The conceptualization and development of a high-stakes video listening test within an AUA framework in a military context

by

Nancy Powers

Department of Integrated Studies in Education
McGill University, Montreal
October 2011

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Arts

© Nancy Powers, 2011
Abstract

The concepts of *justification* and *accountability* are being promoted as a new added value in the field of language testing. Bachman and Palmer’s (2010) Assessment Use Argument (AUA) provides a theoretical framework that can ensure the validity of a test. This study implements an AUA in a high-stakes context to justify the inclusion of videos in a general proficiency listening comprehension test intended for international military personnel studying English in Canada. It follows a three-phase exploratory sequential mixed methods research design. The first phase includes a needs analysis and the development of a prototype. Qualitative data was collected that provided a basis for the next phase. Phase Two includes the development of a computer-delivered video listening test, which follows each stage of test development. In the final phase, Phase 3, the test was trialled on three groups of stakeholders (test developers, teachers and students) and their perceptions of the usefulness of the videos were collected through both qualitative and quantitative methods. The results show that stakeholders perceived the videos as being helpful for comprehension and they appreciated the authenticity of the listening texts. The stakeholders also reported that the videos reduced student anxiety. These data were used as evidence for Claim 1 of the AUA, which states that the use of a test must produce beneficial consequences for the test taker. Though the present study focuses on Claim 1, it does clearly articulate the other three claims of the AUA, which refer to the Decisions, Interpretations, and Assessment Records, as explained by Bachman & Palmer (2010).
Résumé

Les concepts de justification et de fiabilité sont perçus de plus en plus comme une valeur ajoutée dans le domaine de l’évaluation linguistique. Bachman et Palmer (2010) en font la démonstration dans leur cadre théorique en présentant un argument en faveur de l’utilisation de l’évaluation (Assessment Use Argument - AUA) comme moyen de valider un test. La présente étude applique cette notion (AUA) dans un contexte d’enjeux élevés afin de justifier l’inclusion de vidéos lors de l’administration d’un test de compréhension auditive de compétence générale, lequel est destiné à une clientèle militaire internationale apprenant l’anglais au Canada. Cette étude repose sur un modèle de recherche composé de méthodes mixtes s’articulant en trois phases exploratoires séquentielles. La première phase comprend une analyse de besoins et l’élaboration d’un prototype pendant laquelle des données qualitatives ont été recueillies afin de servir à la phase suivante. La deuxième phase porte sur le développement d’un test de compréhension vidéo-auditive généré par un ordinateur. Enfin, la dernière phase présente les résultats obtenus à l’aide de méthodes quantitatives et qualitatives lors de l’essai mené auprès des trois groupes de participants (élaborateurs de tests, enseignants et étudiants), c’est-à-dire leurs perceptions quant à l’utilité des vidéos. Les résultats démontrent que les vidéos ont été perçues par les participants comme étant utiles pour la compréhension et que l’authenticité des tests de compréhension auditive a été appréciée. Les participants ont également affirmé que la présence de vidéos réduisait l’anxiété des étudiants. Ces données positives ont servi à confirmer le Postulat 1 du cadre théorique (AUA) selon lequel l’utilisation d’un test doit produire des conséquences bénéfiques pour le candidat testé. Bien que le Postulat 1 soit le point focal de la présente étude, celle-ci met clairement en lumière les trois autres postulats du cadre théorique AUA de Bachman & Palmer (2010) : les décisions, les interprétations et les rapports d’évaluation.
Acknowledgements

I would like to extend a heartfelt thank you to my supervisor, Dr. Carolyn Turner, who encouraged me from the beginning and never stopped. The idea of this study started many moons ago in her Assessment class, and she has been with me through the many stages of development, always with sound advice and wisdom. I also would like to thank Dr. Mela Sarkar, who agreed to be my second reader. Thank you for your very pertinent and timely comments that helped move my thesis forward.

Next, I would like to thank Mr. Dan Staskevicius, Chief of the Multimedia Production Center at the Canadian Defence Academy (CDA). Without him and his staff, Mr. Dave Hamel, Mr. Jérôme Lebel, and Mr. Valerio Marques de Paula, I would never have been able to film the videos that are integral to this project. I would like to thank Mr. Eric Reneault, Assistant Deputy Chief of Standards, who allowed me the time to create and produce the videos. Next, I would like to thank all those at CDA and CFLS who have supported me throughout the entire process.

A big thank you goes to my “actors”: Ms. Jacqueline Asselin, Ms. Kimberly Batten, Mr. Rod Broeker, Mr. Pierre Coté, Mrs Joanne Cunningham, Ms. Jane Davis, Mr. Andrew Ide, Mr. Sandip Mehta, Ms. Lucie Premont, Dr. Hope Seidman, without whom I would not have been able to make the videos. Thank you, too, to Jocelyne Clermont who helped me with some of the formatting. And to my colleague Camil Villeneuve, who took the time to help with the translation of my abstract. I truly appreciate it.

Last, but certainly not least, I would like to thank my husband, Peet Fortin, whose love and support knows no bounds. With him by my side, everything is possible. Thank you to my children, Sam and Maggie, who have been so patient with Mommy, as she works on her thesis. To all my family – a great big “thank you”! I couldn’t have done it without you.
# Table of Contents

Abstract ............................................................................................................................................... ii
Résumé ................................................................................................................................................ iii
Acknowledgements ............................................................................................................................ iv

**CHAPTER ONE: INTRODUCTION** ................................................................................................. 1
Setting the scene ................................................................................................................................. 1

*The construct of listening comprehension* ...................................................................................... 2
Introduction to the context .................................................................................................................. 2
Motivation for the study ...................................................................................................................... 3
Statement of purpose .......................................................................................................................... 4
Organization of thesis .......................................................................................................................... 5

**CHAPTER TWO: LITERATURE REVIEW** ..................................................................................... 6
Introduction ........................................................................................................................................ 6
The communicative power of visuals and gestures ............................................................................ 6
Videos in testing listening comprehension ....................................................................................... 9
Language anxiety ............................................................................................................................... 16
Computer-based assessments .......................................................................................................... 19
Assessment Use Argument Framework ............................................................................................. 22
Summary ............................................................................................................................................ 27

**CHAPTER THREE: METHODOLOGY: INCLUDING SEQUENTIAL INTER-PHASE RESULTS** ........ 28
Introduction ........................................................................................................................................ 28
Rationale for Study ............................................................................................................................. 28
Military Context .................................................................................................................................. 29

*Language training and testing in NATO* ......................................................................................... 29
*NATO STANAG 6001 Language Proficiency Rating Scale* .............................................................. 30
*The Military Training and Cooperation Program* ......................................................................... 32
Method .................................................................................................................................................. 34
Research Question and Objectives ........................................................................................................ 34
Phase One
(a): Needs Analysis ................................................................................................................................. 37
  Purpose ................................................................................................................................................. 37
  Context .................................................................................................................................................. 37
  Participants ............................................................................................................................................ 37
  Instrument ............................................................................................................................................ 38
  Procedure ............................................................................................................................................. 38
  Data Analysis ....................................................................................................................................... 38
(b): Development of a Prototype Video Listening Test ............................................................................. 39
  Purpose ................................................................................................................................................. 39
  Context .................................................................................................................................................. 39
  Participants ............................................................................................................................................ 39
    Group 1 .............................................................................................................................................. 40
    Group 2 .............................................................................................................................................. 40
  Instrument ............................................................................................................................................ 40
  Procedure ............................................................................................................................................. 42
  Data Analysis ....................................................................................................................................... 43
Phase Two: Development of a multi-level Video Listening Test within an AUA framework ................. 43
  Purpose ................................................................................................................................................. 43
  Context .................................................................................................................................................. 44
  Participants ............................................................................................................................................ 44
  Instrument ............................................................................................................................................ 44
  Procedure ............................................................................................................................................. 44
    Procedure: Stage One: Initial Planning ............................................................................................... 48
    Procedure Stage Two: Assessment Design ......................................................................................... 53
    Procedure Stage Three: Operationalization ..................................................................................... 61
  Data Analysis ....................................................................................................................................... 62
Phase Three: Trial of Video Listening Test ............................................................................................... 62
Purpose ........................................................................................................ 63
Context ......................................................................................................... 63
Participants .................................................................................................. 63
  Group 1: Test developers ........................................................................ 63
  Group 2: MTCP teachers .......................................................................... 64
  Group 3: MTCP students ......................................................................... 64
Instruments .................................................................................................. 64
Procedure .................................................................................................... 65
  Group 1 ..................................................................................................... 65
  Groups 2 & 3 ............................................................................................ 66
Data Analysis ............................................................................................... 66
Summary ...................................................................................................... 67

CHAPTER FOUR: PRESENTATION OF RESULTS: INCLUDING AN AUA
EXPLANATION AND DISCUSSION .............................................................. 68
Introduction .................................................................................................. 68
Phase One (a) Results .................................................................................. 68
  Results of needs analysis ......................................................................... 68
  Summary of needs analysis ..................................................................... 69
Phase One (b) Results .................................................................................. 70
  Results from prototype trial ..................................................................... 70
  Summary of prototype trial ..................................................................... 71
Phase Two Results ........................................................................................ 71
Phase Three Results ..................................................................................... 72
  Quantitative .............................................................................................. 72
    Summary of quantitative results ............................................................. 74
  Qualitative ................................................................................................. 75
    The Assessment Use Argument ............................................................. 76
Summary ...................................................................................................... 99
CHAPTER FIVE: FINAL DISCUSSION: THE RESEARCH QUESTION AND OBJECTIVES

Introduction ......................................................................................................................... 100
Research Objectives revisited ........................................................................................... 100
Research Question revisited .............................................................................................. 104

CHAPTER SIX: CONCLUSION ............................................................................................ 106
Introduction ......................................................................................................................... 106
Summary of findings ........................................................................................................... 106
Implications ......................................................................................................................... 107
Limitations .......................................................................................................................... 108
Future research ................................................................................................................... 108
Contribution ......................................................................................................................... 109

REFERENCES ...................................................................................................................... 110

APPENDICES
Appendix A: STANAG 6001 Level Descriptions for Listening Comprehension ........ 122
Appendix B: Focus group meeting questions ................................................................. 124
Appendix C: Focus group meeting consent form ............................................................ 125
Appendix D: Trisection: Listening comprehension .......................................................... 127
Appendix E: Video Listening Test Blueprint ................................................................. 129
Appendix F: VLT Test Item Specifications ........................................................................ 132
Appendix G: Questionnaire for students .......................................................................... 133
Appendix H: Test developers’ consent forms ................................................................. 141
Appendix I: MTCP teachers’ consent forms ................................................................. 143
Appendix J: MTCP students’ consent forms .................................................................... 145
List of Tables

Table 1: Walma van der Molen’s (2001) coding system used by Cross (2010)
Table 2: Summary of Cross’ findings
Table 3: Summary of Wagner’s research
Table 4: Attributes of Stakeholders
Table 5: Describing the intended beneficial consequences
Table 6: The decisions, stakeholders affected by decisions, and individuals responsible for making the decisions
Table 7: TLU task characteristics
Table 8: Summary of the focus group meeting
Table 9: Stakeholders’ responses: Frequency counts in percentages
Table 10: Individual group responses: Frequency counts in percentages
Table 11: Example of the structure of an AUA
Table 12: The decisions, stakeholders affected by decisions, and individuals responsible for making the decisions
List of Figures

Figure 1: Links from test taker’s performance to intended uses (decisions, consequences)
Figure 2: Framework of AUA
Figure 3: Three phases of the research design
Figure 4: Exploratory sequential design
Figure 5: Phase One: Needs Analysis and the Development of a Prototype
Figure 6: The interface of the prototype Video Listening Test
Figure 7: Phase Two: Development of a multi-level video listening test within an AUA framework
Figure 8: Interface of the Video Listening Test
Figure 9: Controls for the video
Figure 10: Phase Three: Trial of the Video Listening Test
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUA</td>
<td>Assessment Use Argument</td>
</tr>
<tr>
<td>ALTS</td>
<td>Advanced Language Testing Seminar</td>
</tr>
<tr>
<td>BILC</td>
<td>Bureau for International Language Coordination</td>
</tr>
<tr>
<td>CAT</td>
<td>Computer-adaptive tests</td>
</tr>
<tr>
<td>CBA</td>
<td>Computer-based assessments</td>
</tr>
<tr>
<td>CBT</td>
<td>Computer-based tests</td>
</tr>
<tr>
<td>CDA</td>
<td>Canadian Defence Academy</td>
</tr>
<tr>
<td>CFLS</td>
<td>Canadian Forces Language School</td>
</tr>
<tr>
<td>ESL/EFL</td>
<td>English as a Second/Foreign Language</td>
</tr>
<tr>
<td>FNTP</td>
<td>Foreign National Training Plan</td>
</tr>
<tr>
<td>ILR</td>
<td>Interagency Language Roundtable</td>
</tr>
<tr>
<td>ILTA</td>
<td>International Language Testing Association</td>
</tr>
<tr>
<td>IRB</td>
<td>Item Review Board</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>LTS</td>
<td>Language Testing Seminar</td>
</tr>
<tr>
<td>MPC</td>
<td>Multimedia Production Center</td>
</tr>
<tr>
<td>MSLTP</td>
<td>Military Second Language Training Plan</td>
</tr>
<tr>
<td>MTCP</td>
<td>Military Training and Cooperation Program</td>
</tr>
<tr>
<td>NATO</td>
<td>North Atlantic Treaty Organization</td>
</tr>
<tr>
<td>PfP</td>
<td>Partnership for Peace</td>
</tr>
<tr>
<td>PSC</td>
<td>Public Service Commission (of Canada)</td>
</tr>
<tr>
<td>QS</td>
<td>Qualification Standard</td>
</tr>
<tr>
<td>SLP</td>
<td>Standard Language Profile</td>
</tr>
<tr>
<td>STANAG 6001</td>
<td>Standardization Agreement 6001</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>---------------------</td>
</tr>
<tr>
<td>TLU</td>
<td>Target Language Use</td>
</tr>
<tr>
<td>VLT</td>
<td>Video Listening Test</td>
</tr>
</tbody>
</table>
CHAPTER ONE
INTRODUCTION

Setting the scene

Anyone who has ever studied a second or foreign language has a funny or embarrassing story to tell because they did not understand what had been said to them in the target language. These learners may have given inappropriate responses, they may have done something they should not have done, or vice versa, which all make for some humorous stories after the fact. Yet all these stories and situations involve one thing – communication breakdown caused by not understanding what the speaker was saying.

Competence in listening comprehension is among the most important factors that come into play when determining a language learner’s success or failure in learning a second or foreign language. Guo and Wills (2005) stated that, “language learning depends on listening, since it [listening] provides the aural input that serves as the basis for language acquisition and enables learners to interact in spoken communication.” In other words, listening is the skill that is the basis for the development of all other skills. Yet, history has shown us that teaching listening comprehension has been either neglected or poorly taught due to the belief that it was a “passive” skill. In the past, educators believed that merely exposing students to the spoken language provided adequate instruction in listening (Call, 1985; Canale & Swain; 1980), or more specifically, comprehensible input (Krashen, 1985; Long 1996). Baell et al. (2008) reviewed the recent state of the skill of listening in education, and discovered that not much has changed – it is still not being taught to any great extent in the classroom and students are still having difficulty with it.

What makes listening comprehension so difficult? Perhaps it is the heavy processing load that contributes to L2 language learners losing their concentration quickly (Ma, 2005). Perhaps it is the lack of control the language learners feel when having to listen to the target language. According to Rubin (2008), the listener has almost no control over what is going to be said, how it is going to be said, or how quickly it is going to be said and this lack of control is a key reason why learners of a second or foreign language have such difficulty. Another reason that makes L2 listening comprehension such a challenge
may be in the differences between an aural text and a written text, which differ in three main ways: speech is (a) encoded in the form of sound; (b) linear and takes place in real time, with no chance of review; and (c) linguistically different from written language (Buck, 2001).

The construct of listening comprehension

In recognizing the complexity of the processes involved in listening comprehension, it is understandable that assessing someone’s listening comprehension has its challenges. For one thing, a single acceptable definition of the construct of listening does not exist. Several researchers (Buck, 2001; Wagner, 2002) have suggested that the definition of the construct should be dependent on the Target Language Use situations (TLU) and on the purpose for listening. Further, Bachman and Palmer (1996) argue that test tasks should resemble the “real life” tasks that are found in the TLU.

What has been emerging in the listening assessment literature is the growing acceptance of the importance of the visual and non-verbal aspects that are inherent in most listening situations, and the roles they play in comprehension (Beattie & Shovelton, 1999a; Kellerman, 1990, 1992; Kelly et al; 1999; Kendon, 2004; Tyler and Warren, 1987; von Raffler-Engel, 1980; Wagner 2002, 2007, 2008, 2010). Wagner (2007) has even stated that the definition of listening should be expanded to include visuals since most of TLU situations involve face-to-face interactions, except, of course, situations that involve the use of the telephone or listening to the radio. He argues that to preclude non-verbal information on listening tests could be seen as a threat to the validity of the inferences made about a person’s L2 listening ability based on those tests. Despite this acceptance, there have been few studies that have looked at the role of visuals in assessing listening.

Introduction to the context

Developing the skill of listening is difficult for many L2 learners. It is no different for the military, whose very lives depend on communication with one another. In today’s global world, militaries work together in different missions, on military exercises, or at the NATO Headquarters, and English has become the common language of communication. The situations that these people find themselves in can be very
dangerous and they must rely on communicating with each other to survive. If they fail to understand what is being said to them, the consequences may be dire. Proficiency in the English language has become an extremely important feature for the world’s militaries.

The mission of the Canadian Defence Academy (CDA) is “to lead Canadian Forces professional development, uphold the profession of arms and champion lifelong learning to enable operational success” (http://www.cda.forces.gc.ca/index-eng.asp). With support from CDA, the Canadian Forces Language School (CFLS) offers two programs of language training: the Military Second Language Training Program (MSLTP) and the Military Training and Cooperation Program (MTCP). The MSLTP program is offered to members of our Canadian military who are in need of improving their second official language (English or French). The MTCP program is an intensive language immersion program (English and French) that is offered to members of foreign military who come to Canada from NATO and Partnership for Peace (PfP)\(^1\) countries in order to improve their language skills. CDA is responsible for providing the curriculum and the testing for these different programs. It is the responsibility of the Testing department to develop, administer and maintain the language tests that are associated with both the MSLTP and the MTCP.

**Motivation for the study**

My motivation for this study stems from the years I have spent working as a language tester at CDA and observing how the students who study English under the MTCP program have great difficulty with the listening test that is administered. As Chief of the English Testing department, I have been concerned about this and the resulting low listening scores of our foreign students for quite some time. It is difficult to understand how so many of our students score poorly on the listening comprehension test after spending 19 weeks in an English immersion setting when they clearly show signs of comprehension during the speaking test. One possible explanation of this situation may be related to the test method. The students have mentioned on numerous occasions that

---

\(^1\) According to MilitaryDictionary.com, the definition of Partnership for Peace (PfP) is “an agreement between NATO and various non-NATO countries to cooperate in the interests of peace and security, especially in Europe.”
listening to a disembodied voice over a loud speaker is challenging – it affects their concentration, their focus and their comprehension of the context.

I had been thinking of developing a different method of testing listening that would enable us to more accurately measure the students’ listening abilities; one that would allow the students to focus on something concrete and meaningful, would allow them to concentrate on what is being said, and would reduce their feelings of anxiety to enable them to better understand the speakers. I believe that the addition of visuals to a listening test would meet these criteria listed above and would have beneficial consequences for test takers in that visuals will help them better comprehend the speakers. Consequently, I developed a prototype of a computer-delivered, multi-level, general proficiency English listening test that uses video to deliver the listening passages. The extremely positive response I received from colleagues encouraged me to continue exploring this method of testing listening comprehension. I decided to develop a test that incorporated many of the suggestions that were made to improve the prototype test, but was also grounded in theory. I decided to develop it within Bachman and Palmer’s (2010) Assessment Use Argument theoretical framework.

**Statement of Purpose**

The purpose of this exploratory study is to investigate the perceptions and performance of three groups of stakeholders with regard to using videos in a computer-delivered English general proficiency listening test. This study is grounded in the theoretical framework of an Assessment Use Argument, and the results will provide backing to support the claim that using videos will have beneficial consequences for the stakeholders. This study will enrich the language testing community in that this perspective has not been taken with respect to video listening tests. The implications of this study range beyond the context of CDA and CFLS, and may extend to other NATO nations and PfP countries who are working to ensure that their listening tests yield valid results, as well as educators who are interested in testing listening comprehension in general.
Organization of thesis

This chapter has provided an introduction to the context in which the issue under investigation is situated. It has explained my motivation for doing this research as well as stating its purpose. In the next chapter (Chapter Two), I review the literature that provides a rationale for my study and give an explanation of the Assessment Use Argument (AUA). In Chapter Three, I explain the specific context in which my study is situated, the exploratory, sequential research design and the three phases of the study. I also state the research question and the research objectives. As well in Chapter Three, I present the results from Phases 1 and 2, given that each phase builds upon the previous one. In Chapter Four, I report the final results from Phase 3 in relation to the AUA framework. Support is given that elaborate the claims stated in the AUA, which culminates in a fully articulated AUA. All results are then discussed with respect to the research question and objectives in Chapter Five. I conclude the thesis in Chapter Six with a summary of the findings, draw some conclusions, and discuss implications and limitations of the study. The thesis ends with some recommendations for further research and the contribution this study will make to the language testing field.
CHAPTER TWO
LITERATURE REVIEW

Introduction

In this chapter, I review the literature on the communicative power of visuals and gestures, which is essential in order to justify the premise of including videos in a listening comprehension test. I then review studies that have been conducted on assessing listening with the use of videos. This review of the literature situates this study and helps give the study purpose. I then discuss the concept of language learner anxiety, with the purpose of demonstrating that it can be a debilitating factor to one’s performance on a test. I then discuss how computer-delivered tests, in general, may reduce test anxiety, and some important considerations that must be kept in mind when developing a computer-based test. Finally, I explain the theoretical framework of an Assessment Use Argument which serves as the foundation for this study.

The communicative power of visuals and gestures

With advances in technology, such as iPods, iPads, SKYPE with video and videoconferencing, it is obvious that the world’s communication system is becoming much more visually enhanced. Computers are now commonplace and the young generation of today have even been dubbed “Digital Natives,” because their “native language” is that of computers and technology (Prensky, 2001). Many areas of education are making use of these new technologies and language classrooms are no exception. Language assessors are slowly starting to incorporate these new technologies into their tests in all skills. As Okey (2007) noted, “computer-based listening tests will no doubt continue to replace audio-based listening tests, and some type of visual stimulus will almost certainly be included in these tests.” Test developers need to decide how visuals should be incorporated and what benefits they will hold for the students.

