist areas. Demonstrates strong competence in formulating private letters, job-related texts, reports, position papers, and the final draft of a variety of other papers. Shows the ability to use the written language to persuade others and to elaborate on abstract concepts. Topics may come from such areas as economics, culture, science, and technology as well as from the writer's own professional field. Organises extended texts well, conveys meaning effectively, and uses stylistically appropriate prose. Shows a firm grasp of various levels of style and can express nuances and shades of meaning.

LEVEL 5 (NATIVE/BILINGUAL)

LISTENING COMPREHENSION

21. Comprehension equivalent to that of the well-educated native listener. Able to fully understand all forms and styles of speech intelligible to the well-educated native listener, including a number of regional dialects, highly colloquial speech, and language distorted by marked interference from other noise.

SPEAKING

22. Speaking proficiency is functionally equivalent to that of a highly articulate well-educated native speaker and reflects the cultural standards of the country or areas where the language is natively spoken. The speaker uses the language with great flexibility so that all speech, including vocabulary, idioms, colloquialisms, and cultural references, is accepted as native by well-educated native listeners. Pronunciation is consistent with that of well-educated native speakers of a standard dialect.

READING COMPREHENSION

23. Reading proficiency is functionally equivalent to that of the well-educated native reader. Able to fully comprehend all forms and styles of the written language understood by the well-educated native reader. Demonstrates the same facility as the well-educated, non-specialist native when reading general legal documents, technical writing, and literature, including both experimental prose and classical texts. Can read a wide variety of handwritten documents.

WRITING

24. Writing proficiency is functionally equivalent to that of a well-educated native writer. Uses the organisational principles and stylistic devices that reflect the cultural norms of natives when writing formal and informal correspondence, official documents, articles for publication, and material related to a professional specialty. Writing is clear and informative.

Appendix B: Oral Proficiency Interview Information.

The OPI consists of four mandatory phases[42]: *warm-up, level checks, probes, and wind-down.* Each phase must be present in order to obtain a ratable sample. The OPI training manual defines a ratable sample as containing: the four phases, all of the required tasks performed with the necessary accuracy for the relevant level, 2-4 probes, a role-play and a variety of topics. Each phase serves a specific purpose, which needs to be viewed from three different perspectives: the psychological, the linguistic and the evaluative.

Phase 1, the *warm-up,* consists of polite informal conversation at a level that is designed to be very easy for the candidate. There are three purposes for the warm-up: putting the candidate at ease (psychological); reacquainting the candidate with the language and with the tester's pronunciation or way of speaking (linguistic); and giving the tester(s) a preliminary indication of the candidate's level (evaluative). This first hypothesis must be confirmed in the next phase of the interview.

Phase 2, *Level checks*, serves to determine the candidate's highest sustained level of proficiency to allow him/her to demonstrate what s/he can do with the language (psychological). In the level check phase, testers have the candidate perform the

[42] The source of the information pertaining to the OPI is the OPI 2000 Tester Certification Workshop Training Manual (1999).

tasks/functions and content areas associated to a given level (linguistic). Level checks provide a floor for the rating (evaluative).

Phase 3, that of *probes*, aims at establishing the ceiling. The purpose of the *probes* phase is to show the tester(s) whether the candidate has reached his/her highest level of speaking proficiency, and demonstrate what s/he cannot do with the language (psychological). To probe, testers have the candidate attempt to perform tasks one level above the level of level check. Probes serve to determine the functions, content/context areas the examinee cannot handle with the required degree of accuracy, resulting in linguistic breakdown (linguistic). Probing results in either confirming the level established during level checks; establishing a higher floor level if the probes are considered successful; or demonstrating that they examine is at the plus level if the probes are neither completely failed or passed (evaluative). The level check and probes are interwoven, so that the candidate is being alternately challenged and relaxed, not continuously challenged needlessly.

Phase 4, the last phase of the general structure of the interview is called the *wind-down.* The purpose of this phase is to leave the candidate with a feeling of accomplishment (psychological), to give testers a last chance to check any aspect of the candidate's speaking ability that may be incompletely assessed (linguistic) however, it does not add anything new to the speech sample (not evaluative).

This process ideally should take between 10 to 30 minutes, depending on the level of proficiency of the examinee.

