THE USEFULNESS AND ETHICAL STATUS OF LANGUAGE ENTRANCE/PLACEMENT TESTING AT A POST-SECONDARY INSTITUTION

by

Katherine Jane Golder

B.A., Queen's University, 1995

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

in

The Faculty of Graduate Studies (Language and Literacy Education)

UNIVERSITY OF BRITISH COLUMBIA

April 2006

© Katherine Jane Golder 2006
Abstract

Often language entrance and placement testing at post-secondary institutions is used as a gatekeeping mechanism—sorting people in and out, keeping out those who don’t exactly fit their desired profile. This is troubling in a country as multicultural as Canada since testing language can be seen as testing culture. Testing, therefore, may be perceived as a tool for restricting minority access. This has been highlighted by recent media coverage of foreign-trained professionals and their restricted access to post-secondary institutions to gain Canadian credentials.

However, testing can and should do more than gatekeeping: among other purposes, testing can be an educational, diagnostic tool, and feedback on results can give test takers insight into their strengths and weaknesses. While the classic criteria of reliability and validity are important in evaluating testing, we must also include criteria beyond these. For testing to be considered ‘good’ it must be useful (e.g. demonstrate reliability, validity, practicality) and ethical (ensure all test takers are treated equally, demonstrate respect for persons, and maximize the benefits to stakeholders).

This research project is an exploratory qualitative examination of the usefulness and ethical status of language entrance/placement testing at a Canadian post-secondary institution. Opinions and perceptions of testing were obtained from a variety of stakeholders in the testing process to determine how ethical and useful testing was and was perceived to be. The stakeholder groups involved were test takers (both successful and unsuccessful), test score users (including instructors and administrators), and test developers and administrators. Data was obtained through focus groups, interviews, and questionnaires between September and November of 2005.

The main finding of this research was that while the testing process met certain criteria for ethics and usefulness, there were a variety of areas that could be improved. Overall, a lack of communication between test developers/administrators and other stakeholder groups, as well as a general lack of transparency in the testing process, led to widespread misunderstandings of the tests’ content and purposes. This, in turn, created a lack of respect for the authority of testing, and a lack of face validity.
# Table of Contents

Abstract ......................................................................................................................... ii
Table of Contents ........................................................................................................... iii
List of Figures ................................................................................................................ v
Acknowledgements ....................................................................................................... vi

1  Introduction .............................................................................................................. 1
   1.1 Background to the Research Problem .............................................................. 1
   1.2 Conceptual Framework ................................................................................... 4
   1.3 Purpose of the Study and Research Questions ............................................... 4
   1.4 Research Questions ....................................................................................... 5
   1.5 Background: The Institution ......................................................................... 6
   1.6 The Institution’s Uses of Language Testing .................................................... 7
   1.7 Potential Contribution to Research Knowledge and/or Educational Practice .... 9
   1.8 Organization of Study ................................................................................... 9

2  Current thinking in language assessment ................................................................ 11
   2.1 Test Development ......................................................................................... 11
   2.2 Climate of Change ....................................................................................... 12
   2.3 The Ethical Test ............................................................................................ 13
   2.4 Computer-Based Testing .............................................................................. 16
   2.5 Potential of Dynamic Assessment (DA) ......................................................... 18
   2.6 Testing of Foreign-Trained Professionals in Canada ....................................... 19
   2.7 Glossary ........................................................................................................ 20

3  Research Method ..................................................................................................... 24
   3.1 General Approach ........................................................................................ 24
   3.2 Participants .................................................................................................... 26
   3.3 Procedures ..................................................................................................... 26
   3.4 Data Collection .............................................................................................. 27
   3.5 Protection of Human Subjects ....................................................................... 28
   3.6 Summary of Research Methods .................................................................... 29

4  Findings .................................................................................................................... 30
   4.1 Data Analysis ............................................................................................... 30
   4.2 Results .......................................................................................................... 31
      4.2.1 Ethical Considerations ............................................................................ 31
      4.2.2 Psychometric Considerations ................................................................. 33
      4.2.3 Influences on and of the Test ................................................................. 36
      4.2.4 Institutional Considerations ................................................................. 42
   4.3 Summary of Results ..................................................................................... 45

5  Discussion and Conclusions .................................................................................... 46
   5.1 Answers to Research Questions ..................................................................... 46
5.1.1 How is the language ability construct defined? In other words, are the conditions for construct validity adequately met? ................................. 46
5.1.2 Are the tests currently being used seen to be free from bias that might unfairly disadvantage certain people? ......................................................... 46
5.1.3 Are the skills tested relevant to the skills required to succeed at the institution? .... 47
5.1.4 Are the scores obtained adequate reflections of the construct being tested, and, therefore, useful in determining whether a candidate should be able to enrol at the institution? ........................................................................ 47
5.1.5 Are the scores useful beyond simply determining if a candidate is ‘in’ or ‘out’? For example, can test takers use the feedback they receive to target areas of language they need to improve? .......................................................... 47
5.1.6 Can the institution use the scores to more accurately place students in programs where they will receive the support they need? .......................................................... 48
5.2 Lessons Learned and Implications for Further Research .................................. 48
5.2.1 Lessons for the Researcher ................................................................. 48
5.2.2 Limitations of the Research Design .................................................. 49
5.2.3 Lessons Regarding Testing in General .............................................. 50
5.2.4 Lessons for the Institution ............................................................... 50
5.3 Recommendations .................................................................................. 53
5.3.1 Re-examine Construct Validity ......................................................... 54
5.3.2 Consider the Changing Demographics in the Design Statement .......... 55
5.3.3 Educate Stakeholders About the Test ............................................... 56
5.3.4 Provide Opportunities for Test Takers to Practice Test-Taking Skills .... 58
5.3.5 Make the Purpose of the Test More Open and Transparent ............... 58
5.3.6 Involve Stakeholders in the Testing Process .................................... 59
5.3.7 Make the Criteria for Success on the Test Available to Test Takers .... 60
5.3.8 Establish/Adopt a Code of Ethics (and Perhaps a Code of Practice) for Testing/Evaluation ................................................................. 61
5.3.9 Provide More Extensive Feedback to Test Takers .............................. 61
5.3.10 Make the testing conditions more conducive to optimum performance .. 62
5.4 Summary ............................................................................................... 62

References ...................................................................................................... 64

Appendix A ..................................................................................................... 67

Focus group questions – test takers............................................................ 67
Focus group questions – program heads, deans, and instructors ............... 67
Focus group questions – test developers.................................................... 68
List of Figures

Figure 1.1 Testing process at the institution ........................................................................... 7
Acknowledgements

I am grateful for the help and support of the faculty and staff at UBC, as well as that of my fellow classmates.