Hostetter (2011) recently conducted a meta-analysis of studies that have looked at whether or not the use of gestures is beneficial to comprehending the speaker’s message. She identified four different categories of gestures: representational, deictic, iconic and

---

2 Also referred to as the Net Generation
Representational gestures are “movements that depict a spatial or motor referent by pantomiming a particular action, but demonstrating a spatial property or by creating such a referent for an abstract idea” (p. 298). Deictic gestures refer to when a speaker points to an object or location in the environment that is relevant to what is being said. Iconic gestures refer to when the speaker makes a gesture that repeats what is being said; for example, if the speaker is talking about scales, and s/he makes the gesture of a scale. Finally, beat gestures are “small, rhythmic movements that do not convey any obvious semantic content”. Hostetter (2011) concluded that gestures do indeed communicate and that listeners have better comprehension of speech when that speech is accompanied by visible gestures than when it is not.

In addition to conducting a synthesis of the studies, Hostetter (2011) investigated the question of when gestures communicate, given that Goldin-Meadow (2003) concluded that the effectiveness of gestures depended entirely on the type of speaking situation. Hostetter (2011) found that “gestures may glean their communicative power in a number of non-mutually exclusive ways”: they (1) convey information about spatial ideas, spatial relations, and motor events; (2) convey information that is not present in the accompanying verbal description; (3) provide additional cues when speech comprehension is difficult (especially for the L2 learner) in that gestures may be more helpful to listeners with weak verbal skills than to listeners with strong verbal skills because the gestures can provide a nonverbal means of acquiring the same information (p. 299). These three ways exemplify direct influences of gestures on communication. Hostetter (2011) suggests that gestures may also have some indirect influence: gestures (4) help the speaker provide more fluent and rich descriptions; (5) capture and maintain the attention of the listener and build rapport between the speaker and the listener; and finally (6) provide cues that may actually promote learning (p. 298-299) Broaders & Goldin-Meadow’s (2010) findings support the second way of the communicative power of gestures. They suggest that “listeners are quite good at noticing information that is conveyed through non-redundant gestures and using it to inform their knowledge of the speaker’s meaning”.

Marricchiolo et al. (2009) suggested that L1 listeners rate speakers who gesture as more competent and composed than speakers who do not gesture. Moreover, Kelly &
Goldsmith (2004) concluded that listeners also report liking speakers who gesture more than speakers who do not gesture. They found that listeners seem to pay attention more to speakers who gesture, and they concluded that the role of the gestures may be just that – to gain the attention of the listener, which can then help ensure comprehension. Hostetter (2011) concluded that “one implication of this view is that gestures should have their attention-getting, and thereby communicative, power regardless of the topic being discussed”. She concluded her meta-analysis by stating that the question “of whether gestures communicate cannot be addressed without also considering aspects of the particular speaking situation in which they occur” (p. 312). In fact, the use of gestures changes according to different situations. For example, a speaker will gesture more when they know the listener can see them (Alibali, Heath, & Myers, 2001), or if the listener is naïve about the topic of discussion (Jacobs & Garnham, 2007). There are also studies that have concluded that the gestures used by speakers did not provide any help at all for comprehension (Kelly & Goldsmith, 2004; Krauss, Dushay, Chen, & Rauscher, 1995). This variation across contexts is something to consider in teaching and in testing situations.

It is interesting that even in different disciplines, researchers have studied how gestures and speech affect the brain. Hubbard et al (2009) conducted a study that investigated the effect on the brain when an L1 speaker was presented with beat gestures and speech simultaneously. The auditory cortex saw increased activity when the two were presented together when compared to speech alone. “These findings suggest a common neural substrate for processing speech and gesture, likely reflecting their joint communicative role in social interactions” (Hubbard et al, 2009). Hubbard et al (2009) concluded that the brain reacts when speech is accompanied with gestures, which suggest that using gestures when speaking is a natural phenomenon that we may not consciously decide to do. A question to be asked is whether these findings could be generalized to contexts that suggest that gestures play a similar role in L2 listening comprehension. This question is outside the scope of this paper, but could be addressed in future research.

In this section, I have discussed the importance of visuals and gestures on comprehension and the communicative power gestures seem to hold. Depending on the context, the significance of the gestures may be different, but they do appear to help
communicate. In the next section, I review the literature on using videos to test listening comprehension. There is an ever growing literature on cross-cultural pragmatics and gestures (Le Guen, 2011; Li, Abarbanell, Gleitman & Papafragou, 2011; Sotaro, 2009). However, this area of study is outside the scope of this thesis.

**Videos in testing listening comprehension**

In this computer and technological age, it is not surprising that there is an abundance of computer-based/generated educational activities that are used regularly in English as a Second Language/English as a Foreign Language (ESL/EFL) classrooms. Using videos in ESL/EFL is commonplace and several researchers have found favourable results when video is used to teach listening comprehension (Canning-Wilson, 2000; Secules et al., 1992); the previous section discussed the communicative power of gestures, which can easily been seen through videos. Okey (2007) looked at how either moving images (video) or still pictures affected six ESL test takers’ comprehension of two lectureettes. He found that the test takers barely attended to the still pictures and that these had provided little help with comprehension. With regard to the moving images, he found that half his test takers found them helpful and half found them distracting, which seems to suggest that there is a lot of individual variation in the apparent usefulness of the videos.

Wagner (2008) found that most test takers reported that hand gestures helped the most in comprehending the videotext. Further, Cross (2010) reported that in at least one of the BBC news videotexts he had shown to test takers, they mentioned that the hand gestures used by the reporter when comparing and contrasting plastic bags versus biodegradable bags helped to orient them to the aural content, which then helped them with the comprehension of that content. As Cross (2011) reports, “this is in line with Wagner’s (2008) findings that hand gestures can help learners to interpret information in videotexts, and supports the perceptions of the learners in Coniam’s (2001), Okey’s (2007) and Sueyoshi and Hardison’s (2005) studies regarding the usefulness of a speaker’s gestures in aiding listening comprehension.” These findings lend further support to claims that were made as far back as 1990, which stated that L2 listeners are able to more easily construct the meaning of a spoken text that includes non-verbal input than a spoken text
that does not include non-verbal input (Gruba, 1997; Kellerman, 1990, 1992; Progosh, 1996). (Non-verbal input refers to facial movements, body movements and gestures.)

Berk (2009) looked to other disciplines to underline the benefit of using videos as teaching materials. He reviewed studies on how videos may affect the brain and the three core intelligences, which are defined as: verbal/linguistic (learn by reading, writing, speaking, listening, debating, discussing and playing word games); visual/spatial (learn by seeing, imagining, drawing, sculpting, painting, decorating, designing graphics and architecture, coordinating colour and creating mental pictures); and musical/rhythmic (learn by singing, humming, listening to music, composing, keeping time, performing, and recognizing rhythm) (p.3). Gardner (2000) stated that videos can tap into all three core intelligences. Berk (2009) also points out how videos can engage both sides of the brain, which can allow for greater learning potential in students, if the learning activities following the videos attend to these different parts of the brain (for a fuller account of the effect of videos on the brain, see Berk, 2009).

Yet, despite the evidence of the benefits of using videos on learning in general, and in ESL/EFL in particular, there have been relatively few studies that have looked at using videos in assessing listening comprehension; of those studies, the results reported are not conclusive. Some researchers concluded that videos were simply distractions for the students (Brett, 1997; Gruba, 1993), and others have stated that perhaps we would be testing something other than listening comprehension if videos were used (Buck, 2001; Rost 2002). Yet, there are researchers who have concluded that videos do/can help students understand what is being said and their inclusion on tests can lead to increased performance (Baltova, 1994; Shin, 1998; Sueyoshi & Hardison, 2005; Wagner, 2010b). These conflicting results suggest that more research is needed in this area.

Over the past decade, Wagner (2002, 2007, 2008, 2010a, 2010b) has conducted a series of studies on how videos in listening tests may affect the L2 test taker. In 2002, he “explored the listening process when the aural input is delivered through the use of videos” (p. 1). He hypothesized a model of listening as being a two-stage model of bottom-up and top-down processing, which was supported by several researchers (Buck, 2001, Brindley, 1998). His results, however, did not support this construct definition, and instead supported a two-factor model of listening as the ability to comprehend explicit
and implicit information. He concluded that there should be more research on this two-factor model in an attempt to validate it. As Wagner (2010) pointed out, this model of listening is analogous to Buck’s (2001) default listening construct, which is defined as the ability “to understand the linguistic information that is unequivocally included in the text” (p.114) (ability to listen for explicitly stated information) and the ability to make whatever inferences are unambiguously implicated by the content of the passage’ (p.114) (the ability to listen for implicitly stated information). He also acknowledged that the chosen item format may have had an effect that was inherent in the assessment instrument. Wagner stated that “by their very nature, limited-production items may be more suitable for testing a listener’s ability to comprehend inferential information, while multiple-choice items may be better suited to assess a listener’s ability to comprehend explicitly stated information” (p. 26). Because there is no agreed upon definition of listening comprehension, Wagner (2007) suggested that the construct definition of listening may depend more on the purpose of the listening situation as opposed to a global one.

To recap the above discussion, researchers have not been able to agree on a universal construct definition of listening comprehension, which may be due, in part, to the complexities of the processes involved in listening comprehension. Perhaps Wagner’s (2002) suggestion that the construct of listening should depend on the Target Language Use situations is a more useful way of looking at the skill, and will allow for more valid interpretations.

There are several researchers who do not share the view that videos used in assessment tools are potentially beneficial. Bejar et al (2000) in their working paper devoted to creating a listening framework for the TOEFL 2000 test (Test of English as a Foreign Language) stated “there is no doubt that video offers the potential for enhanced face validity and authenticity, although there is a lot of concern about its potential for distraction.” In 2001, Coniam conducted a study in which he compared the use of audio and video as an assessment instrument in the certification of English language teachers and found that the majority of teachers did not feel the video helped in their comprehension. Actually, some of the participants commented that the audio format may actually be less distracting for test takers if the video consists essentially of “talking
heads”. Wagner (2007) conducted a study to see just how test takers viewed the video when it is part of an assessment tool; whether the presence of videos really was a distraction for ESL test takers or not. He discovered that they watched the video 69% of the time, and that this behaviour did not change over the course of the test. He concluded that the test takers did not find the video distracting. This supports Ginther’s (2002) statement that the presence of visuals results in facilitation of performance, and therefore is not a distraction, when the visuals bear information that complements the audio portion of the stimulus. In other words, if the video is interesting and relevant to the audio track, then it should not be a distraction.

Cross (2010) studied how L2 learners use visual content to help understand news videotexts. He examined five BBC news videotexts and categorized them using Walma van der Molen’s (2001) coding system. The categories were: (1) Direct, (2) Indirect, (3) Divergent, and (4) Talking Heads. Audio and visual content that had the same meaning were placed in the Direct category. Indirect meant that the audio and visual content were only partly related. Divergent referred to meanings that were not related at all and Talking Heads referred to when only the head was seen on screen, and no relation of any kind could be detected between the audio and visual content (Cross, 2010). See Table 1 for a summary of the four categories used by Cross (2010).

<table>
<thead>
<tr>
<th>Walma van der Molen’s (2001) coding system used by Cross (2010)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Direct</strong></td>
</tr>
<tr>
<td>Same propositional meaning between audio and visual content</td>
</tr>
<tr>
<td><strong>Divergent</strong></td>
</tr>
<tr>
<td>No correspondence in meaning between audio and visual content</td>
</tr>
</tbody>
</table>

He found that independent/irrespective of the degree of relatedness between the visual and audio content, the presence of the visual created an added strain on the learner’s limited cognitive resources. He suggested that perhaps teaching decoding or awareness strategies may help, since many of the learners did not “recognize congruence and discrepancies
between the aural and visual elements as they strove for understanding” (Cross, 2010). He did not distinguish students with differing levels of proficiency, yet he suggested that learners’ attention be drawn to a speaker’s lip movements and facial expressions, as some researchers have found that attending to these features aids in comprehension (Ockey, 2007; Sueyoshi & Hardison, 2005). Regardless of these findings, Cross (2010) did conclude that both the Direct and Indirect categories had positive influences on comprehension, whereas the divergent and Talking Head categories did not (Cross, 2010). Table 2 summarizes his findings.

<table>
<thead>
<tr>
<th></th>
<th>Direct</th>
<th>Indirect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Facilitative of comprehension</td>
<td>Influenced comprehension positively</td>
</tr>
<tr>
<td>Divergent</td>
<td>Problematic with facilitating comprehension</td>
<td>Talking Head</td>
</tr>
<tr>
<td></td>
<td>Little facilitative ability</td>
<td></td>
</tr>
</tbody>
</table>

Buck (2001) suggested that test developers should focus on testing language ability “rather than the ability to understand subtle visual information” (p.172). In addition, Buck stated that because research has suggested that people “differ quite considerably in their ability to utilize visual information” (p. 172), it is better to emphasize comprehension of the aural rather than the visual. In 2008, Wagner conducted a study that focused on the effect of nonverbal information on individuals. His findings supported Buck (2001) in that he found that, through participants’ verbal reports, there was a great deal of variation in how the test takers attended to and utilized the nonverbal information that was inherent in the videos. However, contrary to what Buck stated, Wagner (2008) concluded that because people vary in their ability to utilize the nonverbal information in “real life”, then it “can be seen as construct relevant variance if this ability is included in the construct definition of L2 listening ability”. In this study, Wagner (2008) also discussed the idea that videos can provide extensive contextual information that allows for test takers to interpret information such as the speaker’s age, status, use of
sarcasm, irony and humour. This fits well with Pupura’s (2004) definition of pragmatic knowledge and can be considered construct-relevant variance on tests of L2 listening ability.

In 2010, Wagner conducted two studies: (a) the first “explored test takers’ behaviour and attitudes toward the use of video texts” on an L2 listening test and (b) the second studied the effect of videos on test takers’ performance. In 2010(a), the videos were projected on a screen at the front of a classroom and there were three video cameras that were set up to record the test takers’ behaviours. Wagner acknowledges that the conditions were not ideal in that the overhead lights needed to remain on for the video camera, but the lighting then adversely affected the quality of the viewing screen for the videos. The results showed that the test takers viewed the videos less than half the time, which is a much lower rate than in his previous study (2007), where test takers watched the videos 69% of the time. However, despite the variation in individual viewing behaviours, test takers reported positive attitudes towards using videos in listening tests, which supports the findings from previous research (Dunkel, 1991; Parry & Meredith, 1984; Progosh, 1996). Interestingly, Wagner also found a weak negative correlation between viewing rate and performance on the test, meaning that those who watched the videos more often scored worse on the test than those who did not view the video.

Wagner (2010b) then looked more closely at what effect using video texts had on test performance. He divided students, who represented numerous and diverse cultural backgrounds, into two groups: audio-only and video. Within six weeks of administering a pre-test, the two groups were given the post-test. The videos were played on a video monitor at the front of the class for the video group, where all students could see it. The same was done for the audio-only group, except that a paper was put over the screen, which prevented the students from seeing the videos. Wagner found that students in the video group performed 6.5% better than those in the audio-only, and this difference was statistically significant. These findings support other studies which have shown an increase in test taker performance when videos are used (Baltova, 1994; Shin, 1998; Sueyoshi & Hardison, 2005). Gruba (2006) found that the learners in his study thought that the videos helped reduce their anxiety and improve their motivation.
Wagner contends that there is strong theoretical justification for using videos in listening tests. He argues that whether or not videos should be included in a test of listening comprehension should really be dictated by the target language use situation (TLU) and the purpose of the test. According to Bachman and Palmer (1996), test task characteristics should resemble the tasks of the TLU as closely as possible, if we want to make inferences about the student’s ability to perform in the “real world”. It follows, then, that if a test is to measure one’s ability to understand spoken language over the phone, or through a radio device, then having videos in the test would not be appropriate. However, the purpose of most tests of listening comprehension is to test a much larger domain, and the majority of situations in which students find themselves where they need to use their second/foreign language include seeing the person speaking – therefore non-verbal cues are present (Wagner, 2007). Thus, if we want to include the ability to understand the visuals that are inherent in listening situations, then we must expand our definition of listening comprehension (Wagner 2010b). Wagner supports Messick’s (1989, 1996) argument that:

threats to construct validity include not only construct irrelevant variance, but also construct underrepresentation. If the purpose of the L2 listening test is to assess listener’s ability in a TLU domain that includes the non-verbal components of spoken language, to exclude them in a listening test task might threaten the validity of the inferences made from the results of that test because of the underrepresentativeness of the task. (p. 509)

For a summary of Wagner’s research, see Table 3 below.

<table>
<thead>
<tr>
<th>Year</th>
<th>Research Topic</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>Explored the listening process when the aural input is delivered through the use of videos (Video Listening Test)</td>
<td>Results did not support the model of top-down/bottom-up processing; rather, data suggested a two-factor model of listening as the ability to comprehend explicit and implicit information</td>
</tr>
</tbody>
</table>
In this section, I have reviewed the literature of previous research that has been conducted using videos in listening tests. Although not all researchers believe that the effect of videos will be beneficial to students, Wagner has provided solid arguments for their inclusion in the construct definition of listening, as well as in tests of L2 listening comprehension. In the next section, I will discuss anxiety felt by students when taking language tests.

**Language Anxiety**

At the Canadian Forces Language School (CLFS), the MTCP students become extremely anxious when faced with listening tasks in the classroom or on a test. One
contributing factor to their feelings of anxiety in the classroom is the fact that these students are mature, experienced people with complex ideas and thoughts, yet they cannot express them due to their linguistic deficiencies. Learning a language relies on communicating ideas, and, therefore, necessarily involves risk taking and exposing one’s weaknesses to others. Depending on how these weaknesses are dealt with in the classroom, and depending on personal traits, a learner’s feelings of anxiety may be shaped. In a learning context, past emotional reactions to classroom experiences will affect future performance. Goleman (1995) reports that "anxiety undermines the intellect" (p. 83); and this can be explained neurobiologically because anxiety "can create neural static, sabotaging the ability of the prefrontal lobe to maintain working memory" (p. 27). A precise definition of Foreign Language (FL) anxiety was offered by Horwitz, Horwitz, and Cope in 1986, and it is still relevant today:

a distinct complex of self-perceptions, beliefs, feelings, and behaviours related to classroom language learning arising from the uniqueness of the language learning process. It may arise from self-doubt, frustration, and perceived (or fear of) failure. When anxiety is associated with learning a FL, it can manifest itself in altered performance, lower test scores, and final grades (p. 128).

Several researchers have studied anxiety and its relationship to listening comprehension (Bacon, 1989; Gardner, Lalonde, Moorcroft, & Evers, 1987; Lund, 1991). The consensus is that anxiety impedes listening comprehension, along with all the other skills. Research on listening anxiety has found an association with language competence (Chen & Chang, 2004; Elkhafaifi, 2005a; Liu, 2006; Mills, Pajares, & Herron, 2006; Vogely, 1995). In general, language competence is undeniably an essential factor that affects anxiety. Therefore it follows that if learning how to listen is not focussed on in the classroom, then one can argue that the student’s competence in listening is lower than his/her competence in other skills, which can then add to a student’s feelings of anxiety when it comes to listening.

Hasan (2000) asked EFL learners at Damascus University in Syria about their perceptions of what their problems in listening comprehension were. Fifty-four percent of listeners reported that they felt nervous and worried when they failed to understand the spoken text.
In 2008, Yan and Horowitz studied Chinese learners’ perceptions of how anxiety, along with other personal factors, may influence their achievement in English. They reported that whatever the learners’ levels of anxiety, their comments were entirely associated with listening and speaking. They also found that anxiety and motivation were inversely correlated and that motivation was a strong predictor of language learning success. As far back as Gary (1975), the notion of a psychological advantage in focussing on developing listening comprehension skills at the early stages of language learning could reduce anxiety, which could lead to higher motivation, which can then lead to more language learning success was examined. Yan and Horowitz’s (2008) findings also support this claim.

Onwuegbuzie et al, (2000) found that learners reported higher levels of output anxiety than other forms of anxiety (Input Anxiety [listening], and Processing Anxiety and Output Anxiety [speaking]). Interestingly, input anxiety was found to be the most closely related to global foreign language anxiety, explaining slightly more than 40% of the total variance in the latter (Onwuegbuzie et al, 2000). This finding is understandable. Although learners may feel less anxious in a classroom because the teacher has provided an atmosphere conducive to collaboration and risk-taking, the fact that the learners still have to communicate in situations that are outside of the classroom may intensify their feelings of anxiety because they may not understand what is being said in these natural, authentic situations. According to Shang (2008), many listeners report that they experienced difficulty in making the transition from understanding classroom talk to understanding natural language. One way to reduce this anxiety, and to help learners function outside the classroom and away from the teacher, is to teach them strategies for listening and learning in general.

In a test situation, feelings of anxiety can be greatly exacerbated. Debilitating test anxiety has two components: the cognitive, involving worry, which Eysenck (1979) defines as "concern about one's level of performance, negative task expectations, and negative self-evaluation," and the emotional, which includes the feelings of "uneasiness, tension and nervousness" (p. 364) that people experience as a result of worry. Both negatively affect performance. This definition is still applicable today.
According to Elkhafaifi (2005a), students who reported higher listening anxiety had lower listening comprehension grades than students who reported lower anxiety. Rotenberg (2002) investigated whether the increasing use of standardized testing methods could have different effects on learners across language proficiency levels. The results of the study confirm that performance anxiety varies inversely with language proficiency, i.e., the lower the performance anxiety, the higher the language proficiency, and vice versa.

Time limits during test administrations are another significant variable that cause and affect the level of test anxiety among foreign language learners (Arnold, 2000). Searching for a means to reduce anxiety and improve listening comprehension, Arnold (2000) found that visualization strategies were successful in reducing test anxiety. Several test variables may also affect learners’ listening anxiety to a certain extent (Chang, 2008). In a test situation, test task characteristics are also important variables that affect test-takers’ performance (Bachman & Palmer, 1996). The characteristics of test tasks include previewing questions, multiple listening, sufficient background or linguistic knowledge, and being familiar with the test format.

In this section, I have reviewed the literature on language anxiety and students’ anxiety when it comes to listening. In the next section, I will discuss how anxiety may be reduced by having the tests delivered by a computer, which is a familiar format to many of the students at CDA.

**Computer-based assessments**

McLuhan’s (1964) phrase “The medium is the message” lays out a challenge for language teaching and testing personnel: does the manner in which information is presented truly affect the way it is understood (Gruba, 1994)?