There are a variety of question types that constitute OPI elicitation procedures. Each question posed must have a purpose related to one of the four phases, the set of tasks, the three perspectives and the content. Another important factor, which must be kept in mind by the testers, is the focus of the question: does it have a broad or a narrow focus? Is it meant to facilitate a topic or elicit a task? To elicit level-specific tasks and functions, testers use questions and role-play situations as the main elicitation techniques. A variety of question types are recommended for a particular level or levels of speaking proficiency. For level 0+, yes/no and choice questions are required; for levels 1 and 2, information questions; for levels 3, 4, and 5, hypothetical and supported opinion questions. In addition to the types of questions indicated above, the OPI elicitation techniques include role-play situations. At the higher levels (levels 3, 4, and 5), the testers are encouraged to use several role-play situations to elicit the required tasks, such as convincing, advising, persuading, etc. Therefore an OPI for a higher-level candidate may include more role-plays than for one at lower levels.

Appendix C

Questionnaire: General Information about Participants Rater # _____

This information will be useful in understanding the background of the raters in their respective testing programmes and countries. Your name should **NOT** appear in the questionnaire, only your rater number. All findings will be coded in order to preserve your anonymity.

Please answer the following questions.

The number on your envelope says you are rater # _____

1. Age group: *Please circle one:* 19-28, 29-38, 39-48, 49-58, 59+
2. Mother tongue:_____
3. Additional languages in which you are proficient (including English): *Please circle one:* 1, 2, 3, 4+
4. Self-assessed English speaking proficiency based on the STANAG 6001: *Please circle one:* level: 0, 1, 2, 3, 4, 5, don't know.
5. Are you presently an English teacher? *Please circle one* Yes No
6. Have you taught English in the past? *Please circle one* Yes No
7. Level of education: *Please circle highest level achieved* Undergraduate degree, Master's degree, Doctorate.
8. Have you ever taken a university course in language testing? *Please circle one* Yes No
9. Do you conduct Oral Proficiency Interviews (OPI) full-time? *Please circle one* Yes No
10. Do you conduct OPIs part-time? *Please circle one* Yes No
11. How many OPIs would you **estimate** you conduct per year? *Please circle one* 0-20, 21-50, 51-100, 101-300, 300+
12. How long have you been a tester? *Please circle one* 0-1 year, 2-3 years, 4-5 years, 5+ years
13. Is the testing system in your establishment separate from the teaching programme? *Please circle one* Yes No.
14. Does your STANAG testing system presently use 'plus levels'? *Please circle one* Yes No

15. Please provide information on the **tester training** you have received in terms of duration of training, number of practice tests, etc:

_____
_____
_____
_____
_____
_____
_____

16. Please provide information specifically on the training you have received on the **interpretation of the STANAG scale.**

17. Do you find the STANAG easy to use and apply? If not, what aspects of its interpretation/application do you find the most challenging? Please elaborate:

18. Please write any other comment pertaining to the STANAG scale that you may have.

THANK YOU!

Please return this questionnaire in the envelope provided.

Appendix D

Questionnaire on Oral Proficiency Interview (OPI) Sample "A"

INFORMATION:

You will need a rating sheet from your school, department or academy's testing section to rate this OPI. A copy of the NATO STANAG 6001 (Edition 2) speaking is included on the CD (or in your envelope if you are using a cassette), since it is the criteria against which the OPI sample should be compared.

First, a few points about the sample OPI: You will hear a real OPI, which took place in Canada in 2005. It was deemed a 'ratable' sample of speech and the test followed the protocol in place. Since there are no perfect tests, this sample may contain some imperfections either in testing elicitation techniques, in the sound quality, or in the miscommunications, which normally arise in tests. You will not be asked questions regarding the tester's skill and/or the way in which the test was conducted. **Your main task is to listen to the OPI and rate the level of language produced by the examinee using your testing system's protocol,** and to answer the questionnaire as best as you can.

It is very important that you **NOT** consult with colleagues and other testers about this OPI and the rating you will assign. If it is not done individually, **it will invalidate my research.**

INSTRUCTIONS:

I. Listen to the full OPI marked sample "A". You can take notes while you listen, and you may listen to the recording as many times as you wish, just as if you were an official second or third rater.

II. Using the rating protocol which is presently in use in your testing programme, fill out the rating grid or sheets you normally use, and place them in the return envelope. Use the speaking descriptors from the STANAG.

III. Now answer this questionnaire. If you need more space than is provided, please add a sheet and identify clearly the question number.

QUESTIONNAIRE: You are rater #____

1. Please give the rating of the candidate in sample 'A', in terms of STANAG levels.

2. How did you arrive at this rating? Explain the steps you took.

3. At what level would you say the examinee performed the tasks or functions very well? Please provide detailed explanations.

4. Where or during which tasks and functions did you think this examinee did not perform well? Please provide detailed explanations.