In particular, my advisor, Dr. Ken Reeder, has been immeasurably helpful. Aside from providing insights and asking the right questions, his optimism and energy have helped me to maintain my motivation. Working with Ken has helped to make my time at UBC not only rewarding academically, but also enjoyable and memorable.

Another faculty member who played an important role in my studies is Dr. Monique Bournot-Trites. Taking her excellent course on Second Language Assessment was integral to my completion of this research project. Had the course not been so thorough, I would not have been able to ask the questions I did. Monique’s support, including providing suggestions on my thesis despite her being in France as an invited professor, was very valuable.

Dr. Elaine Decker has been very helpful and inspirational from the beginning of this project. She has helped to keep me motivated and focused. I am very grateful for her support and advice. I appreciate her taking the time out of her very busy schedule to be a member of my examining committee.

Two classmates who have been especially helpful are Reg D’Silva and Sue Parker Munn. Discussing the research with them helped me to make important decisions. Knowing that their advice and support were readily available over email helped make the writing process less solitary. I am indebted to them for this.

I am also grateful to my sister-in-law, Nicola Hamer, for taking the time to read and offer valuable editorial advice on a draft of this thesis.

Finally, my husband, Alex Hamer, deserves recognition for the invaluable part he played in my completion of my MA studies. Without the countless practical and intangible ways that he supported me in this, this research would not have been completed.
1 Introduction

This research project investigates how well language entrance/placement testing at a post secondary institution meets the needs of the stakeholders in the testing process by examining how useful and ethical testing is. The usefulness of testing is measured by assessing the balance achieved among reliability (similar results at different times, in different environments, on different forms of the test), construct validity (clear and appropriate construct/tasks, controlled sources of bias), authenticity (similarity to 'real life' language tasks), interactiveness (test takers’ topical knowledge, language knowledge, personal characteristics, language functions, metacognitive strategies, affective schemata), impact (on test takers, teachers, society, and education systems), and practicality (ability to put the test into practice). The three principles of ethics as outlined by Hamp-Lyons (1997) are equal treatment, respect for persons, benefit maximization.

This issue was investigated through questionnaires, focus groups, and interviews with the stakeholders. These stakeholders included successful and unsuccessful test takers, test score users (primarily instructors), and test developers/administrators. Participants were asked about their experiences and perceptions of the institution’s in-house language entrance/placement testing. One particular area of interest in conducting this research project was how testing affects foreign-trained professionals.

1.1 Background to the Research Problem

There has been a great deal of public attention to the issue of recent immigrants, specifically non-native English speakers (NNS), with specialized skills being unemployed or underemployed in Canada (Harding, 2003), or having difficulty accessing the necessary resources to enter the job market (Azuh, 2000). Often they need Canadian credentials to get work. Institutions such as the location of my research, a Canadian post-secondary institution, are instrumental in re-credentialing immigrants in this situation. Both the institutions, and the recent immigrants themselves, could benefit from a critical examination of placement testing practices.

Language testing generally falls into one of two categories: institutional, for making program-level decisions about a student, and pedagogical, for making classroom-level decisions about a student. Brown (1996) describes four primary language testing functions: proficiency
and placement at the program-level, and achievement and diagnostic at the classroom-level. It is important to be clear on which of these specific purposes one has for undertaking testing. Without a clear purpose, it is unlikely that a test will exhibit construct validity. If a test actually assesses what it claims to assess, it can be said to demonstrate construct validity. The language ability being tested must be relevant to the purpose of the test in order to make the desired interpretations about the candidate’s language ability.

The testing I investigated at a large Canadian post-secondary institution was developed by the institution to test language proficiency. Brown (1996) describes language proficiency tests as generally being used to make decisions about entry into a program. In this case, based on the candidate’s score on what will be called Test B, he/she is admitted into a mainstream communication class, advised to take language preparation courses (LPCs), or not admitted to the institution at all. If they are advised to take an LPC and wish to do so, they must take another test developed by the institution, which I will call Test A, to be placed in an appropriate level of LPC.

However, even though Brown (1996) classifies proficiency and diagnostic testing separately (proficiency being norm-referenced and diagnostic being criterion-referenced), for a test score to be relevant and useful to a test taker or test score user, it must provide meaningful feedback (Bachman, 1996). In this sense, the test must have some diagnostic properties – it must be able to tell test takers and score users where areas of strength and weakness are.

Ideally, we should be able to use placement tests for their diagnostic potential (Shohamy, 2001). However, testing is often used to create or enforce policy, act as a gatekeeper and, from a Critical perspective, to perpetuate existing power structures. Policy is created de facto by test developers who may not be aware of the policy they are creating or that testing has a reach far beyond the obvious, direct results that they intend (Shohamy, 2001). A test that acts as a gatekeeper and nothing more represents a policy that is unconcerned with any stakeholders aside from the testing institution. Notions of testing that accept this kind of policy uncritically are becoming increasingly anachronistic.

Therefore, instead of simply turning students away, we should be able to offer support to help them attain the necessary level of English to participate in courses and, therefore, in the case of foreign-trained professionals, help them get back into their profession or a similar line of work. This does not necessarily mean the institution must provide these courses, but it should
at least be able to direct them to the appropriate courses. Improved testing practices would not only benefit recent immigrants who need to be ‘re-credentialed’ but all incoming NNS students. The benefits for institutions and society as a whole are extensive.

Entrance/placement testing is crucial in helping recent immigrants get back into the workforce in their desired (or reasonably close to their desired) profession (Cumming, 1994). Our testing practices must be closely examined to see how well they are serving these individuals and, by extension, the institution, the communities, and society as a whole. There is a great deal of research examining the ethics of testing in general (Shohamy, 2001; McNamara, 2001; Lynch, 2001), but less on how highly-skilled recent immigrants are affected.