The use of computers in testing has been around since the 1970s; however, it did not “catch on” in the L2 field until much later. Perhaps the main reason the L2 field has lagged behind in this area is because it has long promoted performance-based assessment, a form of assessment that does not lend itself as readily to computerized administration as do more traditional test formats (Chalhoub-Deville, 2001). However, in recent years, the computerized delivery of tests has become an appealing and viable medium for the
administration of standardized L2 tests in academic and non-academic institutions (Chalhoub-Deville, 2001). The Public Service Commission of Canada is a perfect example of this trend as it now administers their Reading and Writing tests online.

An explanation for the adoption of computerized testing in the L2 field is that it provides many advantages to academics and practitioners, such as test security, cost and time reduction, speed of results, automatic record keeping for item analysis and distance learning (Bugbee, 1996; Drasgow & Olsen-Buchanan, 1999; Mead & Drasgow, 1993; Parshall, Spray, Kalohn, & Davey, 2002; Smith & Caputi, 2005; Thelwall, 2000). It also has the advantage of quick and precise scoring, it can reduce logistical considerations of large-scale administration, and it can provide instant feedback and it can make a test adaptable to the test taker’s ability (Rover, 2001). Research has shown that the results of paper & pencil tests do not significantly change with computer-based testing (CBT) (Baumer, Roded, & Gafni, 2009; Coniam, 2006; Choi, 2003).

However, there are some disadvantages, or at least concerns, associated with computerized testing which Fulcher (2003) brings to light. He discusses different features and elements that test developers must take into consideration when developing a computer-delivered test, or merely when taking an existing paper & pencil test and putting it on the computer. For example, the interface design must be taken into account. As Messick notes, “the primary aim of good interface design is to reduce to a minimum construct-irrelevant variance that could be attributed to test method” (Messick, 1989). Test developers must be careful that they do not introduce construct-irrelevant variance due to test takers’ differing familiarity with computers (Kirsch, Jamieson, Taylor, & Eignor, 1998). Fulcher (2003) draws the test developer’s attention to elements such as icons and navigation through a program, the color used and font size. These are important considerations when developing a CBT because we do not want a student to be put at a disadvantage just because s/he does not understand how to use the computer program. However, that being said, with the proliferation of computers and different media in general in our society today, it is a safe assumption that a multimedia application is probably very familiar to most test takers in developed and/or Western countries and therefore is less anxiety provoking than other, less familiar test methods.
When using multi-media as part of the learning process or as evaluating the product of learning, one must keep in mind that too much information on the screen can lead to cognitive overload in students. According to Liu (2011), there are three factors that influence the effectiveness of using multimedia: the amount of visuals used; the pace of exposure; and layout design. If there are too many visuals, then the students can become overloaded or bored, or it may just be too much for them to effectively process the information. According to Mayer (2001), “when students need to focus their attention on abstract information, the use of pictures or animation creates an external stimulus [that is] competing for cognitive resources”. However, as mentioned earlier, the younger generation of today seems to be much more at ease with the computer, with visuals and with having to process information that is coming from the auditory and visual senses simultaneously. Berk (2009) concluded that “video clips are a major resource for teaching the Net Generation and for drawing on their multiple intelligences and learning styles to increase success in every student”.

Yet, change is never easy. Terzis & Economides (2011) set out to test the acceptance of the use of a computer-based assessment (CBA) among test takers. They based their constructs on other well established models of IT\(^3\) acceptance. The results showed that in order for students to accept using it, a CBA must be playful and easy to use, with careful design of the content. Moreover, the social environment and the facilitating conditions play an important role for the acceptance of CBA (Terzis & Economides, 2011). Some researchers have concluded that students find the use of CBA more promising, credible, objective, fair, interesting, fun, fast and less difficult or stressful (Croft, Danson, Dawson, & Ward, 2001; Sambell, Sambell, & Sexton, 1999). Matsumura and Hann (2004) have suggested that computer testing platforms could diminish students’ anxiety in language testing and, therefore, test takers could perform better.

The effect multimedia have on listening comprehension directly has not been studied extensively. Brett (1997) compared test takers’ success rates on comprehension and language recall through three different types of media: audio, video (answer questions using traditional paper & pencil), and multimedia. He found that higher levels of comprehension and language recall were achieved while listening in the multimedia

\(^3\) IT = Information technology
environment, as long as the task was not too complicated. Brett (1997) also found that multimedia-delivered listening comprehension tasks may be more efficient than the traditional audio only or video plus paper & pencil. Wagner (2010a) also found that fewer than half his participants watched the video, less than what he found in his 2007 study, which could be partially explained by the fact that they had to watch the video on a screen at the front of a class and then answer the questions in a student booklet instead of having a multimedia test.

Ockey (2009) examined the developments in CBT and concluded that, despite the challenges of finding better security features and “developing procedures for employing computer-adaptive testing (CAT) techniques to assess the multidimensionality of language constructs and creating scoring systems capable of measuring meaning and feeling of written and spoken discourse” (p. 845), computer-based testing will continue. He contends that the benefits of CBT outweigh the difficulties of meeting these challenges.

In this chapter, I have reviewed the literature on the communicative power of gestures, on the use of videos in listening comprehension tests, the anxiety felt among students taking listening tests, and finally, I have discussed the features of computer-delivered testing that must be kept in mind when developing a computer-based assessment. In the next section, I will explain the structure of an Assessment Use Argument, as outlined by Bachman and Palmer (2010). My study has been grounded in this theoretical framework.

**Assessment Use Argument Framework**

Why do teachers test their students? Many novice teachers ask when they should test, or how they should test their students, but rarely do they ask why they should test their students. According to Bachman and Palmer (2010), the question of why to test is of the utmost importance. They explain that the reason teachers test students is to collect data/information that will allow them to make decisions. For example, when a classroom teacher gives a quiz, the purpose is to verify whether or not the students are learning the material. If the quiz produces poor results, then the teacher may decide that the students need extra help and that she must revisit her lessons and repeat the concepts. If the quiz
produces good results, then the teacher can conclude that the students are following her and that they understand the concepts. The decision she makes, then, is to move ahead with the class content. This is a simple example. There are, of course, many other variables that need to be mentioned that could affect the students’ performance on a test, such as, what happens if the quiz was not well developed? Perhaps the content on the quiz did not actually reflect what was taught or it was actually measuring something else, or perhaps the teacher did not give the students enough time to practice. Test score interpretation must take into account all these factors that could come into play when a test is given.

The example above shows us how assessment can help shape decisions that need to be made. Presumably, the teacher wants to make the decision that will best help the students: one that will have positive and beneficial consequences for the students. If the test shows that more than half the class failed, then it would not make much sense for the teacher to introduce new content. She needs to review the material and ensure that the class understands. This course of action will be beneficial for the students because they will be able to go over the material an additional time, and, hopefully, will then be able to understand it.

If, for whatever reason, the teacher decides that she must continue despite poor results, one can presume that the consequences will not be beneficial to the students. They will get more confused and lost, and perhaps they will make complaints to their parents, who, in turn, may complain to the principal. The teacher may be asked to justify her decision of continuing on with the course material. She could be held accountable if these students do not pass the final exam and fail the course, because she failed to make the correct decision to review the work that the test confirmed the students had not mastered.

Concepts such as justification and accountability are gaining importance in the world of language testing. More and more, test developers and organizations are being held accountable for the tests that they develop and are being asked to justify their assessments. These two concepts, justifying an assessment and being held accountable

---

4 Assessment comes in many forms (observation, portfolio, conferencing, etc). For the purposes of this thesis, assessment and tests are the same.
for the proper use of that assessment, are the underlying factors of the Assessment Use Argument (AUA) as put forth by Bachman and Palmer (2010). According to them, test developers “need to be able to demonstrate to stakeholders that the intended uses of their assessment are justified.”

Traditionally, decisions were based solely on performance scores, without addressing issues of test use or consequences of test use (Mann & Marshall, 2010). Bachman and Palmer (2010) argue that the consequences of the intended use of tests are important considerations when choosing an existing assessment or when developing a new one. “An AUA provides a framework for investigating the extent to which the intended use of a particular assessment is, in fact, justified” (Bachman and Palmer, 2010).

“The AUA consists of a set of claims that specify the conceptual links between a test takers’ performance on an assessment, an assessment record, which is the score or qualitative description we obtain from the assessment, an interpretation about the ability we want to assess, the decisions that are to be made, and the consequences of using the assessment and of the decisions that are made.” (p.22) The links are illustrated in Figure 1 below. The AUA bridges these links and provides a framework in which the test developer and the test user can justify the assessment.

```
Figure 1: Links from test taker’s performance to intended uses (decisions, consequences) (Bachman & Palmer, 2010, p 23)
```

```
Consequences

Decision(s)

Interpretation(s) about test taker’s language ability

Assessment Record (Score, description)

Test taker’s Performance

Assessment Tasks
```
Using Toulmin’s (2003) argument structure of claims which are supported by data and statements, Bachman and Palmer (2010) have created a framework that provides a rationale and set of procedures for justifying the intended uses of the assessment – also referred to as assessment justification. The structure of the AUA consists of a series of four claims about (1) the beneficial consequences of an assessment, (2) the decisions that are to be made, (3) the interpretations that are made, and (4) the assessment records. Under each claim, there is a series of warrants that are stated. Warrants are statements that elaborate the claims. For example, a claim may state that an end-of-course test is meaningful. The warrant that elaborates this claim may be that the test is an achievement test whose score can be meaningfully interpreted as the level of mastery of the course content. It is doubtful that all stakeholders will accept this warrant merely at face value; therefore, in order to justify this warrant, the test developer needs to collect evidence that will provide the backing for this warrant. In fact, some stakeholders may disagree with the warrants stated and make a counter claim, saying that, in keeping with the example above, that the achievement test included material that was not covered in class. This would act as a rebuttal to the claim and the test developer would need to collect evidence that will convince the stakeholder that this was not the case. Backing of this kind could include showing where in the curriculum the material on the test was covered. This collection of backing to support the warrants and claims is the framework upon which the AUA is based and can be seen in Figure 2 below. Warrants and rebuttals can be stated for all claims.

In 2010, Mann and Marshall tested the suitability of the AUA as a framework for “the development and/or use of a test to assess deaf children’s nonsense sign repetition skills in BSL (British Sign Language). Although they used a partial AUA, they found its structure to be potentially useful in sign bilingual education in that it can inform decision-makers on different forms of intervention needed, as well as provide a transparent framework for assessment developers (Mann & Marshall, 2010).

Colby-Kelly and Turner (2007) articulated a partial AUA in order to attest to the usefulness of formative assessment in the L2 classroom. One claim that was made was that teacher-student feedback is useful. The warrant that was elaborated on this claim states “that teacher-student feedback will contribute to improve learner performance over
time” (p.31). Data was collected in order to provide backing for this warrant: “observations and learner interviews demonstrate that learner oral presentation performance improves over time with feedback” (p.31). A rebuttal to this warrant is “that teacher-student feedback will not contribute to learner performance over time”. The application of the AUA allowed the researchers to draw the conclusion that teacher-student feedback with a motivational component appears to be useful only when the students take it seriously (Colby-Kelly & Turner, 2007). These examples demonstrate that the structure of the AUA can indeed help test developers justify the usefulness of their tests.

Figure 2: Framework of AUA (Bachman & Palmer, 2010, p 104)
Bachman and Palmer (2010) argue that their approach to language assessment provides the following:

- A theoretically grounded and systematic set of principles and procedures for developing and using language tests
- An understanding that will enable test developers to make their own judgments and decisions about selecting, modifying or developing a language assessment whose use can be justified to stakeholders (p. 95).

The present study uses the AUA as a theoretical framework to guide the development of a listening test that uses videos as the medium of delivering the listening texts and to provide the justification of their inclusion in the construct definition of listening comprehension. Because this would be a high-stakes test if used officially, a complete AUA is articulated, and backing has been collected to support the warrants and claims stated, based on the testing environment at CDA. According to Bachman and Palmer (2010):

Assessment development consists of two parallel processes that serve two purposes. The assessment justification process, which includes the articulation of an AUA and the collection of backing, is aimed at justifying the assessment for its intended uses.

The assessment production process, which proceeds through the stages of planning, design, operationalization, and trialing, is aimed at producing an assessment.

These two processes yield two “products” that enable the decision-maker to use the assessment for its intended purpose (p. 430).

At the end of this study, there will be two products: the assessment justification in the form of the articulation of an AUA and a computer-delivered video listening test, which will complete the assessment production process.

**Summary**

In this chapter, I have reviewed the literature on the communicative power of visuals and gestures, using videos in testing listening comprehension, language anxiety in L2 students and I have discussed computer-delivered tests. I have also given an explanation of what an AUA is and how it is useful for test developers to justify the intended uses of the assessment. In the next chapter, I will explain the research design and explain the three phases that this research went through.
CHAPTER THREE
METHODOLOGY: INCLUDING SEQUENTIAL INTER-PHASE RESULTS

Introduction

In this chapter, I begin by explaining the rationale behind the study. Then I describe the general military educational context in which this study is situated and state the research questions. This is followed by an explanation of the research design in which I describe the three phases of the study. For each phase, I will describe the purpose, context, participants, instruments, the data collection and analyses conducted. There is a brief summary at the end of the chapter.

Rationale for the Study

The rationale for this study comes from the growing acceptance of including visuals in order to help L2 learners with the skill of listening. Assessing students’ listening comprehension has been a problematic issue at the Canadian Defence Academy (CDA) for some time and I believe that changing the method of testing will address this issue and provide some answers. Moreover, with the increasing pressure placed on test developers to justify the intended uses of their assessments, developing a test within an Assessment Use Argument (AUA) framework will provide justification for the use of videos in a listening test. Before launching into a new method of testing, it is wise to determine whether it will have beneficial consequences for the stakeholders. The present study gathers together the perceptions of three groups of stakeholders – the test developers, the MTCP teachers, and the MTCP students – of the benefits of using videos in a listening comprehension test. These data will support Claim 1 of the AUA, which states that the test will have beneficial consequences for the test takers. The AUA will be articulated with respect to the MTCP and military environment, and support for the warrants for Claim 2 (the decisions made based on the test will be equitable), Claim 3 (the interpretations made will be meaningful, impartial, generalizable, relevant and sufficient), and Claim 4 (the assessment records will be consistent) will come from documents and procedures that are specific to CDA and CFLS.
Military Context

Language training and testing in NATO

To situate this study, it is important to understand the general context in which the students at CDA and CFLS study languages. French and English are the official languages of NATO, yet English has become the operational language and members of the military are required to attain a certain level of proficiency in English to facilitate communication and interoperability among nations. As new non-English speaking countries join NATO, (particularly since 2004), there has been an increasing need for English language instruction and testing (Dubeau, 2006).

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 staff. 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 in Dubeau, 2006).

In order to compare individuals’ proficiency levels among different nations, a common scale of language proficiency must be implemented. The NATO STANDARDIZATION AGREEMENT (NATO STANAG) 6001 Language Proficiency Levels is the official scale used by NATO to assess language proficiency, and nations that agree to adopt it do so with the understanding that it will be used for:

1. Communicating language requirements for international staff appointments.
2. Recording and reporting, in international correspondence, measures of language proficiency, and
3. Comparing national standards through a standardized table while preserving each nation’s right to maintain its own internal proficiency standards (12 October 2010 NSA(JOINT)1 084(201 0) NTG/6001 ED 4).

Formal testing is used to assign levels of proficiency to individuals, which are then recorded as a Standard Language Profile (SLP). All positions within NATO have been assigned particular SLPs, which a person must meet in order to be appointed. Therefore, if individual members fail to meet the expected linguistic profile in all skills (listening, speaking, reading, & writing) for particular positions, they are at risk of being passed over
for promotion, for NATO postings, and for peacekeeping missions – all of which carry substantial financial benefits. The scores obtained from STANAG tests are high-stakes, in that individual careers hang in the balance.

**NATO STANAG 6001 Language Proficiency Rating Scale**

The NATO STANAG 6001 rating scale has evolved over the years and is based on the Interagency Language Roundtable scale (ILR) that originated in the United States in the 1960s. The STANAG measures general language proficiency, but not necessarily military content, and has evolved into a scale that is based on six levels:

- Level 0 = No proficiency
- Level 1 = Survival
- Level 2 = Functional
- Level 3 = Professional
- Level 4 = Expert
- Level 5 = Highly-articulate native

The scale covers the four skills – listening, speaking, reading, and writing – each with its own holistic description that states the tasks that a person who is at a particular level can perform, and the content areas that are appropriate for that level, as well as the level of accuracy that is required. For example, tasks at level 1 in speaking include maintaining a short conversation, asking and answering basic questions, and exchanging greetings, introducing and identifying themselves and talking about predictable personal and accommodation needs. The content areas for a level 1 speaker include basic needs such as ordering meals, obtaining lodging and transportation, and shopping. The accuracy expected at this level is not precise. Frequent errors in pronunciation, vocabulary, and grammar often distort meaning and time concepts are vague. Yet, this speaker can speak at the sentence level and may produce strings of two or more, simple, short sentences joined by common linking words.

Over the years, different nations had observed that the range of performance within a level was quite large and to award two candidates the same level, though their

---

5 To be referred to as only STANAG from this point on. For a full description of the levels in all skills, readers are invited to visit www.bilc@forces.gc.ca
performances were very different, often caused teachers and students to question the validity of the final score. Therefore, descriptions for plus levels were articulated, eg. Level 1+ or 2+, in order to better describe the proficiency of an individual. Each base level has a plus level that is defined as “a level of proficiency that substantially exceeds a 0 through 4 base skill level, but does not fully or consistently meet all of the criteria for the next higher base level” (12 October 2010 NSA (JOINT)1 084(2010) NTG/6001 ED 4). Canada adopted the plus levels in 2007, but they are not obligatory among the nations using the STANAG.

The Bureau for International Language Co-ordination (BILC) is the consultative and advisory body for language training matters in NATO. It 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 co-ordination 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 clearing house for the exchange of information between participating countries on developments in the field of language training.
- To be custodian of STANAG 6001. (retrieved from www.bilc.forces.gc.ca in Aug 2011)

One manner in which the BILC meets these responsibilities is by offering a two-week Language Testing Seminar (LTS), three times a year. This seminar introduces novice and experienced testers alike to the STANAG rating scale and allows them to practice creating items and to rate samples of written and spoken English, with the purpose of standardizing the interpretation of the STANAG. For graduates of the LTS, a three-week Advanced Language Testing Seminar (ALTS) is offered, which addresses each skill in more depth. The participants in these seminars gain much knowledge on the STANAG and on how to develop test items that reflect the standard. The BILC also holds an annual conference and seminar in order to keep abreast of language issues among nations, and to allow discussions of best practices among testers.
The Canadian Defence Academy (CDA) and the Canadian Forces Language School (CFLS) have been doing their part to help other countries’ militaries obtain their SLPs. The teaching and testing of languages for the Military Training and Cooperation Program (MTCP) are governed by two documents: the Qualification Standard (QS) and the Foreign National Training Plan (FNTP). According to the QS (2006), “principles of the Communicative Approach, adult education and second language acquisition shall be applied.” In the FNTP (2006) it states that “through this [communicative] approach, it is understood that knowledge of the structures and vocabulary of a language does not in itself constitute the ability to communicate in real-life situations. Language is seen, more broadly, as a continuous process of expression, interpretation, and negotiation, which transforms ideas, thoughts, and feelings into speech and writing. Any individual who has attained a measure of competence in this process is said to possess communicative competence.” Hence, the MTCP program is based on the communicative method and it is aimed at teaching lower proficiency students. This program is based on acquiring general English with some military content. Every year, approximately, 170 members of foreign militaries come to Canada for 19 weeks to learn English at CFLS in either Ontario or in Quebec. MTCP students have classroom instruction for six 54-minute periods every day and they also enjoy socio-cultural activities, such as trips to Ottawa and Quebec City, and simple outings such as curling and bowling. These activities allow the students to change their classroom environment while using the language in authentic situations. They also allow the students to become familiar with some aspects of Canadian culture.

Multi-level English general proficiency tests, based on the STANAG, in listening and reading are administered when the students first arrive in Canada and serve two purposes: (1) to help with placement in classes and (2) to obtain a baseline from which we can report on the progress made by each student at the end of the program. All four skills are tested at the end of the course in order for the students to obtain their SLPs. To maximize the number of hours of instruction, testing is reserved for the final weeks of the course. Because of these time constraints, all tests are necessarily multi-level and test STANAG Levels 1-3. Our tests do not test higher than Level 3 because the MTCP program is not
aimed at higher proficiency students; our student population is generally between Levels 0-2 in all skills, with the occasional Level 3.

In my position as the Chief of English Testing, I have been able to observe our foreign students as they take their final tests. Over the last several years, it has become evident that they have a great deal of difficulty with the listening test. Generally, the scores that are awarded to the skill of listening comprehension at the end of the program are lower than those of the other three skills. Although it is understandable that students may not have the exact same level of proficiency across skills, it is still a concern that many students appear to have progressed in the other skills, but not in listening. Since 2003, when CDA first adopted the STANAG, a high percentage of our students have completed their course with a Level 0 (no proficiency) in listening (see Appendix A for the full level descriptions for Listening Comprehension), and at least a level 1 (survival) in the other skills. Having received a Level 1 or higher in Speaking, the students have demonstrated that they do have some level of comprehension; yet this is not being reflected in their listening scores. This is a growing concern among the staff and the upper chain-of-command because, on paper, it looks as though the students have not made any progress in this skill after 19 weeks of being in Canada. The students themselves have mentioned that when their supervisors at home see a final score of 0, their participation in the course is questioned. It appears as though the student has failed part of the course and may be passed over for promotion.

Several factors can account for these low listening results: student motivation, students’ anxiety levels, or even the listening test construct itself, which was developed in a traditional audio-only format. In this thesis, I am going to focus on the test construct of listening. Presently at CDA, the STANAG listening test consists of 65 pre-recorded audio-only texts, with 65 multiple choice questions that increase in difficulty. Due to physical space constraints, the texts are played over a speaker in a testing room and all students answer the questions at the same pace. The testing conditions may contribute to the students’ feelings of nervousness and anxiety just before and during the test administration.

In the next section, I state the research question and objectives which guided the design of the study. I will then discuss the three phases of the study.
Method

Research Question and Objectives

The research question that I addressed in this study is the following:

1. To what extent is the AUA framework suitable for justifying using videos as a means of delivering a listening text in a multi-level video listening test?