5. What did you pay attention to in terms of the examinee's linguistic strengths and weaknesses?

6. Which discourse and delivery features were, in your opinion, most evident in this sample of speech?

7. When comparing the performance from the examinee to the STANAG descriptors, where do you think this examinee fits in, in terms of the overall sociolinguistic aspect?

8. Are there any of the STANAG descriptors which you feel best exemplify this examinee's performance? State them.

9. If your system uses 'plus levels' explain if you think this performance would qualify for a plus rating, or not. Explain. (If you do not use pluses, write *not applicable* in the following space).

10. If you do not use 'plus levels', state whether you think this performance would be rated at a plus level, according to your present understanding of plus levels. Explain your reasons. (If you answered question 9, write *see question 9* in the space below).

11. Please provide any additional information in the space below, which you think will assist me in understanding your rating and the process you followed.

You are rater #____ **Thank you for taking the time to answer these questions.**

Place this questionnaire in the envelope provided.

If you started with sample 'A', please take out OPI sample 'B' and follow the instructions given.

If both samples have been rated, please put all of the contents into the envelope provided and seal the envelope.

Please return it to the person who handed you the envelope.

Appendix E

Protected B (when completed)

MTAP ORAL PROFICIENCY INTERVIEW EVALUATION GRID

Candidate (rank, name, country)			LEVEL:	
Date		Rater(s)	Place	
TASKS/ FUNCTIONS	**LINGUISTIC COMPETENCE Grammar/Voca**	**SOCIOLINGUISTIC/ STRATEGIC COMPETENCE**	**DISCOURSE COMPETENCE (coherence, cohesion)**	**DELIVERY (fluency, pronunciation)**
LEVEL 4 • Tailor language to suit a formal situation • Tailor language to suit an informal situation • Discuss highly abstract concepts at length • Support an opinion on a global or ethical issue • Hypothesize a global or ethical situation	• Uses language with great precision and accuracy	• Can readily tailor language to communicate effectively with all types of audiences • Can express nuances and make culturally appropriate references • Can set and shift the tone of exchanges with a wide variety of NS	• Organizes discourse well, conveys meaning effectively, and uses stylistically appropriate discourse features • Can elaborate on abstract concepts and advocate a position at length • Can easily shift subject matter	• Speaks effortlessly and smoothly with a firm grasp of various levels of style but would seldom be perceived as a NS • Any shortcomings (e.g., non-native pronunciation) do not interfere with intelligibility
LEVEL 3 • Support an opinion on a societal issue • Hypothesize a personal or societal situation • Discuss an abstract topic • Manage an unfamiliar situation	• Flexible and elaborate use of structures • May make errors in low-frequency/highly complex structures in formal speech • Can convey abstract concepts • Occasional errors do not distort meaning and rarely disturb NS	• Appropriate to the situation • Can easily repair the conversation if some cultural reference, proverb, allusion, nuance, or idiom is not fully understood • Makes ideas easily understandable to NS	• Produces extended discourse • Conveys meaning correctly and effectively • Elaborates freely	• Can discuss particular interests with considerable ease • Can use language clearly and relatively naturally • Occasional errors in pronunciation do not distort meaning and rarely disturb NS • Pronunciation may be foreign
LEVEL 2 • State facts • Describe (people, places, things) • Narrate (past, present, future) • Give instructions or directions • Manage a survival situation with a complication	• Appropriate vocabulary for high-frequency topics; unusual/imprecise vocabulary at other times • Typically controls simple structures and basic grammatical relations • Uses complex structures inaccurately or avoids them • Errors may sometimes distort meaning	• Generally appropriate to the situation • Can interact with NS not used to speaking with non-natives, with some adjustment for limitations	• Can speak in complete but simple paragraphs • Can combine and link sentences into paragraph-length discourse • Can often elaborate in common daily communicative situations	• Can confidently handle most normal, casual conversations on concrete topics • Errors in pronunciation may sometimes distort meaning
LEVEL 1 • Participate in short, simple conversations • Ask questions • Manage a basic survival situation	• Erratic word order • Vague time concepts • Often uses only one tense • Tends to avoid certain structures • Frequent errors often distort meaning	• Ineffective reformulations and self-corrections • NS used to speaking with non-natives must often strain, request repetition, and use real-world knowledge	• Can create by combining and recombining familiar, elements • Can speak at the sentence level • May produce strings of 2 or more short, simple sentences joined by common linking words	• Hesitations, frequent pauses • Straining, groping for words (except for routine expressions) • Seldom speaks with natural fluency • Frequent errors in pronunciation often distort meaning

COMMENTS

N.B. This grid has been modified from its original format to fit this page.