Another aspect of testing that should be considered is using computers to administer entrance and placement tests. Many institutions, including the institution studied, are moving this way. Computer-based testing (CBT) presents many challenges, but also a great deal of potential beyond simply transferring pencil and paper (P&P) tests to computers for easier scoring. For example, Poehner and Lantolf’s (2003) discussion of ‘dynamic testing’, discussed in more detail in chapter 2, suggests that dynamic assessment could be well supported by computer technology.

Traditionally, most people, educators included, tend to think of teaching and testing as separate entities: once you are finished teaching, you administer a test. Poehner and Lantolf (2003) emphasize that this is a false dichotomy; teaching and testing (or assessment in general) can and should occur at the same time. This is grounded in the Vygotskian notion of the Zone of Proximal Development: While a student is working on a ‘test’ item, they are given ‘help’, some form of intervention, to complete the task. What is measured is not simply right and wrong, but the amount of intervention it takes for the student to be able to complete the task. While this would be highly labour intensive for human testers, often requiring one-on-one testing, a computer would be able to provide support and measure the support given. This is only one of many potential ‘alternative’ uses for CBT in language assessment.

A similar example can be found in teaching reading to elementary students, where the concept of ‘authentic assessment’ integrates teaching and testing in order to produce a richer understanding of a student’s skills and areas of weakness (Vacca, Vacca and Begoray, 2002). It may not be possible, for many reasons, for high-stakes tests to be truly authentic and provide the same type of rich descriptions as can be found in classroom assessment. However, there is the
potential that (relatively) high-stakes, standardized tests, such as the institution's placements tests, could provide richer descriptions of the test takers' language abilities than they currently do. While the scope of this research project does not permit a deep exploration of either dynamic or authentic assessment, it is nonetheless interesting to see where the future of testing is headed and to consider the impact that this may have on language entrance/placement testing.

Placement and diagnostic testing are crucial in helping recent immigrants get back into the workforce in their desired (or reasonably close to their desired) profession (Cumming, 1994). Our testing practices must be closely examined to see how well they are serving these individuals.

1.2 Conceptual Framework

The conceptual framework for this research consists of two main 'bins,' in Miles and Huberman's (1994) terms: ethics and usefulness. These, in turn, can be subdivided into the qualities that must be present, to a greater or lesser degree, for a test to be considered ethical and useful (these qualities are outlined later). In addition, the focus of the research is bounded by certain parameters: the location (the institution), the testing process (language entrance/placement testing), and the specific group of test takers (non-native speakers, including foreign-trained professionals). This conceptual framework is also grounded in a 'weak' version of the critical research paradigm. In other words, while I am influenced by the critical approach, I am not implementing it in a 'pure' form; it colours my view of the issue and how I will interpret the data. However, I don't wish to be bound to it because, as I will note later, it is not likely possible, nor would it necessarily be valuable, for testing to entirely meet the demands of a critical approach, and I don't wish to be limited in my interpretations.

1.3 Purpose of the Study and Research Questions

The purpose of this study is to look at the case of entrance language testing at this particular institution to determine how well it meets the needs of the stakeholders in the testing process. In order to meet these needs, it is important that the requirements of ethics are met: the test must treat all candidates equally, show respect for persons, and maximize benefits to all parties (Hamp-Lyons, 1997). In addition, the tests must demonstrate a high degree of usefulness as outlined by Bachman & Palmer (1996): usefulness requires achieving the best possible balance of reliability, construct validity, authenticity, interactiveness, impact, and practicality.
It is important to approach this research from a qualitative perspective. In order to really hear the voices of the stakeholders, they must be given an opportunity to be heard in their own words. A quantitative approach, studying only reliability and validity, for example, would run the risk of replicating existing power structures inherent in testing. While reliability and validity are still fundamental to testing, they are no longer considered the sole indicators of a ‘good’ test. Messick’s (1989) work on validity was ground-breaking and did a great deal to expand the notion of validity to include far more than just construct validity. For Messick (1989), validity had to take a variety of factors, including the testing environment and test takers themselves, into account. This expanded notion of validity is reflected in Bachman and Palmer’s (1996) concept of usefulness. Usefulness includes reliability and validity, but it goes well beyond the traditional definitions of these, and includes a variety of qualities that must be balanced for a test to be considered ‘good’.

I used focus groups, interviews and questionnaires to find out how stakeholders in the testing process at the institution perceive the current entrance testing practices. I got opinions on the value, ethics, and usefulness, as defined by Bachman and Palmer (1996), of the current entrance and placement tests from test-takers (successful and unsuccessful), test developers/administrators, instructors, and administrators.

1.4 Research Questions

The overarching question explored in this research project is how useful and ethical is language entrance/placement testing at the institution. Some sub-questions I explored through my interactions with the stakeholders include the following:

1. How is the language ability construct defined by the test developers/administrators? In other words, are the conditions for construct validity adequately met?

2. Are the tests currently being used seen to be free from bias that might unfairly disadvantage certain people?

3. Are the skills tested relevant to the skills required to succeed at the institution?

4. Are the scores obtained adequate reflections of the construct being tested, and, therefore, useful in determining whether a candidate should be able to enrol at the institution?
5. Are the scores useful beyond simply determining if a candidate is ‘in’ or ‘out’? For example, can test takers use the feedback they receive to target areas of language they need to improve?

6. Can the institution use the scores to more accurately place students in programs where they will receive the support they need?

However, some other questions underlie these questions, and the next potential phase of research could at some point in the future go further to explore them more explicitly: How can we answer these questions for non-native speakers of English (NNSs), and even more specifically, for highly-skilled recent immigrants? How can technology help us create a useful and ethical test? What would the technology ideally be capable of? While I must limit the questions I am asking in this phase, the questions for the next potential phase, a future research project, strongly influence how I will frame the this phase.

1.5 Background: The Institution

The site of this research project is a very large polytechnic institution with several campuses in the city and surrounding area. The research took place at the largest campus since it houses the greatest number of schools, and it serves the greatest variety of students.