To gather evidence for the beneficial consequences of using videos, I focussed on two objectives:

2. To what extent will different stakeholders (test developers, MTCP teachers, MTCP students) perceive the use of videos, as the medium of delivering listening texts, as being beneficial when testing listening comprehension?
3. To what extent will students report feeling less anxious when taking a video listening test?

The research question guided the design of this study, which is a mixed methods approach that follows an exploratory, sequential design, as described in Creswell and Plano Clark (2011), in three phases:

PHASE ONE
(a) Needs analysis and
(b) the development of a prototype

PHASE TWO
Development of a multi-level Video Listening Test within an AUA framework

PHASE THREE
Trial of Video Listening Test

Figure 3: Three phases of the research design (Creswell and Plano Clark 2011)

This design was chosen because it “[exploratory, sequential design] is particularly useful when the researcher needs to develop and test an instrument because one is not available” (Creswell, 1999; Creswell et al, 2004, Creswell and Plano Clark, 2011). This
was definitely the case, as there are no computer-delivered video listening tests based on the STANAG that exists for the MTCP clientele. The design is sequential in that the data I gathered from each phase informed the next. I gathered qualitative information through observation and a needs analysis. I then used this information to develop an instrument, a video listening test prototype, because one did not exist. Once I trialed the prototype, I collected qualitative information to help improve the test. The data allowed me to move onto Phase Two, which involved the development of the present 24-item video listening test, which is grounded in the theoretical framework of an Assessment Use Argument (AUA). Finally, once the test was developed, I began Phase Three of the study, which involved trialing the test with a sample of the target population, the MTCP students, along with two other important groups of stakeholders (test developers and MTCP teachers). The priority data that were collected were qualitative; quantitative data were collected to support the qualitative information. See Figure 4 for a summary of the steps taken and data collected during each phase of the study.
Figure 4  Exploratory sequential design(Creswell and Plano Clark 2011)
Phase One consists of two parts: (a) a needs analysis and (b) the development of a prototype.

**Phase One (a): Needs Analysis**

**Purpose:**

The purpose of conducting a needs analysis was to determine the appropriateness of a video listening test. If the listening needs of the MTCP students were mainly to listen to English over a radio or over the telephone, then using videos in a listening test would not be appropriate. If, however, the needs analysis showed that the students need to listen to English in face-to-face situations, then using videos in the listening test would be more appropriate, as it would be more authentic to the TLU situation.

**Context:**

Observational data were collected prior to this thesis during test sessions in the testing rooms. These rooms are large rooms that have tables and chairs arranged in rows. There are no windows and the listening test audio track is delivered over speakers in the room. A more formal needs analysis, in the form of a focus group meeting, was conducted at the Canadian Forces Language School (CFLS), located in Quebec. This meeting took place in a classroom at CFLS. The needs analysis took about two weeks to complete.

**Participants:**

Six teachers, five female and one male, from the Military Training and Cooperation Program (MTCP) program volunteered to participate in the focus group – all of whom
have been teaching the program for at least two years. Ideally the students from the
MTCP program would have been included in the focus group meeting; unfortunately, due
to timing issues, they had already left the country to go back home. However, the
teachers spend at least 700 hours with the students over the course of the program, during
which they have ample opportunity to get a sense of the students’ reaction to many
different activities, scenarios, and issues. Due to logistical constraints, the students were
not included in this phase of the study.

**Instrument:**

A series of five open-ended questions (Appendix B) were posed to the teachers for
discussion. The questions were focussed on *when* (in what situations) the MTCP students
needed to listen to others in English. In other words, what kind of listening do the MTCP
students need?

**Procedure:**

I first approached the senior teacher and explained my research. I gained permission
to invite the teachers to a focus group meeting where we discussed the listening needs of
their students. The participants signed consent forms (see Appendix C) that allowed the
session to be recorded for later transcription, analysis and reporting as part of this thesis.
The meeting lasted for 30 minutes.

**Data analysis:**

The data collected from the needs analysis were transcribed and then content analyses
were conducted. The data were categorized according to themes that emerged. The
qualitative data collected were used to inform the next phase: the development of the
prototype.
Phase One (b): Development of a Prototype Video Listening Test

Purpose:

Developing a prototype of a new product is good business sense. A company does not want to manufacture many copies of a new product if there is no proof that it will work. It saves money and time to create a prototype in order to test it out first, before investing too many resources. Language testing is no different, according to Fulcher and Davidson (2007), who state that when creating a new test, it is important to try out a prototype before creating many versions of the test, in order to establish the validity and reliability of this new test. The purpose of this phase is to ensure that the items work the way the test developers intend them to, that the items are valid measures of the language ability that is being tested. Having a multi-level video listening test is a new method of testing at CDA and within the military, and it was important to try it with colleagues to ascertain whether or not the test would work and to get an initial impression of the benefits of such a test.

Context:

The prototype test was developed at CFLS, Quebec. The items took into account the data collected from the needs analysis and were also in accordance with the STANAG. I followed general test development stages, eg. Initial Planning, Design, Operationalization, Trialling, and Assessment Use, as explained by Bachman and Palmer (1996, 2010). The items are general English and represent STANAG Levels 1-3. The videos were shot entirely in the studio of the Multimedia Production Centre (MPC), located at the St. Jean Garrison.

Participants:

There were two groups of participants for the prototype test: Group 1 consisted of those who helped develop the test, and Group 2 consisted of those who took the test and gave some suggestions for improvement.
**Group 1**

Three test developers, two male and one female, and three curriculum personnel, two female and one male, who worked at the Canadian Defence Academy (CDA) participated as “actors” in the videos. All “actors” had been working at CDA from 1-7 years and all were native English speakers. In addition to filming and editing the videos, one male member of the Multimedia Production Centre (MPC) also programmed the prototype using Adobe Flash Player 9, a multimedia platform that is used to add animation, videos, and interactivity to a webpage\(^6\).

**Group 2**

There were two different groups of participants – (a) Canadian teachers, and (b) international test developers. Group 2a consisted of six female Canadian Anglophone teachers who had worked at CDA and CFLS for at least 10 years, teaching English as a Foreign Language to the MTCP clientele.

Group 2b consisted of 14 international test developers who participated in the BILC-sponsored two-week seminar on Language Testing in Garmisch-Partenkirchen, Germany. The two facilitators of this seminar, one native English speaker and one non-native English speaker, who had been working with the STANAG 6001 to test English for at least 10 years, also took the prototype test.

**Instrument:**

With the data collected from the needs analysis and with the tasks outlined in the STANAG for Levels 1-3, texts were created for a computer-delivered multi-level video listening test.

The texts used for the items came from three sources: my professional experience with testing texts in the military context, the Internet, and some revised texts from a computerized version of the Canadian Forces English Curriculum. Each text was rated against the content and tasks statements for the appropriate level as outlined in the

---

\(^6\)See [http://en.wikipedia.org/wiki/Adobe Flash](http://en.wikipedia.org/wiki/Adobe%20Flash) for a more complete explanation. It is referred to as FLASH from hereonin.
STANAG. Therefore, the characteristics of each text met the criteria set out in the
STANAG. Though the “actors” were given scripts to learn, they were encouraged to
improvise during the filming in order to make the texts as natural as possible. The use of
gestures, facial movements, intonation and all non-verbal communication seen in the
videos are natural movements from the “actors”.

The items were created according to the content/task/accuracy statements that are
articulated in the STANAG level descriptions and which reflect the TLU domains\(^7\), as
confirmed by the needs analysis. Within the TLU domains, there are a mixture of
dialogues, monologues and discussions, which are, therefore, represented in the test. The
language elements to be tested are the ability to listen for the main idea, for explicit
information and for implicit information.

The multiple choice items were based on the listening texts after editing was
complete. The items went through an Item Review Board and revisions were made.
According to Fulcher (2007), this is known as Alpha testing. Because the programmer
was learning FLASH while actually programming the test, he found it easier to have
different weights for items at different levels. Together, we decided that the Level 1
items would be worth one point each, the Level 2 items would be 5 points each, and the
Level 3 items would be 10 points each.

Two assumptions were made in order to generate cut-off scores. The first was that
test takers must get at least two items correct at a specific level to be awarded that level.
The second assumption was that test takers must have at least two of the items at the
lower level correct and at least two items correct at the upper level to be awarded the
upper level. This means that if a test taker got only one level 1 item correct, but he got
two level 2 items correct, he would still only be awarded a level 1 because he did not
correctly answer at least two of the lower level. The same logic is applied for levels 2
and 3. Therefore, based on these assumptions, the following cut off scoring grid was
developed

\[
\begin{align*}
\text{Level 0} &= \quad 0 - 1 \\
\text{Level 1} &= \quad 2 - 11 \\
\text{Level 2} &= \quad 12 - 29 \\
\text{Level 3} &= \quad 30 - 48 
\end{align*}
\]

\(^7\) Target Language Use domains are explained in Chapter 2, section: Videos in testing Listening
Comprehension
The final product included nine videos that ranged in length from 0:41 min to 2:20 min, three at each level (STANAG levels 1-3), and one multiple-choice item following each video. It took 20 minutes to complete. The video appears at the top left-hand side of the screen with the item appearing below it. Time is given to the students to preview the question before the video begins. A coloured bar is seen on the right-hand side of the screen, which acts as an item timer that counts down to indicate to students how much time they have to answer the question. Once the item timer is finished, the test moves onto the next item. See Figure 6 for a view of the interface of the prototype test.

The programmer at the MPC put the test on a CD, which can be played on any computer.

![Image of the prototype test interface]

**Figure 6 The interface of the prototype**

**Procedure:**

I asked for volunteers who may be interested in taking the test and giving me some feedback. For the participants in Group 2a, the Canadian teachers, the video test was played on their classroom computer. Comments were given to me during one-on-one
conversations after they had taken the test. For the participants in Group 2b, the international test developers, the test was projected onto a screen in the front of the class and all answers were called out, due to time constraints. Comments were collected through a group discussion. As this trial was informal, all feedback was recorded by hand and the participants were asked orally for permission to report their comments.

**Data Analysis:**

The feedback from the participants was transcribed and categorized. Certain recommendations for improvement were noted, such as giving more time to preview the question, allow the candidate more control to advance to the next item when ready, and some formatting suggestions. These recommendations were used to help create the video listening test in Phase Two.

**Phase Two: Development of a multi-level video listening test within an AUA framework**

![Diagram showing the phases of development]

*Figure 7: Phase Two Development of multi-level Video Listening Test within an AUA framework*

**Purpose:**

The reaction and feedback from the prototype trial prompted me to develop a more complete video listening test and ground it in a theoretical framework, the Assessment Use Argument (AUA) discussed in a previous section. This framework is a tool that will help justify the inclusion of videos in a listening comprehension test. Test development is a principled process and consists of various stages: (1) initial planning, (2) assessment design, (3) operationalization, (4) trialing, and (5) assessment use (Bachman and Palmer,
Due to the high-stakes nature of the test, I went through each of these stages rigorously and I have reported on each of them below. Documents, such as a design statement, a blueprint, and item specifications, were produced and helped guide the activities involved in the following stage. The test used in this study is still in the trialing stage and has not yet reached the final stage of assessment use.

**Context:**

The videos were shot at the MPC both in the studio and on location in and around the Garrison.

**Participants:**

Two test developers and four curriculum personnel helped in the development of the video listening test as “actors”. The two test developers, one female and one male, both aged between 35–45 years, have worked at CDA in the testing section and with the STANAG, for the past 10 years. Members of the curriculum staff, 3 female and 1 male, aged between 30-40 years, have worked with CDA for a minimum of two years. All are native English speakers. No one who “acted” in the videos is a professional actor, and all gestures and body movements are as natural as possible, considering they were quite aware they were being filmed. The personnel who work at the Multimedia Production Center (MPC) shot and edited the videos. They are all male, aged between 25-30 years, and all have experience in filming.

**Instrument:**

The present video listening test consists of one section that includes 24 multiple-choice items that are ordered in ascending levels of difficulty, as defined by the STANAG, and delivered by computer. Each task is designed to test a particular aspect of the listening construct that has been operationalized as the ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video.

The software used to program the test is FastTEST Pro, a commercial program that can be purchased from Assessment Systems Corporation. The videos were put into
Windows Media files, then into AVI files, and were all burnt onto a CD. FastTEST Pro only accepts AVI files; therefore, I was able to simply add the videos to the items as I inputted them into the software. Because this is a general proficiency test, it is necessarily multilevel. The items reflect levels 1, 2 and 3 as defined by STANAG. All items are independent of each other and there is one item per text. After the editing was completed it was deemed that 30 minutes was the time allotted for the entire test. After much thought, it was decided that the videos were to be watched only once, as is the case in the current official listening test (the audio texts are only played once).

The video listening test is criterion-referenced and the target language use domains are those described in the NATO STANAG 6001 Level Descriptions. There are ten Level 1 items, ten Level 2 items, and four Level 3 items. The length of Level 1 items varies between 15 seconds and 1:04; Level 2 items vary between 40 seconds and 1:50; and Level 3 items vary between 1:17 and 2:17. The items are scored automatically by the program. No cutoff scores were calculated.

The semi-contrived and contrived texts include dialogues and monologues on general English, and came from two sources: my professional experience with testing texts in the military context and the Internet. As with the prototype test, the actors were told to improvise during the filming to make the dialogues as natural as possible. Therefore, the use of gestures, facial movements, intonation and all non-verbal communication seen in the videos are natural movements from the “actors”.

The interface was an important element to consider – the screen could not be too busy with too much information, yet the test taker had to be able to navigate through the test independently.

The format of the test is as follows: the test taker sees the cover page and has to enter his/her ID number (which is supplied by the proctor) and the date. To advance to the next page, s/he clicks on the arrow at the bottom right-hand corner. Next, the test instructions are presented, followed by an example. Once the student feels ready, s/he can begin the test. Once the test advances to Item 1, a test timer starts counting down. The timer can be seen on the bottom of the screen, in the middle. A test timer was used instead of an item timer as in the prototype because it gives the test taker more control over how long s/he can stay on any one item. They know that they have 30 minutes to answer all 24
items, but if they spend a little more time on one item over another, then that is a decision made by the test taker. If weaker candidates do not finish the test, that is OK. A consequence of that is that they are not exposed to higher level items that they really do not need to see. They will be able to do what they can in the time given, which gives some control to the test takers. This kind of control is lacking in traditional tests where the audio track dictates the speed at which a test takers completes the test.

The item is on the left-hand side of the screen and the test taker can take the time needed to preview it before beginning the video. One of the criticisms of the prototype was that the test taker did not have enough time to preview the item before the video started. Now, the test taker can take the time s/he needs to understand the question and know the purpose for listening. Then, when s/he is ready to watch the video, s/he clicks on the words “Click for video”, which are written under the item. The video will appear on the right-hand side of the screen. Once the video has finished, the test taker then answers the multiple-choice item by clicking on the button next to the answer choice. S/he can then advance to the next page and the next item. (See Figure 8)

Figure 8: Interface of Video Listening Test
If the test takers feel as though they have already decided on their answer before the video is finished, they can stop it and exit the video to answer the item. They can also pause the video if they feel the need to read the item again. This sense of control is important for the test takers’ sense of accountability for their performance on a test (See Figure 9). The video can only be seen once, so to be able to pause it to re-read the item is an important consideration for the candidate.

*Figure 9 Controls for the video*

**Procedure:**

Due to the high-stakes nature of the test, the creation of the video listening test went through four stages of test development: initial planning, assessment design, operationalization, and trial. The final stage, Assessment Use, which involves the implementation of a test in an official capacity, is beyond the scope of this study and will not be discussed here.
Procedure - Stage One: Initial Planning

As Bachman and Palmer (2010) stated, “careful planning initially will help guide the development of the test”. They also state that “careful planning and implementation will enable the test developer to justify the intended uses of the assessment and hence, to be accountable to stakeholders” (2010). Because this is a high-stakes test, careful planning is extremely important; the decisions that will be made on the basis of these results can affect people’s careers.

The initial planning stage of test development includes a series of questions concerning the beneficial consequences of the assessment and the decisions that will be made on the basis of using this assessment. Further questions regarding the resources that will be required will also be addressed. The answers to these eight questions provide the basis on which the decision of whether to use an existing assessment or to develop a new one is made.

In this section I have answered these eight questions as part of my initial planning to develop the computer delivered video listening test.

1. What beneficial consequences do we want to happen? Who will benefit?
   
i. To make more accurate inferences on test takers’ listening proficiency as defined by NATO STANAG 6001 Level Descriptors
   
ii. The test takers will benefit because a truer picture of their listening ability will be represented; they have the opportunity to take advantage of using non-verbal communication that is inherent in “real life” listening contexts. The listening texts will be more authentic.
   
iii. Test takers will be given the opportunity to focus on the videos, thereby reducing their nervousness, which will lead to better performance.
   
iv. The teachers and senior teachers will benefit, because they will see the tests as being more accurate and fair, and they, in turn, will put more faith in the test. They will also tailor their teaching to help make test takers aware of using non-verbal information in their listening activities. Therefore, there will be a positive washback effect on their teaching and in the classrooms.
v. The Canadian institutions will benefit because their international reputation of fair and true testing practices will increase and they will be seen as being on the “cutting edge” of computerized testing.

vi. Home country institutions will benefit because they will be able to truly rely on the score of the test takers, and can assume that the test taker will understand others in particular circumstances.

vii. The test developers will benefit because they will be able to develop tests that are more authentic and that better reflect the tasks in the TLU domain. The tests will be seen as being “more valid” and fair, which will improve the test developers’ credibility.

2. What specific decisions do we need to make to help promote the intended consequences?
   i. To award a particular level that genuinely reflects a test taker’s true listening ability. Otherwise, if a test taker is rated at a higher level than he really is, then the consequences of that decision could be that this person is promoted to a position that he is not ready for. An even graver consequence could be that if this person is in battle, and does not understand what is being said, it could result in his being killed or someone else being killed.

   ii. If a test taker is rated at a lower level than he really is then he may be passed up for promotion and feel demoted and unmotivated in his job.

3. Who will be affected by these decisions?
   a. Who are the intended test takers?
      i. The test takers are from various NATO and Partnership for Peace countries, who come to Canada for 19 weeks to study English.
      ii. The test takers are generally officers in their own militaries. They are all university educated and have been in the military for at least 5 years.
      iii. The majority of the test takers are men, although there have been up to five women in the same cohort. They are all between the ages of 25-50 years.
b. Who else will be affected?

Teachers, senior teachers, the test takers’ supervisors at home, the upper hierarchy of the MTCP program in Canada, test administrators, and home country administration. All these stakeholders will benefit from having test scores be a better representation of a test taker’s listening ability. Test developers will also be affected by the decisions that are made because their reputations are on the line if the decisions made do not reflect the test taker’s listening ability.

4. **What do we need to know about the test taker’s language ability in order to make the intended decisions?**

Testers need to have an intimate knowledge of the STANAG 6001 Level Descriptions. These level descriptions are what govern our test development in all skills. Test items must adhere to these levels as specific tasks, grammatical accuracy and content areas are defined. Therefore, in order to make the intended decisions, test developers must know where a test taker’s listening ability falls within the STANAG 6001 Level Descriptions.

5. **What sources could we use or are available for obtaining this information?**

   i. Information collected by the classroom teacher

   ii. Profiles obtained in test taker’s home country

   iii. Develop own assessment

6. **Do we need to use an assessment to obtain this information?**

   i. Yes. Since we are measuring general listening proficiency, information gathered in the classroom is not appropriate. This test is not tied to the curriculum followed by the students.

   ii. Not all students are tested in their home countries, and for those that are, there is a problem with international standardization of the interpretation of the levels.

   iii. Students need to obtain a valid SLP before leaving Canada. This is done through formal testing.
7. **Is an existing assessment available?**

Yes, there is an existing listening test that is currently being used. However, it is a traditional audio-only listening comprehension test and it does not provide the non-verbal cues that are so important in listening comprehension.

   a. Is an existing assessment available that provides the information that is needed for the decisions we need to make?
      
      No
   
   b. Is this assessment appropriate for our intended test takers?
      
      No
   
   c. Does this assessment assess the areas of language ability in which we’re interested?
      
      No, because the proposed listening test is attempting to measure listening ability through a different medium – through the use of videos. The rationale behind this new test is that test takers will be able to utilize the verbal and non-verbal communication inherent in real world situations in order to help with their listening comprehension in English. There is no existing general proficiency test of listening comprehension that uses videos.
      
   d. Does the test developer provide evidence justifying the intended uses of the assessment?
      
      A complete Assessment Use Argument has been articulated and the present study will provide evidence justifying the intended uses of the assessment.

   e. Can we afford it?
      
      Yes. The costs involved include the time and services of the Multimedia Production Center, the resources of the testing section, the time given by the MTCP teachers and students.

8. **Do we need to develop our own assessment?**

   a. How will we assure that the assessment results are consistent and the interpretations meaningful?
i. Close adherence of the test items to the level descriptors will be monitored and this will ensure the relevancy to the decisions to be made. The items will go through an Item Review Board to determine their usefulness. They will then go through a native speaker trial, to confirm that what the test developers had intended is what is really understood. Next the test will go through a trial period with the population concerned. It will also go through a validation stage, and cut-off scores will be decided. By going through this process, the test will give meaningful results. The procedure will ensure the impartiality to all groups of test takers. These appropriate procedures for estimating the consistency of scores, and the meaningfulness and relevance of the interpretations, will be put into place and followed during the development of the test.

b. What resources will we need for the development and use of the assessment (including justifying the intended uses of the assessment)?

i. Human: administrator to direct and monitor the progress of the test development effort; people with expertise in the STANAG level descriptors to provide input into the selection of listening passages; people with the expertise in language assessment to guide the development of listening tasks and scoring procedure, people with multimedia expertise to film the videos and edit the texts, along with programming the test to make it user-friendly, people who are willing to “act” in the videos.

ii. Material: copies of listening texts, CDs, paper, personal computers, statistical analysis software, space/computers for administering listening test, cameras, place/studio for filming, and appropriate software for editing the videos.

iii. Time: to develop the listening texts; to moderate items, to film/edit; to program; to trial the test with native speakers and with the target population; to administer the test once it has become official
c. What resources do we have or can we obtain?

All resources, human, material and time, are available at CDA and at CFLS. Permission has been granted to the researcher to invest the time needed in the development of this test.

Based on the responses to these questions, I decided that there was no existing assessment that I could use therefore I would have to develop one. I decided that although many stakeholders that would be affected by this assessment are mentioned in the initial planning, I made the decision to focus my study on only three groups: the test developers, the MTCP teachers and the MTCP students. I felt that these groups were the most accessible and that they would be most directly affected by the assessment.