The institution offers “programs that lead to certificates, diplomas and degrees in technologies, business, health sciences and trades... [it conducts] applied research, technology transfer activities (the taking of ideas to the marketplace), and corporate and industry training and upgrading.” (institution’s website, 2006). Courses are available through full- and part-time studies as well as by distance education over the internet. No specific statistics on the number of NNS students or the number of foreign-trained professionals at the institution were available. However, a recent report on communication skills among the institution’s graduates cites the increasing number of immigrants to the area (42% were considered to speak English, and 44% had a Bachelor’s, Master’s or Doctoral degree) as one of the main factors that made it important to carry out research on communication skills (Hamilton, 2005). Hamilton (2005) suggests that is a “factor that needs to be considered when planning future communication offerings is the increasing proportion of students whose first language is something other than English” (p. 13). Indeed, it is also a factor that needs to be considered when engaging in language testing.

For details on the study’s participants, see below (chapter three).
1.6 The Institution’s Uses of Language Testing

The testing I investigated included two tests I called Test A and Test B. Test A is used to place a student in an appropriate language preparation course (LPC). Test B is used to determine if a candidate has a sufficient level of English to participate in mainstream courses at the institution. The following flow chart shows the language entrance and placement testing process at the institution:

Figure 1.1 Testing process at the institution

According to the institution’s website, Test B is an English competency assessment that provides “programs with information about an individual’s skill in using the English Language. Assessments are based on the student's understanding of grammar, reading comprehension, and clarity and structure in written composition.” To take this test costs $95. Most people who take this test are asked to do so because they wish to gain admission to mainstream programs at the
institution but cannot provide evidence of success in grade 12 English. Test takers who are not successful on Test B are recommended to take an LPC. If the test takers decide to take an LPC, they will then have to take Test A ($45), described on the institution’s website as a “written assessment to determine students' eligibility for registration” in an appropriate LPC level.

A test taker may begin the process by taking Test A; however, most people prefer to enter a mainstream program as soon as possible, so they may try to avoid taking LPCs, and start with Test B. Test B is also used as an exit test for students completing the LPCs. Additionally, it has been used to measure the language abilities of people already admitted to the institution. For example, one department recently had the test administered to all its students as diagnostic. As can be seen in the flow chart above, Test B is far more widely used than Test A. While Test A is referred to, the majority of discussions with stakeholders focussed on Test B.

Test A has two components: a multiple-choice grammar section and a writing section. Test B has three sections: a multiple-choice grammar section, a multiple-choice reading comprehension section, and a writing section.

The communication courses that successful test takers will enter, including the LPCs, are focused on business and technical writing. Students in these courses write workplace documents (letters, memos, reports, etc) exclusively and give oral presentations on work-related topics. This strong emphasis on writing for the workplace is what most faculty see as setting the institution apart from other institutions and making it unique. Therefore, the writing assignments on both Tests A and B tend to be more business-like than the typical essay-style writing test. Having marked Test A on several occasions, I am quite familiar with it and can attest to this. However, despite the test administrators’ positive responses to my requests to see a copy of Test B, I was unable to obtain or even see a copy of it, and had to rely on test administrators’ and test takers’ descriptions of it.

Since I began this research, there has been a move to implement a new entrance/placement test: Accuplacer (http://cpts.accuplacer.com). This is still in the piloting stages, and my participants, other than the test developers/administrators, were not aware that Accuplacer may be used in the future. Therefore, unless specifically noted, any mention of the institution’s tests refers to Tests A and B.

Accuplacer is an off-the-shelf computer-based test (CBT) from the College Entrance Examination Board (CEEB) that the institution pays for per use. The skills tested remain the
same as with Tests B: reading, grammar, and writing. The reading and grammar sections on the Accuplacer test are computer adaptive. With computer-adaptive tests (CATs), the test taker is given progressively more difficult questions based on his/her performance on the previous question; the test taker's level is determined when he/she is no longer able to answer the questions correctly. With Accuplacer, the test taker's writing is graded by the computer program and given a band score. Potentially, Accuplacer will be used to administer a listening test in the future.

While the results of my research can no longer be used to make changes to the existing tests, they help highlight the positive aspects of the current tests that should be maintained with Accuplacer. In addition, any areas of weakness or concern that came up can be addressed and improved upon by using Accuplacer. Indeed, much of the discussion among the test developers was about how Accuplacer would be able to do some things better than the current tests.

1.7 Potential Contribution to Research Knowledge and/or Educational Practice

As Bachman and Palmer (1996) point out, it is essential that those impacted by a test are involved in the development of the test. The information gained through these discussions with stakeholders can be used to ensure that tests at the institution have a positive impact on stakeholders.

The impact of tests goes beyond the direct stakeholders and influences educational institutions and society as a whole (Bachman & Palmer, 1996). Beyond this research project but based on its findings, I hope to examine how ethical testing can ensure that foreign-trained professionals are not unnecessarily hindered from re-entering the workforce. This has great potential to benefit both individual immigrants and Canadian society as a whole.

1.8 Organization of Study

Having described the scope and purpose of the study as well as the nature of the research problem in the present chapter, the remainder of this thesis is organized as follows: Chapter two will outline some of the relevant literature on testing and situate this study in the context of the literature. This includes literature on changing thinking in language testing, ethics and testing, computer-based testing, dynamic assessment, and the impact of language testing on foreign-trained professionals. A description of the research approach and methods used in this project
will be provided in chapter three. Chapter four will cover the findings of the study, and in chapter five the findings will be discussed and recommendations made.
2 Current thinking in language assessment

As Bachman and Palmer (1996) note, there are many misconceptions about language testing among stakeholders in the testing process. For example, stakeholders may have unreasonable expectations of tests, or they may trust psychometric measurements blindly. The aim of the following review of some of the literature on language testing aims to provide context for this research project, to dispel misconceptions, and to provide a brief discussion of the complex issues that should be considered in language testing.

2.1 Test Development

Bachman and Palmer (1996) outline the process of test development and point out that it "is not strictly sequential in its implementation" (p. 86). While in general the development moves from one stage to the next, "the process is also an iterative one, in which the decisions that are made and the activities completed at one stage may lead us to reconsider and revise decisions, and repeat activities, that have been done at another stage" (p. 86). In this way, test development is never complete; test developers should always be seeking feedback on the usefulness of testing in order to ensure testing is as useful as possible. Indeed, according to Bachman and Palmer (1996) "all decisions and activities involved in test development are made in order to maximize the overall usefulness of the test" (p. 86).

The first stage of the development, according to Bachman and Palmer (1996) is the design stage in which components of the test are identified, described, and defined. This is an important stage that must be returned to in order to ensure that "performance on test tasks will correspond as closely as possible to language use" (p. 86). This includes, for example, describing the purpose of the test, defining the construct, developing a plan to evaluate usefulness, and describing characteristics of test takers.

In the second stage, operationalization, begins with developing test tasks and a blueprint for how the tasks will be organized on the test. Then the instructions for the tasks are written and a scoring method is specified. Specifying a scoring method requires defining the criteria for evaluation and determining procedures to arrive at a score. At this point, test developers may begin building up a pool, or archive, of test tasks to be used on future versions of the test (Bachman & Palmer, 1996).
The third stage, test administration, is comprised of two phases. The first, try-out, involves administering the test to collect information that will help administrators assess the test’s usefulness and improve the test and testing procedures. The second phase is operational test use. In this phase, the test is used to “make decisions or inferences for which the test was intended” (Bachman & Palmer, 1996, p. 91) and to continue to collect information for assessing usefulness. This stage requires that procedures for administering the test and collecting feedback be established. Of course, there must also be procedures for analyzing the feedback and the test scores (e.g. item analysis).

2.2 Climate of Change

The prevailing trend in literature on language testing suggests that we are going through a “profound change” (McNamara, 2001, p. 333), a change that has caused a certain amount of tension in the field (Hamp-Lyons, p. 330). The shift from a positivist perspective to postmodernism pervades the literature; language and testing are no longer seen to be ‘truths’ in and of themselves, but socially-constructed realities. This change is in evidence in the literature on basic principles in testing, ethics, and CBT. Another area that influences, and is influenced by, these changes is the language assessment of recent immigrants (Cumming, 1994). While this area of testing itself is not new, the “major influx of skilled immigrants in recent years” is (Statistics Canada, 2004a) is a new reality in the Canadian context.

McNamara (2001) provides an excellent overview of the shifts that have taken place in attitudes towards testing practices and our views of language learners. Specifically, he cites a growing recognition of the social aspects of testing, and more broadly, the social nature of language. The very concepts of validity and language proficiency central to language testing have “undergone profound change” (p. 333). McNamara uses a poststructuralist lens to examine testing; it is socially constructed and based on issues of power and control. Testing is inherently political and value-laden. For McNamara, a more ethical test would blur the lines between pedagogy and assessment.

Messick’s (1989) work on validity, as McNamara (2001) points out, is ‘radical’ in its recognition of assessment as social in nature. Traditionally, the positivist view of assessment holds that there is an objective reality to a test (Hamp-Lyons, 1997). Messick’s view posits that “we have no ‘objective’, ‘scientific’, value-free basis for this activity” (McNamara, 2001, p.
This updates the concept of validity to be more in line with current, post-modern, relativistic philosophies of education and assessment.

Bachman and Palmer’s (1996) matrix for evaluating test usefulness, heavily influenced by Messick’s notions of validity, provides testers with a practical, relatively easily-applicable method of determining how ‘useful’ a test is. Reliability, construct validity, authenticity, interactiveness, impact, and practicality are taken into account. Testers can evaluate a test based on the balance that exists between these factors – factors that take into account not only the tester’s and institution’s needs, but the needs and characteristics of the test takers. Because this balance takes a variety of factors beyond reliability and construct validity into account, it can potentially lead to more ethical testing practices.

The development and use of testing should be an iterative process, as Bachman and Palmer (1996) point out. Therefore, it is important that testing practices are continually examined. This research project investigates the testing process at this institution in light of the changes in the notion of validity, and the expansion of the notion of usefulness to include the factors that Bachman and Palmer outline.

2.3 The Ethical Test

Partly as a result of and partly in tandem with this shift in the understanding of basic testing principles, the topic of ethics in testing has arisen. Hamp-Lyons’ (1997) notes that ‘twenty or ten and perhaps even five years ago it [ethics] would not have appeared’ in an encyclopedia of testing such as the one her overview of ethics in testing appears in (p. 328). Hamp-Lyons outlines three principles of ethics: equal treatment, respect for persons, benefit maximization.

Underlying these principles is the necessity for participants in the testing process to be held accountable for their actions: those who develop and administer tests must take responsibility for the intended and unintended consequences of testing. The International Language Testing Association (ILTA) goes so far as to recommend that those involved in the testing process have a moral obligation to refuse to participate in unethical practices (in Alderson and Banerjee, 2001a). This recommendation, while noble, seems rather impractical: although the ILTA notes that those who refuse to participate should not be discriminated against by their employers, it seems unlikely that employees will actually be protected.
The sentiment, however, is well taken. Hamp-Lyons suggests that the prevention of bad testing is as important as the development of new tests (p. 323). To safeguard stakeholders from the misuse and abuse of tests, Hamp-Lyons suggests that ‘those who have had professional preparation in language testing are equipped to appreciate the ethical dilemmas faced by test developers, administrators and score users, and to make ethical decisions themselves as they take on those roles: those who do not, are not” (p329). This admonition is particularly striking: ethical issues facing those involved in any relatively high-stakes testing are too great to leave testing to those who are unprepared to deal with these issues.

The shift towards considering ethics in testing can be seen in Messick’s (1989) seminal work on validity mentioned in section 2.2. In addition to the concept of construct validity, Messick introduced the notion of consequential validity. This means the consequences of a test, what Bachman and Palmer (1997) call ‘impact’, must be considered. Test developers and administrators are called on to be mindful of more than just the traditional psychometric criteria. They must consider the intended and unintended consequences the test will have on those directly and indirectly involved in testing. In addition to the notion of the impact of the test, Bachman and Palmer’s (1997) concept of usefulness includes ethical considerations in the interactiveness of the test; test developers and administrators must consider the test takers’ personal characteristics.