When these two decisions were made, I approached my supervisor at CDA and the Chief of the Multimedia Production Center to explain my research and to ensure that the necessary resources would be available to me. Once I gained the necessary approvals that would allow me to move ahead, I was then able to progress to the next stage: Assessment Design.

*Procedure - Stage Two: Assessment Design*

In this stage, test developers envision what the test will entail and what it will look like. This “vision” of the test is realized in a Design Statement. “The Design Statement is a document that states what one needs to know before actually creating an assessment” (Bachman & Palmer, 2010). This document is used to guide test developers in creating items and assembling tests. It also provides backing for the warrants stated in the AUA.

The Design Statement includes ten parts that should be answered in detail in order to provide item developers the necessary information they will need to complete their task. Much of the information found in the Design Statement is a more fleshed-out response to the questions asked in the initial planning stage.

The following is a completed Design Statement for the present video listening test.
DESIGN STATEMENT

1. **DESCRIBING THE TEST TAKERS AND OTHER STAKEHOLDERS**

Table 4

*Attributes of Stakeholders*

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Test takers</td>
<td>Members of military from NATO and PfP nations, participating in the Military Training and Cooperation Program.</td>
</tr>
<tr>
<td></td>
<td>The students are adults, ranging in age from 25 to 45 years.</td>
</tr>
<tr>
<td></td>
<td>Ninety percent of students are male.</td>
</tr>
<tr>
<td></td>
<td>There are varying levels of English proficiency.</td>
</tr>
<tr>
<td></td>
<td>There are varying levels of familiarity with computers.</td>
</tr>
<tr>
<td>2. Teachers</td>
<td>Teachers who work at the Canadian Forces Language School.</td>
</tr>
<tr>
<td></td>
<td>The majority of teachers are female, aged between 26 and 60 years.</td>
</tr>
<tr>
<td></td>
<td>The majority of teachers have experience in teaching ESL and in teaching the MTCP program more specifically.</td>
</tr>
<tr>
<td></td>
<td>Many have a lot of familiarity with computers.</td>
</tr>
<tr>
<td>3. Test Developers</td>
<td>Come from different NATO and PfP nations.</td>
</tr>
<tr>
<td></td>
<td>The test developers are adults, aged between 36-45.</td>
</tr>
<tr>
<td></td>
<td>Many have a lot of familiarity with computers.</td>
</tr>
<tr>
<td></td>
<td>Differing levels of English proficiency.</td>
</tr>
<tr>
<td></td>
<td>Differing levels of testing experience: from none to 2 years.</td>
</tr>
<tr>
<td></td>
<td>Differing knowledge of and experience using the NATO STANAG.</td>
</tr>
<tr>
<td></td>
<td>6001 Language Proficiency Levels: from none to 2 years.</td>
</tr>
</tbody>
</table>

2. **DESCRIBING THE INTENDED BENEFICIAL CONSEQUENCES**

Although the final decisions are made in the home country, and we have no information on what decisions are actually made there, it is of utmost importance that our tests accurately measure the skill we are testing. The decisions that are made on site, by the test developers, concern the level that is awarded to the students. These are important decisions, because the level that is awarded to the student will have consequences for his career (after 19 weeks in an intensive English course in Canada, the level he obtains in listening can be the deciding factor in whether or not the student will be promoted).

As Bachman and Palmer (2010) explain, the beneficial consequences are stated in Claim 1 of the AUA; however, “in order for these intended beneficial consequences to
guide assessment development and use, they need to be stated in greater detail in the Design Statement.”

Table 5
Describing the intended beneficial consequences

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Intended beneficial consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Of using the assessment</td>
</tr>
<tr>
<td>Test takers</td>
<td>The test takers will be able to use non-verbal cues to help them with comprehending the verbal language. The speakers will be clearer and the context will be clear for the students, thus reducing their level of anxiety and allowing them to focus their attention while they listen to the dialogues and monologues. They will see the video listening test as being an authentic test that more closely resembles the TLU domain. They will view the VLT as a valid approach to testing listening.</td>
</tr>
<tr>
<td>Teachers</td>
<td>Teachers will be able to use more videos in class, which is stimulating and interesting for the students. They will be able to teach strategies that encourage the inclusion of non-verbal cues, which will lead to using more authentic material and tasks that are closer to the TLU domain. They will see the test as being more authentic and therefore fairer for the test takers. They will see this as a valid approach to testing listening comprehension.</td>
</tr>
<tr>
<td>Test developers</td>
<td>Test developers will be able to develop tests that are more authentic that better reflect the tasks in the TLU domain.</td>
</tr>
<tr>
<td></td>
<td>Of the decisions that are made</td>
</tr>
<tr>
<td></td>
<td>If students take a listening test that better reflects the TLU, then the interpretations and generalizations of their proficiency level will be more accurate and valid. This will allow the test developers to award a level, according to the STANAG, with confidence that the students’ scores are just and fair. This will allow the supervisors in the home country to make decisions about promotions for the students.</td>
</tr>
<tr>
<td></td>
<td>Teachers will benefit from having students very interested and engaged when watching videos. They can then teach different listening strategies.</td>
</tr>
<tr>
<td></td>
<td>The scores derived from the video listening test will be accurate measures of the students’ listening ability because of the videotexts will be received favourably by the students, reducing their anxiety levels, thereby allowing them to fully focus on the videotext. They will see the videotext as being more authentic and pertinent.</td>
</tr>
</tbody>
</table>
3. **Describing the Decisions to Be Made and Individuals Responsible for Making These:**

Table 6

*The decisions, stakeholders affected by decisions, and individuals responsible for making the decisions (Bachman & Palmer, 2010, p. 278)*

<table>
<thead>
<tr>
<th>Decision</th>
<th>Stakeholders who will be affected by the decision</th>
<th>Individual(s) responsible for making the decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Award Level 0/0+</td>
<td>Students and teachers in the MTCP program</td>
<td>Test developer</td>
</tr>
<tr>
<td>Award Level 1/1+</td>
<td>Students and teachers in the MTCP program</td>
<td>Test developer</td>
</tr>
<tr>
<td>Award Level 2/2+</td>
<td>Students and teachers in the MTCP program</td>
<td>Test developer</td>
</tr>
<tr>
<td>Award Level 3</td>
<td>Students and teachers in the MTCP program</td>
<td>Test developer</td>
</tr>
<tr>
<td>Career altering decision</td>
<td>MTCP students</td>
<td>Home country</td>
</tr>
</tbody>
</table>

4. **Determining the Relative Seriousness of Classification Errors and Policy-Level Decisions About Standards**

*False positive classification errors.* If students are rated at a higher level of proficiency in listening comprehension, this may have detrimental consequences for students because their supervisors will expect them to be able to understand certain situations that they do not. In battle situations, if a student has been rated at a higher level of proficiency than they are, the consequences could be a misunderstanding that could result in injury or even death.

Students may be promoted to positions for which they are not ready, which can result in higher levels of stress and poor work performance.

*False negative classification errors.* If students are rated at a lower level of proficiency in listening comprehension, this may have detrimental consequences because these students may be passed over for promotion. This may lead to feelings of disappointment and less motivation in their job. It may also support an erroneous view of
their own listening level, as many students underestimate themselves in terms of their listening capabilities.

One way of mitigating the detrimental consequences of both decision classification errors if they occur is to compare the test taker’s listening and speaking scores. If the test taker receives a Level 0 in listening and a Level 1 or higher in speaking, then a different test developer from the one who administered the interview can relisten to the speaking test (which has been recorded) to ensure the correct rating was given. This relisten is done right away, before the official results are given to the candidates. However, because of the tight testing schedule that we must work under, there really is no room for error: the test takers receive their results and then they get on a plane the next day to return to their respective home countries. For this reason, we need very precise and accurate evaluation tools to minimize these detrimental consequences.

5. DEFINING THE CONSTRUCT TO BE ASSESSED

The following definition of listening as “an active process in which listeners select and interpret information which comes from visual and auditory clues in order to define what is going on and what the speakers are trying to express” (Rubin 2007) is the overarching definition of listening. However, in this study it was operationalized as:

The ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information

This construct definition is based on Wagner (2002) and on the tasks that are outlined in the NATO STANAG 6001 Language Proficiency Levels for levels 1 – 3. If one looks at the trisection (see Appendix D) of the level descriptions that are supplied by the BILC, in accompaniment to the full level descriptions, one can see that at level 1, students are expected to understand the main idea of a simple dialogue. At level 2, they are expected to be able to understand concrete, factual information taken from the news or a more complex dialogue. At level 3, students are expected to understand what is between the lines, or implicit information, in professional settings. This construct definition of listening allows the test developer to create assessment tasks that pinpoint these specific areas of listening ability.
6. **IDENTIFYING AND DESCRIBING THE TLU DOMAIN**

The STANAG rating scale is used to measure general proficiency in English. The MTCP course is focussed on general English, with some military content. It is, therefore, easy to conclude that the TLU domain of the students at this level is the use of English in general situations and not specific military contexts. For example, the students may meet members of foreign delegations and engage in general conversation. Specific military content is not necessarily used in these conversations; that would fall under English for Specific Purposes. The STANAG is used to measure general proficiency, and it provides the test developer with the TLU domain at the different levels. The trisection of the level descriptions is a useful tool for test developers to identify the content areas that are specific to each level, the required tasks and the required level of accuracy for each level.

7. **SELECTING TLU TASKS AS A BASIS FOR DEVELOPING ASSESSMENT TASKS**

The TLU tasks that were selected were based on the comments made by the Item Review Board, the number of revisions that needed to be done, and the feasibility for filming, in terms of location and the availability of actors.

8. **DESCRIPTION OF THE CHARACTERISTICS OF TLU TASKS THAT HAVE BEEN SELECTED AS A BASIS OF AN ASSESSMENT TASK (TABLE 7)**

<table>
<thead>
<tr>
<th>Characteristics of TLU Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Setting:</strong></td>
</tr>
<tr>
<td>Physical characteristics: restaurant, workplace, lecture hall, meeting room, lounge</td>
</tr>
<tr>
<td>Participants: the student, colleagues, professor,</td>
</tr>
<tr>
<td><strong>Rubric (all implicit in the TLU task)</strong></td>
</tr>
<tr>
<td><strong>Instructions</strong> Target language: written, visual, or internally generated by the student; specification of procedures and tasks based upon students’ knowledge of general English</td>
</tr>
<tr>
<td><strong>Structure</strong> Number of parts: one part</td>
</tr>
<tr>
<td>Salience of parts &amp; Tasks:</td>
</tr>
</tbody>
</table>
Sequence of tasks: ascending order of difficulty

Relative importance of tasks: all are of equal importance

Number of tasks: 24 tasks: 10 Level 1, 10 Level 2, 4 Level 3

**Time allotment (per task)**

Highly variable

**Recording method**

Criteria for recording: As all tasks are multiple choice, the tasks are scored dichotomously

Procedures for recording the response: Using the mouse, the students need to click on the radio button next to their answer choice on the screen

Explicitness of criteria and procedures for recording the response: fairly explicit as the task is in a familiar format for most students

Recorders: the computer program, FastTEST Pro, records the scores automatically

**Input:**

**Format**

The input is delivered through videos, which allows for aural and visual input. The video texts are given in English, on the computer screen, next to the item with a natural speed of delivery, and natural gestures used

**Language characteristics**

Grammatical: range from simple to more complex

Textual: dialogues and monologues

Functions: survival, functional, professional

Genre: general English

Dialect: Canadian English

Register: informal and formal

Naturalness: natural

Cultural references: variable

Figures of speech: variable

**Topical characteristics**

Those identified in the NATO STANAG 6001 Language Proficiency scale, that are appropriate for Levels 1-3

**Expected Response:**

**Format**

Aural/visual; Lang/non lang; native/target; short; live, natural speed of delivery
### Language characteristics

| **Grammatical** | n/a as expected response is to click a button |
| **Textual** | n/a |
| **Functions** | survival, functional, professional |
| **Genre** | n/a |
| **Dialect** | n/a |
| **Register** | n/a |
| **Naturalness** | n/a |
| **Cultural references** | n/a |
| **Figures of speech** | n/a |

### Topical characteristics

Those identified in the NATO STANAG 6001 Language Proficiency scale, that are appropriate for Levels 1-3

### Relationship between Input and Response

| **Type of external interactiveness** | Non-reciprocal interaction between test takers, video listening text, and answering the multiple choice item, by clicking on the radio button on the screen |
| **Scope** | Narrow |
| **Directness** | Direct, Indirect |

9. **Plan for Collecting Backing and Feedback**

A plan for how to go about collecting evidence that will support the claims stated in the AUA must be well laid out. In order to collect evidence/backing that using videos in a listening comprehension test will have beneficial consequences for the test taker (Claim 1), the plan is to: Develop a prototype, trial it, collect feedback. The data collected from this phase will be incorporated into the next phase, which will be to design a computer delivered video listening test that is composed of many items. Once this test is developed, Phase Three (Trial) can begin.

10. **Plan for Acquiring, Allocating, and Managing Resources**

CDA has approved the development of this test
Once the design of the assessment has been completed, Stage Three can now begin. The Operationalization stage is where the actual test items are created.

**Procedure - Stage Three: Operationalization**

The test developer now turns the Design Statement into a document that is more operational in order to create assessment tasks/items. A blueprint of the test, or test specifications, is drawn up which includes information that will guide the test developer and “includes a description of the overall structure of the assessment and the specifications for each type of task to be included in the assessment” (Bachman and Palmer, 2010). The information found in the blueprint can also provide different stakeholders with interpretative information about the test. The reader is referred to Appendix E to see the blueprint for the present video listening test.

An item that has been created according to the blueprint contains information that needs to be recorded, such as what it is intending to measure, how it relates to the TLU, and what type it is (whether it be multiple choice or short answer). Being able to refer to this type of information allows the test taker to choose items for a test that do not repeat each other and items that test different aspects of the TLU. The item specification form for item 1 in the present test can be found in Appendix F.

During this stage, while following the blueprint, I created many texts and items at different levels. Twenty-four items were chosen to be in the test. The choices were based primarily on feasibility, in terms of location and availability of actors. The videos were shot over a four-month period. The multiple choice items were then developed based on the final version of the videos. These items were revised at an Item Review Board with my colleagues at CDA. After the videos had been filmed, I learned the FastTEST Pro test software in order to assemble the test. Once the videos were edited, they were presented on a CD in both Windows Media File and in AVI formats. FastTEST Pro accepts AVI format only for video.

Great care was taken to ensure that the test tasks reflected the TLU tasks that the MTCP students would be involved in. This was to make the test more authentic and useful. According to Bachman and Palmer (1996), assessment tasks should mirror, as
closely as possible, the tasks in real-life situations. They define “authenticity” as “the degree of correspondence of the characteristics of a given language test task to the features of a TLU task” (p.23). If the test tasks resemble the tasks that the test takers will do in real life, then that adds to the construct validity of the test. Generalizations and inferences can then be made more reliably to a test taker’s performance outside of a test situation.

Data Analysis:

Analyses on the videos included the overall visual appearance of the video, the content (how natural it sounded) and the production quality. IRBs were conducted to review the items and all revisions to the items were done during this phase.

Once the videos had been filmed, edited and added to the items on the computer, I was then able to move to stage four of test development – trialing. This stage is the focus of Phase Three of this study.

Phase Three: Trial of Video Listening Test

Figure 10: Phase Three: Trial of the Video Listening Test
Purpose:

The purpose of trialing this video listening test was to collect evidence from three different groups of stakeholders that would provide backing for the claims made for beneficial consequences in the AUA. It is exploratory in nature because this method of testing listening is unfamiliar in the MTCP context and before any kind of advancements can be made, it is important to first ensure that this new method will be accepted by the stakeholders. Therefore, this study examined the perceptions of three groups of stakeholders - the test developers, the MTCP teachers, and the MTCP students - on the benefits of using videos as a means of delivering the listening text in a test situation.

Context:

The trialing of the video listening test took place in two different locations: (1) at the Language Testing Seminar (LTS), Garmisch-Partenkirchen, Germany and (2) at the Canadian Forces Language School (CFLS). At CFLS, the test was administered individually to the teachers and the students, in an office on a laptop computer. At the LTS, the test was administered to the group of test developers, where the videos were projected on a screen at the front of a classroom and the participants answered the items in a test booklet.

Participants:

There were three groups of participants: test developers, MTCP teachers, and MTCP students.

Group 1: Test developers

Eleven test developers from several NATO and PfP countries from the LTS\(^8\) trialed this video listening test. The participants in this seminar ranged in experience in language testing from no experience to at least two years. Many of them were not familiar with the STANAG rating scale. The age of the test developers ranged from 26 to over 46, with the majority (55%) between the ages of 36-45. Eighty-nine percent of them reported having a lot of familiarity with computers.

\(^8\) Language Testing Seminar
Group 2: MTCP teachers

Ten teachers who work at CFLS participated in this study. The age of the teachers, all female, ranged from 26 to over 46, with the majority being over 46 years (60%). The majority of these participants, 60%, reported having a lot of familiarity with computers. There is a fairly even spread of the number of years teaching ESL, but most of the teachers have been teaching the MTCP for under five years.

Group 3: MTCP students

A stratified random sample of the students was taken, resulting in ten students from CFLS who are members of a foreign military studying English in the Military Training and Cooperation Program (MTCP) that participated in this study. The age of the students ranged from under 25 to 45 years, with half of them being between the ages of 36-45. The majority, 60%, reported having some familiarity with computers. Ninety percent reported as having studied English continuously within the past 5 years, although a few students did mention that they took an English course in high school as well. There were two women and eight men who took the test. This gender breakdown is representative of the student population as a whole.

Instruments:

There were two instruments that were used in this phase of the study. The first was the multi-level video listening test that was developed in Phase Two. The second instrument was a questionnaire that was adapted from Progosh (1996). It contained 10 questions, using a 5-point LIKERT scale, and had space for comments after each question. The students’ questionnaire had two extra questions regarding the situations where they most often needed to use English (See Appendix G). Questions 1-2, 6-10 referred to research objective #1; questions 3, 4, and 5 referred to research objective #2; and questions 11 & 12 referred to the target language use situations the students find themselves in most often.
Procedure:

Procedures to obtain ethical approval to conduct this research were followed. Informed consent forms for the different stakeholders were drawn up that explained the research (please see Appendix H for the test developer’s consent form and Appendix I for the teachers’ consent form). The students’ consent forms (Appendix J), clearly explained what was expected of the volunteers and that they could withdraw from the study at any time. MTCP students were told that their final listening scores would not be affected in any way by participating in this research. The consent forms were given to the participants before beginning the test; no one declined.

The video listening test was administered to all groups, but with two different methods. Due to contextual and logistical constraints, I was not able to bring the FastTEST software to Germany in order to download the test on the computers at the seminar; therefore, the test developers had to take the test under different conditions. That is, the test developers were administered the test as a video plus pen and paper test. The other two groups of stakeholders, the teachers and students, were able to take the test as a computer-delivered test.

Group 1:

Because I was not able to bring the computer software with me to Germany, I created student booklets. Once I explained the research and the constraints put upon us, I asked for volunteers to take the test. Once I obtained their consent, I played the videos on the screen in the front of the classroom. They then answered the items directly in the student booklet. I observed the group and when they were finished answering the item, I would advance to the next video. Once all 24 videos were played, we corrected the test together and then the group completed the questionnaire. We had an informal discussion after all the questionnaires were collected, in which I took notes. This discussion did not last long (approximately 10 minutes) because the test was administered just before the graduation ceremonies for the seminar. This was the only time that was available to administer the test.
Groups 2 & 3

The test was computer-delivered to the teachers and students individually. This was due to technical constraints which imposed restrictions on time, resulting in fewer participants. The final number of participants was 10 teachers and 10 students.

I met with the teachers to explain the research and ask for volunteers. I then took a stratified random sampling of the students, based on the listening scores that they had received at the beginning of the course. I met with these students and explained the research, what the test was and what to expect. I also told them that their official listening score at the end of the course would not be affected by participating in this study. After receiving signed consent forms from the teacher volunteers and the students, I scheduled their tests.

I gave each of the participants an ID number, which allowed them to access the test. I read the instructions to them and ensured their understanding of how the program worked. I went over the example with them. Before leaving the room, I answered any questions they had and reminded them that if they did not want to take the test, they could withdraw from the study. When they finished the test, I gave them their score, which was instantly available from FastTEST Pro. Many of them gave me some comments after the test, but others just left the room without saying anything. These comments were documented and were incorporated with the written comments from the questionnaire. All comments are discussed in the Data Analysis section.

Data Analysis:

I used content analysis to analyze the qualitative data that I collected as a result of the trial. Themes emerged from the data and comments were categorized according to these themes.

Frequency counts were used with the quantitative data that were collected through the questionnaire. The categories were collapsed and are reported as disagreement and agreement. The neutral category is also reported.
Summary

In this chapter I have described the three phases of the study. This mixed-methods research design was adapted from an exploratory sequential model, in that the prototype test was developed on the basis of the qualitative information that was collected from a needs analysis. Then, the prototype was trialed and feedback was collected. This information then contributed to the development of a computer-delivered multi-level video listening test that reflected STANAG Levels 1-3. The test was then trialed in both Germany and in Canada, although in different modalities due to contextual and logistical constraints. Both qualitative and quantitative results were gathered from both locations. These results will be reported in the next chapter.
CHAPTER FOUR
PRESENTATION OF RESULTS: INCLUDING AN AUA EXPLANATION
AND DISCUSSION

Introduction

In this chapter, the overarching results of Phases One and Two will be summarized, in view of the fact that they were presented in the previous chapter. Both quantitative and qualitative data were collected in Phase Three. The quantitative data will be presented first, then a detailed Assessment Use Argument is articulated, and the qualitative data are used as backing for the warrants that elaborate the claim of beneficial consequences of this assessment.