Echoing Bachman and Palmer’s definition of test usefulness and Messick’s concepts of consequential validity, Shohamy (2001) calls for democratic principles to be applied to testing. Tests are sources and symbols of power. Many test takers are so accustomed (socialized) to believing that tests are beyond reproach, that they are unlikely to question the results or use of a test. This maintains the social order. She calls for greater participation in testing among local bodies rather than the far removed testing ‘elites’; power should be shared. Above all, test takers have rights, and those rights should be respected by including test takers’ perspectives in the testing process. Like the ILTA, Shohamy believes tester developers have a moral obligation to prevent misuses of tests – tests are like other products to the extent that manufacturers are responsible for them.

Also connected to Messickian concepts of consequential validity (and Bachman and Palmer’s ‘impact’), Shohamy (2001) describes issues related to Critical Language Testing (CLT): a) the examination of the intentions of tests, “to assess and negotiate knowledge or to define and dictate it” (p. 337), b) the need to rely on other sources of knowledge, including that
of many stakeholders, c) the use of multiple procedures for testing instead of relying on one source, and d) the need to enter into dialogue and debate. The most critical point Shohamy makes is that those involved in testing must be aware that they do not have a monopoly on ‘truth’. She quotes Freire’s (1985) assertion that, “[b]y believing they possess the truth, the evaluators act out their infallibility” (p. 25). Instead, we must always question what we do as testers and why we do it. We must “rethink our priorities and responsibilities in language testing research” (McNamara, p. 333) in order to respond to and serve the needs of testing stakeholders.

Lynch’s (2001) article, as he says, is an act of exploration of the role of Critical Theory in language testing. After going through a short history of Critical Theory, and more specifically, Pennycook’s work in Critical Applied Linguistics, Lynch describes different points of view on ‘alternative assessment’ including Shohamy’s CLT. He comes to the conclusion that “the challenges of the critical perspective have, at times make me think that language assessment of any sort is incompatible with such a perspective. However, I am glad that critical applied linguists like Alistair Pennycook and critical language testers like Elana Shohamy have asked us to take a step back and put our assessment research and practice into the broader social, political and cultural picture” (p. 369). While the state of “constant scepticism” (p. 368) is important and allows for the “best chance at thoughtful engagement” (p. 368), we must not allow it to paralyze our testing practice.

Cumming (2002) provides a good example of how ethical issues must be balanced in large-scale, high-stakes testing. He takes a strong stand on the criticism of the ethics of the Test of English as a Foreign Language (TOEFL), specifically the TOEFL Essay. Cumming (2002) reminds us of what Brown (1996) and Bachman and Palmer (1996) have said: in high-stakes testing we need to balance two conflicting groups of ethical issues, and must continually make trade-offs. On the one hand, tests must be the same so that all candidates get an equal opportunity and so that scores will be comparable. In addition, tests must provide confidentiality and the possibility for prior orientation. On the other hand, alternative assessment calls for personal development, inter-cultural negotiation, various sources of knowledge, and emancipation.

Cumming (2002) illustrates that it is not possible to meet both sets of criteria at once. Once you ensure the second set of issues are dealt with, you lose the confidentiality and sameness of test circumstances; it is a trade off between fairness and freedom from construct-
irrelevant variance, and emancipatory value. While the testing context in this research project is not nearly on the scale of the TOEFL, it is a relatively high-stakes series of tests administered to people representing an enormous variety of cultural, linguistic, educational and age groups, much like the TOEFL. Therefore, it is critical to keep in mind that all the criteria for an ethical test (much like all the criteria for a ‘useful’ test) cannot be met completely. Some reasonable level of balance is the most one can hope for.

It is increasingly clear that tests and testing practices cannot be judged simply by considering the reliability and construct validity of the test. The opinions of stakeholders in the testing process at this institution help provide a clearer picture of the ethical status of the process.

2.4 Computer-Based Testing

Influenced by these changes but a separate source of change itself comes from the rapid changes in technology. Computer-based testing (CBT) is used widely, and although pencil and paper (P&P) tests are often simply transferred to computer to facilitate scoring, the changes in technology make the potential of computers in testing a possible impetus to vast changes in testing practice. CBT constitutes what Chalhoub-Deville (2001) calls, “disruptive technology, that is, technology that changes how we think of and implement our operations” (p. 2).

While Brown’s (1997) article is fairly dated in computer technology terms, it is widely cited and provides a good framework for looking at the major issues in how computers are and can be used in testing. The four issues that Brown notes that come up with some consistency in research in CBT are item banking, computer-assisted language testing, computer-adaptive language testing, and the effectiveness of computers in language testing.

It is interesting to see how far CBT has come since Brown wrote this article. It is still true that questions in which candidates select answers (multiple choice, true-false, matching) are easy to adapt to CBT, and more interesting types of questions (role plays, interviews, presentations) are not. However, advances have been made in what CBT can facilitate. For example, for the Graduate Record Examination (GRE), Educational Testing Services (ETS) has been experimenting with e-rater – computer software that scores essays, with a high degree of inter-rater reliability when compared with trained human raters (Fulcher, 2000). The institution studied in this research project is in the process of piloting Accuplacer, a CBT that uses similar technology to score test takers’ writing.
Brown lists advantages in two areas: test considerations and human considerations, and the disadvantages in two areas: physical considerations and performance considerations. In discussing issues that should be addressed in the future, Brown identifies three categories: design, scoring and logistical issues. All of these are areas that still must be considered in designing a CBT if that CBT is to be useful for institutions as well as ethical and fair to candidates taking the test.

Dunkel (1999) asserts that computer-adaptive testing (CAT) is becoming more and more common, and we are beyond having to ask if we should be using CAT (p. 77). Instead, we should now be asking how we can best use it. She quotes Brown and is clearly building on the questions he posed two years earlier. Dunkel notes that few guidelines exist to help developers. She stresses that there is an obvious need to create guidelines given the number of groups that may be interested in developing their own CAT including educators, licensing boards, corporations such as ETS, departments of education, departments of foreign languages at universities. It is arguable, however, that guidelines or standards for testing are necessary in all testing practice, not just in CBT (Davidson, Turner & Huhta, 1997). Dunkel divides her considerations for developing CAT into four categories: 1) the basic principles of assessment in CAT (reliability and validity), 2) psychometric and technical issues peculiar to CAT (compared to a P&P test), 3) quality and availability of hardware and software, 4) administration of CAT.

Dunkel’s search for the ‘ideal’ CAT and guidelines for developers and users very much mirrors the kinds of questions I am interested in. Where she focuses on CATs, my focus is on entrance and placement tests, be they CBT, CAT, or P&P. However, where she discusses design aspects, reliability, and validity, I would add ethical issues as well.

According to Chalhoub-Deville (2001), CBT is mostly used to facilitate the administration of exams. While this is useful, we need to go beyond this to look at how this ‘disruptive’ technology can lead to fundamental changes in L2 testing. More meaningful, complex test tasks can be created than with P&P tests. This ties in with Dunkel’s assertion that CBT will be used; what we need to do is learn to use it better. In fact, Chalhoub-Deville seems to be taking this idea one step further – not just using CBT better, but re-visioning the way we assess candidates. Dynamic Assessment will likely be an important part of this revisioning.
2.5 Potential of Dynamic Assessment (DA)

A possible way to integrate the changes in these three areas, the changing approach to language testing, the notion of the ethical test, and CBT, is Dynamic Assessment (DA). An in-depth exploration of DA is beyond the scope of this research project; however, this overview of DA demonstrates the potential that DA has help us exploit the ‘disruptive’ potential of CBT.

Poehner and Lantolf (2003) describe various permutations of the DA model in contrast to static assessment (SA). DA is the “simultaneous and dialectical integration of assessment and instruction” (p.19). DA works in the Zone of Proximal Development (ZPD). Vygotsky (1978) defined the ZPD as “the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving” with outside assistance (p. 86). Therefore, DA assesses more than just a person’s actual level of ability, but his/her responsiveness to assistance. This mode of assessment looks at the future potential of the student, not just present abilities.

This has great potential in entrance and placement tests such as those given at the institution being studied. For example, part of the current assessment involves writing. We know different rhetorical styles in different cultures could put some students at a disadvantage. However, new rhetorical styles could be learned with some training, so this is not an insurmountable ‘deficiency’ but an area of difference the student must be made aware of.

Poehner and Lantolf outline the Vygotskian notion that stages in cognitive development are not culture-free, ability is not a stable trait, and by gaining perspective on a person’s future we can help them attain that future (Poehner & Lantolf, 2003). As educators, that should be one of our main goals. It is far more ethical than simply acting as ‘gatekeepers.’

Poehner and Lantolf (2003) describe two ways of implementing DA: sandwich (pretest-intervention-posttest) and cake (instruction during assessment, standard or individualized). In addition, there are two approaches to DA: interactive and interventionist DA. Interventionist tends to follow the ‘cake’ model and tends to be more quantitative and psychometric. Standardized support can be implemented, and this type of DA stands up to traditional tests of reliability and validity more easily. Although I have been unable to find any research on DA administered via computer, it seems logical that a computer program could provide standardized support and measure the amount of support needed rather than just looking at right or wrong
answers. This would allow for more depth in the information obtained about students and their ability to succeed in the institution's programs.

According to Poehner and Lantolf, little research has been done on DA in second language (L2) contexts. At the time of writing they were aware of only 5 studies. Most relevant to this research project is the study they note about the use of DA to place students in advanced Spanish language courses: “Students who were able to revise under prompting were considered to be at a more advanced stage than students who could not” (p. 13). The ability of DA to ‘fine tune’ placement is clear.

While DA does not require the use of a computer, integrating the two presents an excellent example of how computer technology can be disruptive; CBT could make DA much more practical and cost-efficient. Although the testing in this study is not currently computer-based, nor is it dynamic, considering the potential of DA provides perspective on what is possible in language testing beyond the traditional, easily scorable (e.g. multiple-choice) tests. Moving tests to a computer-based format, as will soon be done at this institution, should not mean simply transferring a P&P test to CBT. The biggest obstacle to implementing DA is likely the increased difficulty of developing tests, especially if they were administered via computer. If DA tests were not delivered via computer, the administration of the tests would certainly be more time consuming and expensive as they would have to be delivered by a trained tester one-on-one with the test taker with care taken to maintain standardization and sameness of circumstances for all test takers.

2.6 Testing of Foreign-Trained Professionals in Canada

A final area of change in testing that I wish to explore is the assessment of highly-skilled recent immigrants to Canada. Statistics Canada reports that in recent years, the number of highly-skilled new immigrants has increased, but that there is a gap (12-22%) between employment of immigrants and non-immigrants (Statistics Canada, 2004a). Of those who do find employment, they are often not working in their chosen field. Not finding suitable employment influences their decision to stay in Canada (Statistics Canada, 2004b).

Cumming's (1994) work on language assessment of recent immigrants points to ways in which assessment presents often needless barriers to participation in Canadian society. While he is talking about assessment Canada-wide (the reasons for setting up Canadian Language Benchmarks (CLB)), some of the same issues must be considered on a smaller scale as well,
specifically his question, “does the process of language assessment help or hinder certain people trying to achieve certain things?” He finds the answers are “disturbingly negative” (p. 118). Often, he says, assessment may pose barriers, be too limited in scope to be really useful, and put the burden of responsibility on the immigrants and neglect the responsibilities of majority populations and societal institutions. Immigrants, he notes, are underrepresented in certain professions because of unfair testing practices (e.g., using TOEFL when it is inappropriate), a lack of a common definition of language proficiency (an issue that was likely overcome to some extent by the CLB), inadequate testing for specific groups (sometimes demanding knowledge of grammar rather than more relevant aspects of language), and the existence of bias and the lack of appropriate support at some institutions. He points to studies that suggest that unique achievement criteria that don’t include “local cultural references, allusions, or schemata” (p. 123) may be needed to address this last issue.

2.7 Glossary

*Authenticity*: This refers to the extent to which the test tasks relate to the Target Language Use (TLU). In other words, the tasks a candidate completes on a test should resemble the kinds of tasks he/she would have to do in practice. For example, on a placement test, the tasks should resemble the kinds of activities that will be encountered in the classroom.

*Canadian Language Benchmarks (CLB)*: “national standards in English and French for describing, measuring and recognizing second language proficiency of adult immigrants and prospective immigrants for living and working in Canada” (Centre for Canadian Language Benchmarks, n.d.)