Phase One (a) results:

Results of Needs Analysis

The MTCP teachers were asked to discuss the listening needs of their students in a focus group forum. They mentioned that some students require English for specific purposes (e.g. certain telephone responses, or some specific vocabulary for Air Traffic Controllers). However, they did agree that the majority of situations where the students find they need to communicate in English are when they are engaged in face-to-face conversations. Three distinct themes emerged from the content analysis: non-verbal language; students’ anxiety, and the students’ ability to listen. Table 8 provides the categories and examples of the comments made by the teachers. The comments are direct quotes from the teachers and represent their voices.
### Summary of the focus group meeting

<table>
<thead>
<tr>
<th>Categories</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Non-verbal language</strong></td>
<td>“It’s a given that people who don’t necessarily understand each other, they naturally use gestures to get their message across.”</td>
</tr>
<tr>
<td></td>
<td>“I think we all understand the non-verbal cues are an important part for communication but I don’t find that I really have to draw students’ attention to those things very much, they do it naturally.”</td>
</tr>
<tr>
<td></td>
<td>“Students pick up on it on their own – do not necessarily need to overtly point it out.”</td>
</tr>
<tr>
<td><strong>Students’ feelings of anxiety</strong></td>
<td>“In my opinion, they feel like they are deprived of one aspect of communication by just having the phone ‘cause in the classroom, they are always seeing their colleagues and teacher using a lot of non-verbal communication.” [referred to an English certification interview that was conducted over the phone and not face-to-face]</td>
</tr>
<tr>
<td></td>
<td>“One student began to panic, even though he already had his exam…”</td>
</tr>
<tr>
<td><strong>Students’ overall ability to listen</strong></td>
<td>“There is a lot of blocking at the beginning”</td>
</tr>
<tr>
<td></td>
<td>“I had a student who wanted to speak to somebody and he wasn’t in his office so I said “phone him”. And she wouldn’t do it. She said “You do it. I can’t! I can’t!” yeah but I said you just…just phone and leave your name and number cause …it was administrative stuff that had to be taken care of…and she absolutely refused to do it. It just reminds me now that…cause I remember kind of being taken aback cause she could speak English well enough easily to communicate but…as soon as she thought she may have to start speaking on the phone, she refused to do it.”</td>
</tr>
<tr>
<td><strong>Students’ overall ability to learn</strong></td>
<td>“They are not good listeners overall. When some irony or sarcasm is pointed out to the students, they then see it oh yeah!!!”</td>
</tr>
<tr>
<td></td>
<td>“Requires a huge amount of training – [a student] lost her focus for one moment and she missed a whole chunk.”</td>
</tr>
</tbody>
</table>

**Summary of needs analysis**

After the focus group meeting, I concluded that the MTCP students’ listening needs were general English that is used primarily in face-to-face situations, which closely
resembles the STANAG rating scale. Therefore, I concluded that if I based my test items on the STANAG, I would meet the listening needs of the students.

The meeting with the teachers also confirmed my initial observations that students had difficulty listening in general. The teachers commented that during listening activities, the students had difficulty focussing on the listening passage; once they lost their focus, they would miss a lot of what was being said, which then exacerbated their feelings of anxiety. This information, added to the question of whether using videos in a listening test would be engaging or distracting, led me to the development of the prototype video listening test.

Phase One (b) Results:
Results from Prototype trial

The participants in the prototype trial were teachers and test developers. Their results are reported below. These data were analyzed and can be placed into two broad categories: what they liked about the test and how/where the test can be improved.

The main attribute of the prototype that the participants liked best was that they did not feel stressed. It lowered their anxiety levels.

“\textquote I was encouraged to go through the test because I wasn’t stressed. Had I done this test in an audio-only format, despite the easy item at first, I would have been stressed.\textquote”

“\textquote Scores may not be different from video vs audio-only, but the psychological impact on the student may be great. The students would perceive the video as a fairer test since they were not stressed and more relaxed; therefore, they feel they did better on the test.\textquote”

They also found the item timer to be advantageous and the fact that the listening texts were not too scripted was helpful, which made them sound authentic.

“\textquote Time limit bars were very helpful.\textquote”

“\textquote The dialogues were very natural.\textquote”

The main area that the participants all agreed upon that needed improvement was the fact that there needed to be more time to preview the items before the video started.
“Give longer time to preview the question before the text/video starts to give students a clearer purpose for listening (more time to read the question)”

One disadvantage to this prototype test was that the students did not have any control over the videos. They had to wait for them to start and wait until the item timer finished before going on to the next item. This resulted in time wasted on easier items when more time was needed for the more difficult items.

“Allow the students to go through the test at their own pace. This would allow the stronger students to get through the easy items faster and get to the more difficult ones.”

**Summary of prototype trial**

It is interesting to note that many of the participants in Group 2b, the international test developers, did not provide many comments, except that they liked the test. When asked to give more concrete examples of what they liked or did not like, they just repeated that they liked it and thought using videos would be a good idea. The teachers in Group 2a gave more constructive feedback.

One comment that was made on several occasions was that, when using videos, I had to be careful that the visual element did not solely give away the answer to the item. In other words, the item needed to ensure that the test taker had to listen to the speakers in order to answer it correctly, and that they were not able to identify the correct answer only by looking at the visual context.

The feedback that I received was positive, which encouraged me to continue to the next Phase of the study.

**Phase Two Results**

The results of the four stages of test development produced several documents: a Design Statement, a Blueprint, Item Specifications, an Assessment Use Argument (AUA), and a computer-delivered, multi-level, English general proficiency video listening test. The AUA will be reported together with the results of Phase Three, as the data that were collected in this phase act as backing for several warrants in the AUA.
Phase Three Results

Quantitative

Both quantitative and qualitative data were collected during this phase. First, the results from the questionnaire will be presented. Next, the qualitative data will be presented within the framework of the Assessment Use Argument (AUA).

A 5-point Likert questionnaire was administered to the participants after they had taken the video listening test. For more meaningful results, the Strongly Disagree and Disagree categories were collapsed, as were the Strongly Agree and Agree. They will be reported as either Disagreement or Agreement and questions that were left unanswered were grouped together with the Neutral category. Frequency counts of the total number of responses were calculated and are reported in Table 9 as percentages. Frequency counts were also calculated for each group individually, as seen in Table 10, and are reported as percentages.

Table 9
Combined stakeholders’ responses: Frequency Counts in Percentages (N=31)

<table>
<thead>
<tr>
<th>Questionnaire question</th>
<th>Disagreement</th>
<th>Neutral</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>This was an interesting test taking experience</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>The sound was clear</td>
<td>0%</td>
<td>30%</td>
<td>94%</td>
</tr>
<tr>
<td>I was able to focus my attention on the listening passages</td>
<td>0%</td>
<td>13%</td>
<td>81%</td>
</tr>
<tr>
<td>The videos helped me to understand what was being said</td>
<td>23%</td>
<td>23%</td>
<td>55%</td>
</tr>
<tr>
<td>The videos were distracting</td>
<td>65%</td>
<td>3%</td>
<td>32%</td>
</tr>
<tr>
<td>Using videos is a good way of testing listening comprehension</td>
<td>6%</td>
<td>19%</td>
<td>74%</td>
</tr>
<tr>
<td>This test was easier than an audio only test</td>
<td>13%</td>
<td>29%</td>
<td>58%</td>
</tr>
<tr>
<td>Listening to audio-only passages makes me nervous</td>
<td>45%</td>
<td>22%</td>
<td>32%</td>
</tr>
<tr>
<td>Having videos in the listening test makes me less nervous</td>
<td>13%</td>
<td>29%</td>
<td>61%</td>
</tr>
</tbody>
</table>
**Table 10**  
*Individual Group responses: Frequency Counts in Percentages (N=11, 10, 10 respectively)*

<table>
<thead>
<tr>
<th>Question</th>
<th>Groups</th>
<th>Disagreement</th>
<th>Neutral</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>This was an interesting test taking experience</strong></td>
<td>Test</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>The sound was clear</strong></td>
<td>Test</td>
<td>0%</td>
<td>9%</td>
<td>91%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>I was able to focus my attention on the listening passages</strong></td>
<td>Test</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>The videos helped me to understand what was being said</strong></td>
<td>Test</td>
<td>36%</td>
<td>18%</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>The videos were distracting</strong></td>
<td>Test</td>
<td>64%</td>
<td>0%</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Using videos is a good way of testing listening comprehension</strong></td>
<td>Test</td>
<td>18%</td>
<td>18%</td>
<td>64%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>This test was easier than an audio only test</strong></td>
<td>Test</td>
<td>18%</td>
<td>36%</td>
<td>46%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Listening to audio-only passages makes me nervous</strong></td>
<td>Test</td>
<td>55%</td>
<td>27%</td>
<td>18%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Having videos in the listening test makes me less nervous</strong></td>
<td>Test</td>
<td>18%</td>
<td>46%</td>
<td>36%</td>
</tr>
<tr>
<td></td>
<td>Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Students</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Overall, the results show a positive view of using videos in a general proficiency listening test. One hundred percent of participants in all three groups agreed that this was an interesting test experience. Ninety-four percent of participants reported that the sound was clear, although two participants mentioned that the use of headphones would have been appreciated. Interestingly, 81% of the participants reported that they were able to focus their attention on the listening texts, yet only 55% said that the videos helped them to understand what was being said. However, of that 55%, 90% came from the students. Interestingly, 50% of teachers remained neutral on the idea that the videos helped them understand what was being said.

There was a split among the participants when asked whether or not they found the videos distracting. Eighty percent of students and 50% of teachers reported they did not find them distracting. However, 40% of teachers and 36% of test developers agreed that the videos were, in fact, distracting.

Seventy-four percent of the participants agreed that using videos is a good way of testing listening comprehension. Still, teachers had the greatest reservations, with 40% of them remaining neutral with this statement. Eighteen percent of test developers disagreed that using videos is a good way of testing listening comprehension.

When asked if the video listening test was easier than an audio-only listening test, 80% of students agreed, yet only 30% of students reported being nervous when they have to listen to audio-only passages. Despite this low percentage, 61% of all participants said that having videos in the test made them less nervous.

Ninety percent of students reported that they most often used English in face-to-face situations and not over the phone.

**Summary of quantitative results**

Generally, the participants thought that using videos in a listening test was a good idea and during an informal discussion with the test developers many of them said they would like to try using videos with their students. Many agreed that having the visual aspect of the situation available to students during a listening test would have beneficial consequences.
Qualitative

In this section, I articulate an Assessment Use Argument (AUA) for the video listening test. As explained earlier in this paper, the AUA is a structured approach to collecting evidence that will act as justification for the use of an assessment. The structure of the AUA consists of a series of four claims about (1) the beneficial consequences of an assessment, (2) the decisions that are to be made, (3) the interpretations that are made, and (4) the assessment records. Under each claim, there are a series of warrants, which are statements that elaborate the claims (see Table 11).

Table 11
Example of the structure of an AUA (Bachman & Palmer, 2010, p. 158-159)

<table>
<thead>
<tr>
<th>Claim 1 CONSEQUENCES: The consequences of using an assessment and of the decisions that are made are beneficial to stakeholders.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: Warrants about the <strong>beneficence</strong> of the consequences of <em>using the assessment</em>:</td>
</tr>
<tr>
<td>1. the consequences of using the assessment that are <strong>specific to each stakeholder group</strong> will be beneficial.</td>
</tr>
<tr>
<td>2. Assessment reports of individual test takers are treated confidentially.</td>
</tr>
<tr>
<td>3. Assessment reports are presented in ways that are clear and understandable to all stakeholder groups.</td>
</tr>
<tr>
<td>4. Assessment reports are distributed to stakeholders in a timely manner.</td>
</tr>
<tr>
<td>5. In language instructional settings, the assessment helps promote good instructional settings, the assessment helps promote good instructional practice and effective learning, and the use of the assessment is thus beneficial to students, instructors, supervisors, the program, etc.</td>
</tr>
<tr>
<td>B: Warrant and rebuttal about the <strong>beneficence</strong> of the consequences of the <em>decisions that are made</em>:</td>
</tr>
<tr>
<td>1. Warrant: The consequences of the decisions will be beneficial for each group of stakeholders.</td>
</tr>
<tr>
<td>2. Rebuttal: Either false <strong>positive</strong> classification errors or false <strong>negative</strong> classification errors, or both, will have detrimental consequences for the stakeholders who are affected.</td>
</tr>
</tbody>
</table>

This structure is followed for the other three claims of the AUA. The claim is stated and is followed by the warrants and rebuttals (if any). Backing that supports the warrants and refutes the rebuttals is then presented. The qualitative data that were collected in Phase Three of the present study is reported as backing to the warrants that elaborate Claim 1: Using this assessment will have beneficial consequences for the stakeholders.
ASSESSMENT USE ARGUMENT

SETTING

The students studying English under the Military Training and Cooperation Program at the Canadian Forces Language School have been having difficulty with the current listening test. This test was designed to measure general proficiency in listening comprehension, yet the listening texts are delivered through an audio-only format. To be denied the visual channel during listening is to limit sources of information that can help EFL learners understand the context of the listening text. The reported listening scores at the end of each 19-week course have been significantly lower than other scores, and do not reflect the reality of the students’ listening comprehension (as seen during the Oral Proficiency Interview).

I decided to embark on a project that uses videos as a means of delivering the listening text in order for the MTCP students to be able to utilize the visual aspect of the situation to help their comprehension. Being able to see the speakers and their gestures is more in line with the target language use situations that our students find themselves in when having to use English.
CONSEQUENCES

CLAIM 1  The consequences of using a video listening test and of the decisions that are made are beneficial to the test developers, the MTCP teachers, and the MTCP students.

WARRANTS: CONSEQUENCES OF USING THE MULTI-LEVEL VIDEO LISTENING TEST

A1: The consequences of using the assessment that are specific to the test developers, the MTCP teachers and the MTCP students will be beneficial.

i. Test developers will be able to develop tests that are more authentic and that better reflect the tasks in the TLU domain. The test will then be better accepted as a valid measure of listening comprehension among stakeholders.

ii. Teachers will be able to use more videos in class, which may be more stimulating and interesting for the students. They will be able to teach strategies that encourage the inclusion of non-verbal cues, which may lead to using more authentic material and tasks that are closer to the TLU domain. They will see the test as being more authentic and therefore fairer for the test takers. They will see this as a valid approach to testing listening comprehension.

iii. Students will be able to use non-verbal cues to help them with comprehending the verbal language. The context will be clear for the students, thus reducing their levels of anxiety and allowing them to concentrate on the listening passages. They will see the video listening test as being an authentic test that more closely resembles the TLU domain. They will view the video listening test as a valid approach to testing listening.

Rebuttal:

The consequences of using the assessment that are specific to the test developers, the MTCP teachers and the MTCP students will not be beneficial.
Forty percent of teachers were not convinced that having videos in the listening test would be beneficial to the students. The comments made by the teachers that act as a rebuttal to this warrant are the following:

“For weak ESL students, many of the questions will be just guessing”

“Need to have a good short-term memory and be able to “juggle” multiple cognitive “levels” simultaneously”

“In some ways, having to attend to multiple sources of stimulation (visual + auditory) is more tiring, demanding.”

“I felt I had to consider an additional element…the coffee pot, the newspaper, the staffroom…the concrete doorway, outside weather…What relevance did they have to the linguistic content? Would this help me to understand Chinese any better?”

“More difficult at first. There are 3 skills here – reading and understanding the different answers, listening and watching for background information.”

Forty percent of teachers and 36% of test developers agreed that the videos were distracting. Almost half of the teachers also remained neutral on whether or not they found the videos helpful. The comments that support these percentages follow:

“The novelty of it was a bit…or made it a bit distracting, but as I was under time pressure, I had to keep myself focused and managed to concentrate on the task.”

“To me, it’s a bit difficult to focus my attention on listening only where I have a video as well; it’s due to my character. Sometimes when I watch sth [something] I forget that I’m suppose[d] to listen as well”

“As the items became more difficult I have to give up watching them.”

“At the beginning I felt more distracted. As the test progressed, I tried to concentrate more on what was written and on listening and glanced at the video from time to time.”

“Once I had the answer I would notice other things in the video. For example: clothes, location, people.”

“It was only distracting for the news reports. I had the option of closing my eyes to help me concentrate on all the details.”
Backing:
The students reported overwhelmingly (90%) that the videos were helpful and would
be a great idea for future students.

“it is a best way for the listening test”

“I thought it was good way for future students”

“I hope this test becomes used because it will help other students”

“I think this is very good idea to learn English. Also this video exam better than audio”

“I would like to thank you for giving me an opportunity of passing this test, I am
sure that it will be very helpful for students who are going to take it.”

“Because the video helps me to understand the context or the situation”

“Much better when you are listening & watching (visual) who’s talking”

“It is helpful”

“Yes, because I can see who is/are the person talking and I can see the
object/thing they are talking about. Unlike when I’m just listening, I have to
figure out the object & have to internalize how the object works”

“Listening while watching is easier. I need not to internalize about the subject
matter. I think listening with video will help the students to understand the topic
easier.”

“It gave focus to me, therefore allowing me to listen – often, when listening to
audio-only – my mind wanders, i.e. I think of something else, therefore missing
the listening text.”

“talking on the phone is okay too, but sometimes when I didn’t see a person
talking, it’s hard to decipher or understand what she’s/he’s talking about,
especially when there are things that needs to explain or describe about
something. Seeing the actual object or the subject matter makes it easier to
understand.”

Fifty-eight percent of the student population reported that the video listening test
lowered their anxiety levels. The following are comments made by the students:

“It is helpful”
“Yes, because I can see who is/are the person talking and I can see the object/thing they are talking about. Unlike when I’m just listening, I have to figure out the object & have to internalize how the object works.”

“It felt less like a test”

“They were relaxing; therefore there was no mental block to listening because of nervousness.”

Despite the high percentages of teachers who reported the videos to be distracting, or questioned whether they were helpful, the majority of teachers reported that they thought the videos were a good addition to the listening test and would be a good way of testing listening comprehension (64%). They made the following comments:

“It felt more natural and helped put me at ease. (Perhaps it kept me occupied at a higher level and not overly focussed on the listening)”

“I believe that it lowers the student’s affective filter to a level where they feel comfortable and this would give us a more accurate score.”

“This was fun!”

“I thought the test-maker did a great job at creating realistic conversations. The actors seemed comfortable in their roles and projected a natural speech pattern, such as their intonation, rate of delivery, etc”

“Fascinating – Enjoyable – fun (affective benefit for students)”

Eighty-percent of students and 50% of teachers reported they did not find the videos distracting.

“It’s easier for me seeing things while listening”

“It’s helpful having visual aids while listening”

A2 Assessment reports, which include the (1) scores from the video listening test and (2) the proficiency level decisions made on the basis of them, are treated confidentially.

**Rebuttal:**

No rebuttal

**Backing:**
Follow established procedure at CDA
Test scores are designated Protected “B”, meaning that only authorized personnel are allowed to see the scores. Scores are reported to the senior teachers who, then in turn, inform the students.

A3:  Assessment reports, which include the (1) scores from the video listening test and (2) the proficiency level decisions made on the basis of them, are presented in ways that are clear and understandable to all the test takers.

Rebuttal:  
No rebuttal

Backing:  
Follow established procedure at CDA

A4:  The Test Administration Center at CDA distributes the assessment reports to authorized personnel at the Canadian Defence Academy and Canadian Forces Language School in time for them to be used for the intended decisions. The senior teachers give the reports to the test takers.

Rebuttal:  
No rebuttal

Backing:  
Follow established procedure at CDA. The results of the STANAG tests, from all four skills, are given to the students just prior to their departure from Canada. The decisions that are made based on these scores are done so by the home country after the students’ return.

A5:  The video listening test helps promote good instructional practice and effective learning, and the use of this is thus beneficial to the test developers, MTCP teachers, and MTCP students.
i. **Test developers**: the accuracy of rating the students’ listening comprehension will improve.

ii. **Teachers**: the classroom teaching of instructors will improve. (positive washback effect)

iii. **Students**: their performance on the test will improve; thereby scores will reflect their true listening comprehension ability.

**Rebuttal:**

The video listening test helps promote good instructional practice and effective learning, and the use of this is thus beneficial to the test developers, MTCP teachers, and MTCP students.

**Backing:**

**Test Developers**: The use of videos can be theoretically justified in that it introduces construct-relevant variance if nonverbal information is included in construct definition (Wagner, 2002, 2007).

Similarly, if the test task characteristics are similar to the TLU characteristics, then the test can be seen as having construct validity (Bachman and Palmer, 1996)

**Teachers**: Using videos on the listening test can be pedagogically justified, in that the test will better reflect what is being used in the classroom.

**Students**: Previous research has demonstrated improved performance on video listening tests as opposed to aural-only listening tests (Baltova, 1994; Shi, 1998; Sueyoshi & Hardison, 2005; Wagner, 2010)

**WARRANTS: CONSEQUENCES OF THE DECISIONS THAT ARE MADE**

**B1**: The consequences of the proficiency level decisions that are made will be beneficial for the test developers.
Rebuttal:

The consequences of false positive and false negative classification errors will be different, as follows:

1. **False positive classification errors.** Being rated at a too high a level of proficiency in listening comprehension will have detrimental consequences for students because their supervisors will expect them to be able to understand certain situations that they do not. In battle situations, if a student has been rated at a higher level of proficiency than they are really at, the consequences could be a misunderstanding that could result in injury or even death.

   Students may be promoted to positions for which they are not ready, which can result in higher levels of stress and poor work performance.

2. **False negative classification errors.** Being rated at too low a level of proficiency in listening comprehension will have detrimental consequences because these students may be passed up for promotion. This may lead to feelings of disappointment and less motivation in their job. It may also lead them to develop an erroneous view of their own listening level, as many students underestimate themselves in terms of their listening capabilities.

**Possible ways of mitigating the detrimental consequences of decision classification errors if they occur**

1. False positive classification errors.
2. False negative classification errors.

**Backing**

One way of mitigating the detrimental consequences of both decision classification errors if they occur is to compare the test taker’s listening and speaking scores. If the test taker receives a Level 0 in listening and a level 1 or higher in speaking, then a test
developer, different from the one who conducted the interview, can relisten to the speaking test (which has been recorded) to ensure the correct rating was given. No room for error: MTCP students receive their results and then they get on a plane the next day to return to their respective home countries. For this reason, we need very precise and accurate evaluation tools to minimize these detrimental consequences.

B2: The consequences of the proficiency level decisions that are made will be beneficial for the MTCP teachers.

Rebuttal:
No rebuttal

Backing:
There will be fewer complaints from the students about their final listening scores. There will be fewer anxious feelings to deal with.

B3: The consequences of the proficiency level decisions that are made will be beneficial for the MTCP students.

Backing
The students will see the video listening test as a more authentic way of testing their listening comprehension. The videos will allow them to focus their attention on the speakers, and will make them feel less anxious about the test. Therefore, their scores will better reflect their proficiency levels and they will be able to go back to their countries with a realistic understanding of their listening comprehension in English. If their SLP levels meet the linguistic profiles of specific jobs that they are interested in, their supervisors will have evidence of their true listening ability.

Comments from the stakeholders concerning the authenticity of the listening texts follow:
“\textquote{I think during our life, we won’t use headphones and audio to talk each other. This item is very good to learn English}”
“The video aspect helped to ground the task, making it more authentic than just an audio test”

“The dialogues/monologues were more natural than those on other listening tests I’ve encountered, which I think is wonderful.”

“Although I haven’t taken many audio-only tests, I thought this would help put a student at ease by engaging more senses in the task. This made it more realistic or closer to an authentic interaction.”

“The visual component adds to the comprehension. Great effort was put into authentic settings”

“Depending on the situation, visual clues are more important. In professional settings, it is rare that there is no visual support”

“The videos are more engaging than a purely audio-based listening test. Students today, especially those using computers, which are almost universally used in western education, are used to having visual support in online activities. This mode of delivery is more in line with what they are used to and therefore is likely more comfortable – at least familiar – than a disembodied voice”

“I thought the test-maker did a great job at creating realistic conversations. The actors seemed comfortable in their roles and projected a natural speech pattern, such as their intonation, rate of delivery, etc”
The decisions to award a proficiency level reflect existing educational and societal values (see explanation below) and the content/task/accuracy statements as stated in the NATO STANAG 6001 Language Proficiency Levels and are equitable for those students who are placed at different proficiency levels. These decisions are made by the test developers and refer to which proficiency level the students belong. The individuals affected by these decisions are the students and the teachers of the MTCP program.

The decisions, stakeholders affected by the decisions and the individuals responsible for making the decision are provided in Table 11 below.

Table 12
The decisions, stakeholders affected by decisions, and individuals responsible for making the decisions

<table>
<thead>
<tr>
<th>Decision</th>
<th>Stakeholders who will be affected by the decision</th>
<th>Individual(s) responsible for making the decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Award Level 0/0+</td>
<td>Students and teachers in the MTCP program</td>
<td>Test developer</td>
</tr>
<tr>
<td>Award Level 1/1+</td>
<td>Students and teachers in the MTCP program</td>
<td>Test developer</td>
</tr>
<tr>
<td>Award Level 2/2+</td>
<td>Students and teachers in the MTCP program</td>
<td>Test developer</td>
</tr>
<tr>
<td>Award Level 3</td>
<td>Students and teachers in the MTCP program</td>
<td>Test developer</td>
</tr>
<tr>
<td>Career altering</td>
<td>MTCP students</td>
<td>Home country</td>
</tr>
</tbody>
</table>
WARRANTS: VALUES SENSITIVITY

A1 Relevant educational values of CFLS and CDA are carefully considered in the proficiency level decisions that are made.

Rebuttal:
   No rebuttal

Backing
   At CDA and CFLS, the teaching and testing of languages for the MTCP program are governed by two documents: the Qualification Standard (QS) and the Foreign National Training Plan (FNTP). According to the QS (2006), “principles of the Communicative Approach, adult education and second language acquisition shall be applied.” In the FNTP (2006) it states that “through this [communicative] approach, it is understood that knowledge of the structures and vocabulary of a language does not in itself constitute the ability to communicate in real-life situations. Language is seen, more broadly, as a continuous process of expression, interpretation, and negotiation, which transforms ideas, thoughts, and feelings into speech and writing. Any individual who has attained a measure of competence in this process is said to possess communicative competence.”
   The video listening test follows the task/content/accuracy statements for each proficiency level as defined by the NATO STANAG 6001.

A2 Existing educational values and guidelines of the NATO STANAG 6001 Language Proficiency Levels are carefully considered in determining the relative seriousness of false positive and false negative classification errors.

Rebuttal:
   No rebuttal

Backing
   The test developers refer to the ILTA Code of Ethics guidelines and the Code of Fair Testing Practices in Education, prepared by the Joint Committee on Testing Practices. This document is available through the BILC.
A3 However, no cut-off scores have been set at this point.

(a) *Relative seriousness of classification decision errors:* both types of errors are serious, as they will affect future employment decisions made by the students’ home countries. However, false negative errors may be less serious as the students may have the chance to prove that their listening comprehension is at a higher level than the level awarded to them in Canada.

(b) *Policy-level procedures for setting standards:* Standards are set by the Bureau for International Languages Coordination (BILC) through the NATO STANG 6001 Language Proficiency Levels.

**Rebuttal:**

No rebuttal

**Backing:**

The NATO STANDARDIZATION AGREEMENT 6001 Language Proficiency Levels

---

**WARRANT: EQUITABILITY**

B1. The same cut-off score is used to classify all students taking the English General Proficiency Video Listening Comprehension Test; no other considerations are used.

**Rebuttal:**

No rebuttal

**Backing:**

All the test takers are administered the same test at the same time.

B2. Test takers, teachers and other individuals within CDA and CFLS are fully informed about how the decision will be made and whether decisions are actually made in the way described to them.

**Rebuttal:**

No rebuttal
Backing:  
All the information is contained in the Candidate’s Guide booklet, which is available for all teachers and students.

B3. For proficiency level decisions, test takers have equal opportunity to learn or acquire the ability to be assessed.

Rebuttal:
No rebuttal

Backing:
All test takers have participated in, and completed, the 19-week intensive English course offered through the MTCP program
CLAIM 3  The interpretations about the students’ ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information are **meaningful** in terms of listening to and comprehending general English, **impartial** to all groups of test takers, **generalizable** to tasks that resemble the TLU, and **relevant** to and **sufficient** for the proficiency level decisions that are to be made.

WARRANTS: MEANINGFUL

A1. The interpretations about the students’ “ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information” are **meaningful** in terms of listening to and comprehending general English, **impartial** to all groups of test takers, **generalizable** to tasks that resemble the TLU, and **relevant** to and **sufficient** for the proficiency level decisions that are to be made.

- The definition of the construct is the “ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video”.
- The definition is based on research on listening comprehension and the tasks that are outlined in the STANAG 6001.

Rebuttal:
No rebuttal

Backing:
Wagner (2002) found that the video listening test in his study suggested a two-factor model of listening as the ability to comprehend explicit and implicit information.
NATO STANAG 6001 Trisection of the level descriptors

A2. The assessment task specifications clearly specify that the test takers will watch a video and answer a multiple-choice question that will require them to listen for the main idea, explicit information or implicit information.
Rebuttal:
No rebuttal

Backing:
Follow established procedure at CDA. All items must be written up using an Item Specification form which explicitly states what this item is intended to measure.

A3. The procedures for administering the video listening test enable test takers to perform at their highest level on the “ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video”.

Rebuttal:
No rebuttal

Backing:
The instructions for the video listening test are written on the computer screen and are also read aloud by the proctor. An example is provided in order to show test takers what is expected from them.

A4. The scoring key is included in the computer program that delivers the video listening test; therefore the scoring is done automatically.

Rebuttal:
No rebuttal

Backing:
This is part of the FastTEST Pro computer software capability.

A5. The video listening test engages the “ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video”.
Rebuttal:
No rebuttal

Backing:
Wagner (2002) found that the video listening test in his study suggested a two-factor model of listening as the ability to comprehend explicit and implicit information.

CDA Item Specification form

A6. The scores on the video listening test are interpreted as “ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video”.

Rebuttal:
No rebuttal

Backing:
Based on the trialling of the video listening test, cut off scores will be calculated. Once the items have been validated, then the scores will reflect this construct definition.

A7. The testing section of CDA communicates the definition of the construct in non-technical language via the instructions for the video listening test. The construct definition is also included in the candidate’s guide in non-technical language for the test takers and other stakeholders. The Candidate’s guide is a document that provides the stakeholders with information about the tests in order to help prepare the students.

Rebuttal:
No rebuttal

Backing:
The Candidate’s Guide
Warrants: Impartiality

B1. The video listening test does not include response formats or content that may either favour or disfavour some test takers.

Rebuttal:
No rebuttal

Backing:
The response format requires a test taker to click on a radio button with a mouse.
This is an objective response format.

B2. The video listening test does not include content that may be offensive to some test takers.

Rebuttal:
No rebuttal

Backing:
The content of the items is based on the content statements at the different levels of proficiency according to the NATO STANAG 6001.

B3. The procedures for producing an assessment record for the video listening test are clearly described in terms that are understandable to all test takers.

Rebuttal:
No rebuttal

Backing:
Followed established procedure

B4. Test takers are treated impartially during all aspects of the administration of the assessment.
(a) Test takers have equal access to information about of the assessment content and assessment procedures.

(b) Test takers have equal access to the assessment, in terms of cost, location, and familiarity with conditions and equipment.

(c) Test takers have equal opportunity to demonstrate their knowledge of utilizing verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video.

**Rebuttal:**

No rebuttal

**Backing:**

All the information is in the Candidate’s Guide, which is available to all teachers and students.

All test sessions are organized by the testing section, and tests are administered in the same computer lab for all students.

All students will take the same test, thereby having equal opportunity to demonstrate their knowledge.

---

B5. Interpretations of the test takers’ “ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video” are equally meaningful across students from different first language backgrounds and academic disciplines.

**Rebuttal:**

No rebuttal

**Backing:**

The test is criterion-referenced, according to the NATO STANAG 6001

All students are given a student ID to log onto the video listening test. Therefore, no names are used, nor are countries identified on the test.
Warrants: Generalizability

C1. The characteristics of the tasks in the video listening test correspond closely to those tasks outlined in the STANAG 6001 at different levels of proficiency.

Rebuttal:
No rebuttal

Backing:
The assessment tasks were created according to the task and content statements in the NATO STANG 6001.

C2. The criteria and procedures for evaluating the responses to the tasks in the video listening test correspond closely to those that are typically used in the TLU.

Rebuttal:
No rebuttal

Backing:
Do not need this warrant as all items are multiple-choice.

Warrant: Relevance

D. The interpretation of the “ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video” provides the information that is relevant to the test developer’s decisions about proficiency levels.

Rebuttal:
No rebuttal

Backing:
Research shows that non-verbal movements, which include gestures and visuals, are a natural and important part of listening comprehension (Kellerman, 1990, 1992, Okey, 2007, Hostetter, 2011)
**WARRANT: SUFFICIENCY**

E. The assessment-based interpretation of the “ability to utilize verbal and non-verbal behaviour to comprehend the main idea, explicitly stated information and implicit information from a text delivered through video” provides sufficient information to make the proficiency level decisions.

**Rebuttal:**

No rebuttal

**Backing**

The interpretations will be based on listening texts that include all aspects of the listening situation – visuals and auditory.
ASSESSMENT RECORDS

CLAIM 4 The scores from the video listening test are consistent across different forms and administrations of the test, across students from different military trades, and across groups with different nationalities and first languages.

WARRANTS: CONSISTENCY
1. The video listening test is administered in a standard way every time it is offered.
   Rebuttal: No rebuttal

2. The scoring criteria and procedures for the computer scoring algorithm are well specified and are adhered to.
   Rebuttal: No rebuttal

3. Raters undergo training and must be certified
   Not needed, as this is a multiple-choice test
   Rebuttal: No rebuttal

4. The cut score was developed through trialling with several different groups of test takers
   Rebuttal: No rebuttal

5. Scores on different tasks in the video listening test are internally consistent.
   Rebuttal: No rebuttal

6. Ratings of different raters are consistent
   Not needed, as this is a multiple-choice test
Rebuttal:

No rebuttal

7. Different ratings by the same rater are consistent
   Not needed, as this is a multiple-choice test
Rebuttal:

   No rebuttal

8. Scores from different forms of the video listening test are consistent.
Rebuttal:

   No rebuttal

9. Scores from different administrations of the video listening test are consistent.
Rebuttal:

   No rebuttal

10. Scores on the video listening test are consistent across different groups.
Rebuttal:

   No rebuttal

Backing

   Evidence needs to be gathered
Summary

In this chapter, I have reported and discussed the results from each of the three phases this research went through. The results of the initial needs analysis confirmed the observations that had taken place over many years – that the MTCP students have difficulty with the listening skill, and that they are very anxious when it comes to performing listening tasks – whether it is in a test situation or not. A prototype video listening test was developed and informally trialed with colleagues both at CFLS and at the LTS in Germany. The results of the prototype trial allowed me to progress through four out of five stages in test development as outlined by Bachman and Palmer (2010). Results of the trialing of the video listening test have been reported in this chapter. Documents were produced after each stage and all culminated in an AUA. The AUA has been articulated, with the qualitative data that was collected from Phase Three of this research acting as backing to the warrants that elaborate Claim 1: the use of the video listening test will have beneficial consequences for the stakeholders. Backing that has been provided for the other warrants that elaborate the other claims in the AUA come from the context of the MTCP program that is given by CDA and CFLS.

In the next chapter, I will discuss the results with respect to the research question and the research objectives.
CHAPTER FIVE
FINAL DISCUSSION: THE RESEARCH QUESTION
AND OBJECTIVES

Introduction

In this chapter I will discuss the findings from this research study in terms of how they related to the two research objectives. I will then show how, in relating to the research objectives, the findings have addressed the main research question.

Research Objectives Revisited

Research Objective #1 To what extent will different stakeholders (test developers, teachers, students) perceive the use of videos as the medium of delivering listening texts as being beneficial when testing listening comprehension?

Overall, the perception of the stakeholders was positive towards using videos in a listening comprehension test.

The majority of test developers reported that they believed it would be a good idea to include videos in a listening comprehension test. Many of them also reported that they did not find the videos distracting and that they helped them understand the spoken passages. However, a high percentage of test developers (36%) reported a certain amount of reservation. This reservation can be accounted for by the actual test method that was used. Due to technical difficulties, the test developers were not able to take the test on the computer. Instead, they saw the videos on a screen in the front of a classroom and they answered the items in a student booklet. Several of them reported that the videos were distracting, especially as the passages got more difficult. The test developers found watching the videos on the screen, then looking down to answer the questions, and then looking up to watch the next video was very distracting, which made many resort to not watching the videos at all. These comments support several researchers’ concerns that test takers would be so busy looking at their test papers and answering the questions, that they would not even bother watching the videos (Alderson, Clapham, and Wall, 1995; Brett, 1997, and Gruba, 1994). This finding also supports Wagner (2010a) where he found the test takers watched the video less than half the time when they had to watch the
videos on a screen in front of the class, and answer the items in a test booklet. Brett (1997) found that multi-media delivered listening comprehension tasks may be more efficient than the traditional audio-only or video plus pen and paper.

One of the test developers suggested putting the items on the computer and having them and the video side by side on the same screen. This is exactly what was administered to the teachers and the students. Due to this difference in test method, it is not surprising that some of the test developers had these reservations about including videos, yet only two test developers actually disagreed with the question “using videos is a good way of testing listening comprehension” that appeared on the questionnaire. Unfortunately, no one wrote any qualitative comments regarding this question. Despite the concerns, the test developers did express that they thought this was an interesting idea and said they would be willing to try it with their own students. They could see the benefits of using a video, but the test method interfered with their view. However, they did say that it was useful to be able to see the context and have the speakers clearly distinguished – especially when there were only men or only women speaking. It would be interesting to see whether their impressions and perceptions would change if they were able to see the final product as it is.

The majority of teachers agreed that it would be good to have videos included in the listening test, but there were a number of reservations. Some teachers were not convinced that the video added any value to the assessment of a student’s listening ability. They reported that they found they were looking for clues in the video, or were distracted by something in the video that would make them then miss what was being said. A few teachers said that they found the context did not necessarily match the content, which made them question the usefulness of video, which supports Cross’s (2010) findings in that if there is no correspondence in meaning audio and visual content, there will be a problem with facilitating comprehension. Yet despite this criticism, the majority of teachers did not report this incongruence of context and content, and were excited by this method of testing listening comprehension and could see the benefits for the students. Instead, several teachers commented on how well the visuals provided the context, which allowed them to concentrate on the listening passage. Some teachers also commented that they thought the conversations and all the situations were realistic and authentic,
which made the videos engaging and allowed them to focus their attention on the
speakers. The concerns expressed by the teachers must be taken into account when a
video is being used for a listening comprehension test. The test developer must ensure a
high congruence of audio and visual content in order to maximise the positive influence
on comprehension for the students (Cross, 2010). The test developer must also keep in
mind that perhaps the aim of the video is merely to get the attention of the test taker
(Kelly & Goldsmith, 2004) in order for him/her to focus on what the speakers are saying.

The students unanimously agreed that the inclusion of videos in a listening test would
be beneficial to their performance. They reported being less nervous and more engaged
in the listening passages, which enabled them to listen more closely. Many students
reported that they felt this kind of listening test was a great improvement to the traditional
listening test experience

Most of them reported that the video test was easier than an audio-only test, but when
examining the comments made by the students, the test was only easier because they did
not have to imagine the context. Perhaps having to listen to audio-only passages and
having to imagine not only the context but the speakers as well, and to listen to what they
are saying puts more of a strain on the cognitive load than having the context and
suggests that humans can process dual sources of information concurrently if the two
sources are in different modalities (i.e. visual and oral)”, and are complementary sources
of information, as are the verbal and non-verbal components of spoken language
(Anderson, 2004). Consequently, if students do not have to imagine these factors, then
perhaps they are able to more easily listen to what is actually being said and therefore
allow them to truly perform at their best. This supports Hostetter’s (2011) findings that
state that visuals provide additional cues when comprehension is difficult (especially for
L2 learners).

The fact that none of the students expressed similar concerns as the teachers is very
telling. The students were very enthusiastic with the inclusion of videos and several of
them reported that this method of testing listening is better than the traditional method
and were quite convinced that it will help future students. Despite the concerns expressed
by the test developers and teachers, the majority of these participant groups and all the
students liked the idea of using videos in the listening comprehension test, which supports several studies (Baltova, 1994; Dunkel, 1991; Progosh, 1996; Seuyoshi and Hardison, 2005; and Wagner, 2002) that found video was preferable to audio-only.

There were some contradictions in the results. One student who reported that the videos were distracting also strongly agreed that using videos is a good way of testing listening comprehension. This student may not have had a firm understanding of what “distracting” meant, even though I had explained all the vocabulary that appeared on the questionnaire.

Another student disagreed that the videos helped him understand what was being said and agreed that the videos were a distraction, yet he still agreed that using videos is a good way of testing listening comprehension. He also agreed that the videos become more helpful as the items become more difficult. An explanation for this anomaly is that perhaps because the student was not very proficient in English, he did not really understand all the questions.

All groups saw the test as being more authentic than a traditional audio-only listening test. The idea of the authenticity of the items – a concept that Bachman and Palmer (1996) discussed thoroughly – emerged from the data. According to Bachman and Palmer (1996), we, as test developers, must consider the target language use (TLU) situation and try to have our assessment tasks resemble real life tasks as much as possible. When we do so, the inferences made about the students’ performance can be seen as having high construct and content validity. Ninety percent of the students said that they use English in face-to-face situations more often than they use the language over the phone. Therefore, if visual support is present in the TLU, then it follows that our tests should include visual support in order to be representative of the TLU and allow us to confidently make inferences about the students’ performance outside the test situation.

Research Objective #2: To what extent will students report feeling less anxious when taking a video listening test?

In 1996, Progosh reported that one’s affective filter goes down when one feels less nervous, which in turn, allows the students to process more information. Although only 30% of the students reported that they feel nervous when listening to audio-only passages,
this is a difficult characteristic to identify. However, despite what they reported, 80% reported that having the videos would make them less nervous. Some participants even mentioned that taking the test was fun and that it did not even seem like a test. One participant continued watching the videos after she had chosen her answer, just because she was enjoying them. Comments such as these support the suggestion made by numerous researchers that if the test is fun and enjoyable, students will want to do it, which could, again, reduce their anxiety levels, thereby allowing the test takers to really listen to what is being said (Croft et al, 2001, Sambel et al, 1999, Matsumura & Hann, 2004).

This discussion of the research objectives leads to a discussion of the question that guided this research.

**Research Question Revisited**

**RQ: To what extent is the AUA framework suitable for justifying using videos as a means of delivering a listening text in a multi-level video listening test?**

The data collected in this study provides backing for the claim that using the video listening test will have beneficial consequences for the stakeholders. I believe that the perceptions from the test developers, the MTCP teachers and the MTCP students, show that there are indeed beneficial consequences. The students are not highly stressed, which will have an effect on the teachers and the classroom environment as well. Also, the students see the test as being more authentic and probably fairer as an assessment of their listening comprehension. Ultimately, though, the students feel that the videos helped them in their understanding of the texts. All these consequences of using the video listening test are documented in the AUA and can be referred to at any time by any stakeholder. The AUA also allows the test developer to address any rebuttals that may arise. For example, one rebuttal from the teachers was that they felt the videos were distracting. Yet, not one of the students reported that. On the contrary, the majority of them said that they were helpful.

One of the biggest advantages of using the AUA is that it can help the test developers with the kind of information they need to collect in order to justify the assessment that was either developed or to justify the use of an existing assessment. In addition to this,
the AUA allows the test developer to address any rebuttals that are made and present evidence against them in a clear and concise manner.

The framework of the AUA is extremely useful in allowing the test developer to keep their evidence together in one document. As Bachman and Palmer (2010) state, “to be competent in language assessment, means being able to demonstrate to stakeholders that the intended uses of their assessments are justified.” The AUA provides a clear framework in which the test developer can clearly justify the development/selection and the uses of the assessment. The AUA framework also allows the test developers to collect evidence of construct validity for their test. This is important if the test developers are held accountable for the uses of their tests and the decisions the developers made are easily justified within this framework.

The present study found, too, that the AUA is a pertinent and relevant framework in language testing. It guides the test developer in test construction, while at the same time, it provides evidence for the construct validity of that test.

In this chapter, I have provided a discussion of the results with respect to the research question and the objectives. In the next chapter, I will provide a summary of the findings, the implications and limitations of the study as well as suggest future research. I also mention the contribution that my study will make to language testing.
CHAPTER SIX
CONCLUSION

Introduction

In this chapter, I will summarize the findings from this research study. I will then discuss the implications and some of the limitations. I make some recommendations for future research and close the chapter by summarizing my contribution to the field of language testing.

Summary of findings

An Assessment Use Argument framework was the foundation of this study. I articulated a complete AUA and collected evidence in order to back the claim stating that this test will have beneficial consequences for the test taker. In this study, a computer-delivered listening test that uses videos as the medium of delivery for the listening passages was developed. After conducting a needs analysis and developing a prototype test, I was able to synthesize the information gathered from these two tools in order to develop a 24-item multi-level video listening test. I went through four out of the five stages of test development, as described by Bachman and Palmer (1996). The fifth stage of test development is the official implementation of the test, which is out of the scope of this thesis, as more research is needed to support the inclusion of visuals in a listening test. Nevertheless, I was able to gather evidence of beneficial consequences from three different stakeholders: the test developers, the MTCP teachers, and the MTCP students. The results showed a positive view of using videos from all three different groups of stakeholders. Many of the test developers thought that the use of videos would benefit the students in that the students would feel that the test better reflected the TLU situations in which they find themselves.

Many of the teachers were more reserved with their support of using the videos. They were concerned with the videos being distracting and with their general usefulness. However, most of the teachers and all of the students overwhelmingly approved of the videos. The students felt that the consequences of using this test would be beneficial to their performance. They felt that their anxiety would be reduced, they would not have to
rely on their imaginations for the contexts of the listening passages and they felt that the listening texts reflected more authentic situations than a traditional audio-only listening test.

A complete AUA was articulated, and many of the documents and procedures that are used at the Canadian Defence Academy, coupled with research, provided backing for the many warrants that support the claims made: that the use of the test will have beneficial consequences for the test taker, that decisions are equitable, that interpretations of the students’ performance are meaningful, impartial, generalizable, relevant and sufficient. They also provide backing for the assessment records that are kept. In having articulated the AUA, the use of videos in a listening test can be justified in that this study has shown that the inclusion of videos will have beneficial consequences for the test taker. It has also strengthened the construct validity of a video listening test.

Implications

This study raises questions about the construct validity of the traditional audio-only listening tests that are currently being used at CDA and elsewhere. If the majority of situations in which our MTCP students find they need to use English is when they see the other person, then our tests need to reflect that reality. The beneficial consequences of using videos in a listening comprehension test seem to outweigh those of a traditional audio-only test. Our militaries often work on the world stage and have to communicate with people from other countries in face-to-face situations. Due to the high-stakes nature of the STANAG tests, it is imperative that they reflect the TLU in order that the proficiency levels that are awarded to the students genuinely reflect what they can and cannot do in the language. The consequences of a wrong level could be the difference between life and death.

The AUA provides a sound framework in which the validity of a test and its use can be justified and the test developers can be accountable for their test. If other nations adopt this framework and can provide evidence that their tests support the claims stated in the AUA, then perhaps the mission of the BILC (Bureau for International Languages Coordination) to ensure that all nations have a common interpretation of the STANAG 6001 will be facilitated.
Limitations

There are some limitations to this study. Due to technical difficulties with the software and due to time constraints, only a small sample of stakeholders was able to participate. Although generalizations cannot be made on their performance, there is an indication that perhaps this method of testing listening comprehension with our foreign students may be an interesting alternative and may be able to address the problems that we encounter with testing this skill. This is backed by the AUA framework and results.

Another limitation to this study was that the test itself could not provide generalizations on performance, given that only 24 non-validated items were used. This is not enough of a sample of the TLU at the different levels that would allow for reliable interpretations. Time constraints prevented the inclusion of a validation period for the items. Some military contexts could have been used, which would have made the listening passages that much more authentic.

Another limitation is the fact that the videos could only be listened to once. This was done in order to reflect the current listening test used with the foreign national students, where the listening passage is only played once. In the future, the test taker should have the opportunity to listen a second time if necessary.

A further limitation is that the production of a video listening test requires more resources than a traditional audio-only one. It takes longer to film and edit videos than it does to make an audio recording of a passage.

Future Research

This study has provided an example of a detailed AUA that can be used to justify the inclusion of videos on a listening comprehension test. It can be used as a departure point for future studies that can address the limitations mentioned above. More research is needed to continue providing evidence to justify the use of videos in listening comprehension tests; research in areas such as the effect of videos on performance on a multi-level listening test and the usefulness of videos on item difficulty and on students’ level of proficiency. More research is needed on validating a video listening test. An interesting research study would be the usefulness of videos in a listening test for visual learners as opposed to those who are not, and the extent to which the inclusion of videos
in an L2 listening comprehension test would have beneficial consequences on these learners.

**Contribution**

This study has demonstrated the development of a high-stakes instrument in a mixed methods framework, using the AUA as a further justification for construct validity. I have used the AUA as a sound theoretical structure that has shown that the inclusion of videos in a listening comprehension test will have beneficial consequences for the students, and this can be justified. Nowhere in the literature have I found a complete AUA articulated with respect to listening comprehension. This AUA can be complementary to those studies that have looked at using videos in assessing listening, such as the research conducted by Wagner (2002, 2007, 2008, 2010).

The study will also contribute to the literature on testing listening comprehension and perhaps influence future test development projects.
REFERENCES


www.elsevier.com/locate/COGNIT


Qualification Standard, Language Training Course for Foreign Nationals, Military Training Assistance Programme (MTAP). Issued on authority of the Chief of Defence Staff, Managing Authority: Canadian Defence Academy, 27 March 2006.

Rost, M (2002). Teaching and researching listening. Harlow, UK: Pearson Education/Longman.


Sotaro Kita (2009): Cross-cultural variation of speech-accompanying gesture: A review. *Language and Cognitive Processes, 24*(2), 145-167. To link to this article: [http://dx.doi.org/10.1080/01690960802586188](http://dx.doi.org/10.1080/01690960802586188)


TOEFL: Test of English as a Foreign Language: [http://www.ets.org/toefl](http://www.ets.org/toefl)


Training Plan, Language Programme for Foreign Nationals. Issued on authority of the Chief of Defence Staff, Managing Authority: Canadian Defence Academy, 19 June 2006.


Appendix A

STANAG 6001 Level Descriptions for Listening Comprehension

LEVEL 0 (NO PROFICIENCY)

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

LEVEL 0+ (MEMORIZED PROFICIENCY)

Understands isolated words and some high frequency phrases and short sentences in areas of immediate survival needs. Usually requires pauses even between familiar phrases and must often request repetition. Can understand only with difficulty even people used to adapting their speech when speaking with non-natives. Can best understand those utterances in which context strongly supports meaning.

LEVEL 1 (SURVIVAL)

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.

LEVEL 2 (FUNCTIONAL)

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.

LEVEL 3 (PROFESSIONAL)

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.

LEVEL 4 (EXPERT)

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 counseling. 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.

LEVEL 5 (NATIVE/BILINGUAL)

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.
FOCUS GROUP QUESTIONS

Thank you for coming to this focus group meeting. I have asked you here because you have worked closely with the MTCP students and are in a better position than I to understand their listening needs.

I would like to discuss the following questions:

1. What kind of listening do the MTCP students need for their job(s)? in other words, in their jobs, what are some of the tasks that require listening in English?

2. How often do they find themselves in these situations (or doing such tasks?)

3. How anxious do they report themselves as being when having to listen in English? To what extent do they report anxiety when having to listen?

4. How do you, as teachers, help prepare your students for their listening test? What kinds of activities do you engage the student in, in order to practice their listening skills?

5. Do you, as teachers, draw students’ attention to the non verbal behaviour of a speaker? Do you view this as an important part of some listening tasks?

Thank you for your participation

Nancy Powers
Appendix C

Faculty of Education
Integrated Studies in Education
3700 McTavish Street
Montreal, Quebec
Canada H3A 1Y2
Tel: 398-4527  Fax: 398-4529

Project: Master’s Thesis exploring the use of videos when testing listening comprehension
Principal Investigator: Nancy Powers
Program: Second Language Education
Supervisor: Dr. Carolyn Turner (514) 398-6984
Date: July 7, 2010

Purpose of the research:
This research is an exploratory study that will investigate the impressions from different stakeholders on the use of videos as a means of delivering listening comprehension texts. By participating in a focus group discussion, I will ask you questions pertaining to the listening needs of the MTCP clientele. This information will be used as a basis for the construct definition of listening as the video listening test is developed.

What is involved in participating:
You will be asked several questions in order to identify different situations in which the MTCP students need to use their listening skills in English. Your participation is voluntary and you may choose not to participate or withdraw at any time or decline to answer any question you don’t want to. Your name will never be revealed in written or oral presentations and no record will be kept of your name. The focus group discussion will last ½ hour (30 minutes) and it will be audio-taped. The information gained from this focus group will be used solely by the researcher in order to help ensure that the video listening test relates to the needs of the students. Some comments may be reported in the final thesis report, although identities will remain anonymous.

By participating in this research you will be able to contribute to the future research into the evolution of testing listening comprehension.

If you have any questions concerning this research, or would like to give some additional information, you may contact me by phone at 7495, by email at nancy.powers@forces.gc.ca or come by my office in C-214.
I have read and understood all of the above conditions. I freely consent and voluntarily agree to participate in this focus group.  ____YES  ____NO

I agree to be audio-taped.  ____YES  ____NO

I agree to have my comments reported in the final thesis report, with the understanding that my name will not be revealed.  ____YES  ____NO

Participant’s printed name  __________________________________________

Participant’s signature  __________________________________________

Date:  _________________________________________

Researcher’s signature  _________________________________________
## TRISECTION: LISTENING COMPREHENSION

### LISTENING LEVEL 1

| CONTENT: | Familiar phrases and short simple sentences  
Everyday needs such as minimum courtesy, travel, and workplace requirements  
Concrete utterances, simple questions and answers, and very simple conversations  
Topics such as meals, lodging, transportation, time, simple directions and instructions |
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TASKS:</td>
<td>Understand the main idea</td>
</tr>
</tbody>
</table>
| ACCURACY: | 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 speech from media or among native speakers if content is completely unambiguous and predictable |

### LISTENING LEVEL 2

| CONTENT: | Everyday social and job-related conversation  
Concrete topics, such as personal and family news, public matters of personal and general interest, routine work matters  
Descriptions of persons, places, and things  
Narration of current, past, and future events |
| --- | --- |
| TASKS: | Understand factual, paragraph level discourse  
Answer factual questions about texts |
| ACCURACY: | 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 only comprehend general meaning of speech from the media or among native speakers using specialized or sophisticated language  
Unable to understand subtleties of language surrounding the facts |
LISTENING LEVEL 3

CONTENT: Most formal and informal speech on practical, social, and professional topics
Speech on professional specialty
Language used at interactive meetings, briefings, and other extended discourse
Abstract concepts on such topics as economics, culture, science

TASKS: Understand hypothesis, supported opinion, argumentation, statements and defense of policy, other forms of elaboration
Understand both explicit and implicit information
Distinguish between various stylistic levels
Recognize humor, irony, emotional overtones, subtleties

ACCURACY: Can follow accurately the essentials of conversation among educated native speakers, lectures on general subjects, reasonably clear telephone calls, and media broadcasts
Rarely has to request repetition, paraphrase, or explanation
However, may not understand very rapid native speech, slang, regionalisms, or dialect

LISTENING LEVEL 4

CONTENT: All forms and styles of speech used for professional purposes
Highly sophisticated language including most matters of interest to well-educated native speakers
Language used in representation of official policies, lectures, negotiations
Language tailored for various audiences, including persuasion, representation, and counseling
Highly abstract concepts

TASKS: Adjust to shifts of subject matter and tone
Follow unpredictable turns of thought in both formal and informal speech on any subject matter addressed to the general listener
Recognize nuances of meaning and stylistic levels, irony, humour

ACCURACY: Readily understand language in media and in conversations among native speakers, both globally and in detail
Generally comprehend regionalisms and dialects
Appendix E

VIDEO LISTENING TEST BLUEPRINT

The test’s purpose
To test general proficiency in listening comprehension as defined by NATO STANAG 6001 Language Proficiency Levels

Description of the test taker
Members of foreign military in Canada studying English for 4 ½ months under the Military Training Cooperation Programme (MTCP)
Majority are males between the ages of 25-50
Varying levels of English proficiency
Varying levels of computer familiarity

Test level
Designed to test Levels 1, 2, & 3 as defined by the NATO STANAG 6001 Language Proficiency Levels

Construct (theoretical framework for test)
Using non-verbal cues can
- Listen for explicit information
- Listen for implicit information
- Listen for the main idea

Number of sections/papers
1 section – 25 items
10 items at level 1
10 items at level 2
5 items at level 3

Time for test
30 minutes: the FastTEST Pro software that will be used to deliver this test allows for a display of a countdown session timer to the examinee and allows the examinee to hide or unhide the timer at will. Note that FastTEST Pro has 3 timers – a session timer, a test timer and an item timer. Only one should be displayed at one time.

Text Features:
The video text will be listened to only once. Therefore a test timer will be used.
**Weighting for each section/paper**
Each item is dichotomously scored.

**Target language situation**
As defined by the NATO STANAG 6001 Level Descriptors

**Text-types**
Dialogues
Monologues

**Text length**
Level 1 items: range 0:14 to 1:04, average length 32.5 secs.
Level 2 items: range 0:40 to 1:50, average length 73.6 secs.
Level 3 items: range 1:17 to 2:17, average length 1:52 min.

**Language elements to be tested**
Ability to use the non-verbal cues present to help listen for:
  - explicit information
  - explicit information
  - the main idea

**Test tasks**
Listening to short videos and answering multiple-choice questions based on the video texts

**Test methods**
This test will be computer-delivered multiple-choice test. The software being used, FastTEST Pro 2.0, locks examinees out of Windows once a test has begun. The examinees can only respond to the session screens that are presented.

The examinees will indicate their response by clicking on the radio button next to the response they have chosen. The examinees will click on the link provided to begin the videos. In this way, they will have the time they need to preview question and responses before viewing the video. In this way, they will have a clear purpose to what they are listening for.

**Interface Design**
The computer screen will be a vertically split screen. This will allow the examinee to read the item, which will be on the right-hand side, and watch the video, which will be on the left-hand side, on the same screen.
The font will be Arial 12, because I once read that this was the easiest font for the eyes.

**Rubrics**
NATO STANAG 6001 Language Proficiency Levels

**Descriptions of typical performance at each level**
As described in the NATO STANAG 6001 Language Proficiency Levels

**Descriptions of what candidates at each level can do in the real world**
As described in the NATO STANAG 6001 Language Proficiency Levels
Appendix F

VLT TEST ITEM SPECIFICATIONS

<table>
<thead>
<tr>
<th>Test Code:</th>
<th>Item Number:</th>
<th>Intended Level:</th>
<th>Validated Level:</th>
</tr>
</thead>
<tbody>
<tr>
<td>VLT</td>
<td>Item 1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

Author: NP  
Date Created:  
Length of passage: 0:19

TLU (As Per STANAG Level Descriptor 6001 for MTCP)  
Minimum courtesy, can understand spoken language among native speakers if content is completely unambiguous and predictable

Task = listening for the main idea

Check or underline the appropriate category below.

<table>
<thead>
<tr>
<th>Item Type:</th>
<th>Direct Question</th>
<th>Incomplete Sentence</th>
<th>Monologue</th>
<th>Dialogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Context:</td>
<td>Military</td>
<td>Civilian</td>
<td>Formal</td>
<td>Informal</td>
</tr>
<tr>
<td>Target:</td>
<td>Main Idea</td>
<td>Explicit Information</td>
<td>Implicit Information</td>
<td></td>
</tr>
</tbody>
</table>

Patti is ____________.

a. giving some advice  
b. making an introduction  
c. requesting a service  
d. ordering a subordinate
Appendix G

QUESTIONNAIRE for students

Faculty of Education
Integrated Studies in Education
3700 McTavish Street
Montreal, Quebec
Canada H3A 1Y2
Tel: 398-4527 Fax: 398-4529

STAKEHOLDERS’ PERCEPTIONS ON USING VIDEOS AS A MEDIUM OF DELIVERING LISTENING COMPREHENSION TEXTS IN A MILITARY HIGH STAKES TESTING CONTEXT

Country ________________________________

Age:
- under 25 years _____ 26-35 years _____
- 36-45 years _____ over 46 years _____

Rank (if applicable) _____________________________

Familiarity of computers: little _____ some _____ a lot _____

Teacher _____ Student _____ Test Developer _____

Number of years studying English:
- 0-5 years _____ 6-10 years _____
- 11-15 years _____ over 16 years _____

Nancy Powers
MA DISE SLE
McGill University
Supervisor: Dr. Carolyn Turner
<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. This was an interesting test taking experience.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>2. The sound was clear.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. This test was easier than an audio-only test.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Listening to audio-only passages makes me nervous.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Having videos in the listening test made me less nervous.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. I was able to focus my attention on the listening passages.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________

____________________________________________________________________
<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>7. The videos helped me to understand what was being said</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:

____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
____________________________________________________________________

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. The videos were distracting.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:

____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
____________________________________________________________________
<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>9. The videos become more helpful as the items become more difficult.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>10. Using videos is a good way of testing listening comprehension.</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>11. I usually use English in face-to-face situations (when I see the other person)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

<table>
<thead>
<tr>
<th>Question</th>
<th>Strongly Disagree</th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
<th>Strongly Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>12. I usually use English on the phone (when I do not see the other person)</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Additional comments:
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
Any further comments
Appendix H: Test Developer’s Consent Form

Test Developer’s Consent to Participate in Research Study

Project Title: Stakeholders’ perceptions on using videos as a medium of delivering listening comprehension texts in a military high stakes testing context
Principal Investigator: Nancy Powers
University: McGill
Faculty: Department of Integrated Studies in Education (DISE); 3700 McTavish Street, Montreal, Quebec, Canada, H3A 1Y2
Supervisor: Dr. Carolyn Turner (514) 398-6984

Purpose and Procedures:

The purpose of this research is to develop a general proficiency listening test using videos to deliver the listening test. Your participation in this study will entail taking a listening comprehension test that has either videos or audio-only, and it will last approximately 30 minutes. A questionnaire will follow the test in order to get your thoughts and feelings about the test. Participants’ personal information will not be divulged to anyone and anonymity will be maintained in all written and published data resulting from this study.

Conditions of Participants:

Your participation is strictly on a voluntary basis and you may choose not to participate or withdraw at any time or refuse to answer any question you do not want to. Under no circumstances will any of your personal information be disclosed and anonymity will be maintained in all written and published data resulting from this study. All participants will receive a randomly selected identification number to assure anonymity. There are no risks involved in participating in this study. Please note, that by participating, you will be making a contribution to the future research into the evolution of testing listening comprehension.

All data collected will be kept in our Protected B computer, as well as a locked filing cabinet. Only I, the researcher, will have access to the data, which will be destroyed once the final thesis has been officially submitted.
You may contact me by email at Nancy.Powers@mail.mcgill.ca or at Nancy.Powers@forces.gc.ca

If you have any questions or concerns regarding your rights or welfare as a participant in this research study please contact the McGill Research Ethics Officer at 514-398-6831 or lynda.mcneil@mcgill.ca.

Nancy Powers
MA Second Language Education
DISE

I have read and understood all of the above conditions. I freely consent and voluntarily agree to participate in this study.

YES ____  NO ____

Participant’s printed name

__________________________________________

Participant’s signature:

__________________________________________

Date:

__________________________________________

Researcher’s signature

_________ Nancy Powers _________
Appendix I: MTCP teachers’ Consent Form

McGill
Faculty of Education
Integrated Studies in Education
3700 McTavish Street
Montreal, Quebec
Canada H3A 1Y2
Tel: 398-4527 Fax: 398-4529

Teacher’s Consent to Participate in Research Study

Project Title: Stakeholders’ perceptions on using videos as a medium of delivering listening comprehension texts in a military high stakes testing context
Principal Investigator: Nancy Powers
University: McGill
Faculty: Department of Integrated Studies in Education (DISE); 3700 McTavish Street, Montreal, Quebec, Canada, H3A 1Y2
Supervisor: Dr. Carolyn Turner (514) 398-6984

Purpose and Procedures:

The purpose of this research is to gather the opinions from different stakeholders on a general proficiency listening test that uses videos to deliver the listening text. Your participation in this study will entail taking a video listening comprehension test and it will last approximately 30 minutes. A questionnaire will follow the test in order to get your thoughts and feelings about the test. Participants’ personal information will not be divulged to anyone and anonymity will be maintained in all written and published data resulting from this study.

Conditions of Participants:

Your participation is strictly on a voluntary basis and you may choose not to participate or withdraw at any time or refuse to answer any question you do not want to. Under no circumstances will any of your personal information be disclosed and anonymity will be maintained in all written and published data resulting from this study. All participants will receive a randomly selected identification number to assure anonymity. There are no risks involved in participating in this study. Please note, that by participating, you will be making a contribution to the future research into the evolution of testing listening comprehension.

All data collected will be kept in our Protected B computer, as well as a locked filing cabinet. Only I, the researcher, will have access to the data, which will be destroyed once the final thesis has been officially submitted.

You may contact me by phone at 7495, email: nancy.powers@mail.mcgill.ca or come by my office in C-214.
If you have any question or concerns regarding your rights or welfare as a participant in this research study please contact the McGill Research Ethics Officer at 514-398-6831 or lynda.mcneil@mcgill.ca.

Nancy Powers  
MA Second Language Education  
DISE  

I have read and understood all of the above conditions. I freely consent and voluntarily agree to participate in this study.  

YES ____  NO ____

Participant’s printed name  

Participant’s signature:

Date:

Researcher’s signature: Nancy Powers
Appendix J: MTCP student’s Consent Form

Faculty of Education
Integrated Studies in Education
3700 McTavish Street
Montreal, Quebec
Canada H3A 1Y2
Tel: 398-4527 Fax: 398-4529

Student’s Consent to Participate in Research Study

Project Title: Stakeholders’ perceptions on using videos as a medium of delivering listening comprehension texts in a military high stakes testing context
Principal Investigator: Nancy Powers
University: McGill
Faculty: Department of Integrated Studies in Education (DISE); 3700 McTavish Street, Montreal, Quebec, Canada, H3A 1Y2
Supervisor: Dr. Carolyn Turner (514) 398-6984

Purpose and Procedures:

The purpose of this research is to develop a general proficiency listening test using videos to deliver the listening text. Your participation in this study will entail taking a listening comprehension test that has either videos or audio-only, and it will last approximately 30 minutes. A questionnaire will follow the test in order to get your thoughts and feelings about the test. Participating in this study will in no way affect your official profile obtained at the end of the course. Participants’ personal information will not be divulged to anyone and anonymity will be maintained in all written and published data resulting from this study. The score you obtain from the official listening test will be used in order to make comparisons.

Conditions of Participants:

Your participation is strictly on a voluntary basis and you may choose not to participate or withdraw at any time or refuse to answer any question you do not want to. Under no circumstances will any of your personal information be disclosed and anonymity will be maintained in all written and published data resulting from this study. All participants will receive a randomly selected identification number to assure anonymity. There are no risks involved in participating in this study. You should not feel any pressure to participate in this study. Your official scores at the end of the course will not be affected in any way by participating or by choosing not to participate in this research study. Please note, that by participating, you will be making a contribution to the future research into the evolution of testing listening comprehension.
You may contact me by phone at 7495, email: nancy.powers@mail.mcgill.ca or come by my office in C-214.

I have read and understood all of the above conditions. I freely consent and voluntarily agree to participate in this study.

YES ____  NO ____

I agree that my official listening score can be used for comparison with the score on the research study listening test.

YES ____  NO ____

Participant’s printed name
__________________________________________

Participant’s signature:
__________________________________________

Date:
__________________________________________

Researcher’s signature  _______ Nancy Powers _______