*College Entrance Examination Board (CEEB)*: Also called the College Board; “a not-for-profit membership association whose mission is to connect students to college success and opportunity” (College Board, n.d.)

*Computer-adaptive test (CAT)*: a specific type of test delivered using a computer. Test takers are given different questions depending on their response to the previous question. If they get one question right, the next question will be slightly more difficult. If they get it wrong, the next question will be slightly easier. The test taker’s level has been assessed when he/she is able to consistently answer questions at a given level of difficulty correctly.
Computer-based test (CBT): any test that is delivered using a computer. Generally, a computer adaptive test (CAT) is considered to be a type of CBT.

Construct validity*: Essentially, does the test actually assess what it claims to assess? The language ability being tested must be relevant to the purpose of the test. This must be reflected in the test tasks and scoring procedures so that resulting scores will allow test users to make the desired interpretations about the candidate's language ability.

Criterion-referenced testing**: a common form of classroom assessment in which a test taker’s ability is measured against a specific set of criteria. These criteria are usually learning outcomes for a course or a unit of study. For example, if a class had just finished a unit on forming and using the present perfect, they would be given a test that measured their ability to do this.

Dynamic Assessment (DA): the “simultaneous and dialectical integration of assessment and instruction” (Poehner & Lantolf, 2003, p.19). The process of administering a test includes offering support to the test taker. Rather than measuring right and wrong answers, the amount of support needed to complete the task is measured. Poehner and Lantolf (2003) describe two types: sandwich and cake. Sandwich refers to the pretest-intervention-posttest model. Cake refers to an assessment model in which standard or individualized instruction is administered during assessment.

Face Validity: the perception that a test is valid and fair. Also called ‘test appeal’. See Bachman and Palmer, 1996, p. 42n6.

Educational Testing Service (ETS): the corporation that produces and sells a range of standardized tests including the Test of English as a Foreign Language (TOEFL), the Graduate Records Examination (GRE), and Scholastic Aptitude Test (SAT).

Impact*: The power of a test goes beyond simple decision making on the part of score users. Impact looks at the broader effects of testing practices. A test affects the test taker. Taking the test may influence the test taker (language knowledge, perception of the target language) and should, therefore, be meaningful; feedback given by the test taker about the experience of taking the test may help the test developers improve the test; the test scores must be relevant to the decisions being made; test takers should be aware of the scoring criteria. Teachers are also affected. Tests should be consistent with teaching materials, characteristics of teaching and learning activities, and with the goals and
values of the instructional program. Beyond that, **society and educational systems** may be affected. Tests should be consistent with the values and goals of the educational system and society (although these may not always be easy to define), and potential consequences, both positive and negative, must be taken into account.

*Interactivenss*: This takes into account the test takers language knowledge, background knowledge, personal characteristics, metacognitive strategies, and affective schemata. It is not simply the psychometric, scientific quality of the test that matters, but how the test engages the test taker.

*Inter-rater reliability*: consistency of scores on the same task between different raters.

*Language Preparation Course (LPC)*: a series of courses offered at the institution being studied that focus on language skills required to complete mainstream Communication classes. They focus heavily on grammar and writing (paragraphs, letters, memos) on business and technical topics, but also include reading, speaking and listening components.

*NS*: Native speaker of English.

*NNS*: Non-native speaker of English.

*Norm-referenced*: usually a standardized proficiency test in which a test taker’s abilities are measured against a norm or ‘bell curve’. A few test takers will do extremely well or poorly, and the majority will fall in the middle.

*Pencil and Paper test (P&P)*: a test that is written with a pencil and paper as opposed to on a computer.

*Placement test*: used to determine the appropriate level of instruction for a test taker.

*Practicality*: How easy is it to implement this test? The amount of resources necessary for the design, operationalization, and administration stages of the test must be considered, along with the resources available for carrying out these stages.

*Proficiency test*: a norm-referenced test used to determine a test taker’s overall abilities.

*Reliability*: refers to the extent to which similar results are produced at different times, in different settings, and on different forms of the test. A candidate should get similar results if they write different forms of the test, on different days, or in different testing centres.
Test of English as a Foreign Language (TOEFL): a test designed and sold by Educational Testing Service (ETS) used to measure the English language proficiency of a non-native speaker (NNS). The test focuses on academic English and English related to American campus life. Often used to determine if an NNS will be granted admission to an English-language college or university.

Target Language Use*: the type of language that the test taker will require outside of the test.

Usefulness: defined by Bachman and Palmer (1996) as the balance of reliability, construct validity, authenticity, interactiveness, impact, and practicality.

Zone of Promixal Development: the difference between what someone can do with guidance (potential level) and what someone can do without guidance (actual level); “the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving” with outside assistance (Vygotsky, 1978, p. 86); the “potential level minus the actual level” that a person demonstrates (Johnson, 2004, p. 109).

* Explanations adapted from Bachman and Palmer (1996)

** See Brown (1996) for a more detailed discussion of these terms
3 Research Method

A qualitative approach was used to explore the usefulness and ethical status of language entrance/placement testing at a post-secondary institution. This was done by obtaining the opinions and perceptions held by stakeholders in the testing process including test takers, test score users, and test developers/administrators. The stakeholders were asked to contribute their opinions and perceptions through focus groups, questionnaires, or an interview. These were conducted with 15 participants from September to November of 2005 at the institution’s main campus.

While the test developers/administrators mentioned Accuplacer, the CBT that was being piloted to replace the tests developed in-house, this was not the focus of this research project. The questions asked in the focus groups, questionnaires, and interview were regarding the testing that was in place at the time using the tests that were developed in-house. Unfortunately, I was not able to have first-hand access to these tests, so rather than analyse and evaluate the tests themselves, I focused on the testing process as perceived by the stakeholders. The apparent lack of documentation about the development of these tests was another limitation of this study.

3.1 General Approach

Initially, I intended to use focus groups to explore how stakeholders perceive the English language tests: the purposes; the (face-) validity; the effects they perceive the tests as having; the constructs and content they see the tests testing; generally how satisfied they are with the tests; and how useful the test scores are. Instead, for reasons outlined in the discussion section, I used only three focus groups, two of which were part of the pilot study. These two groups were quite small with only two participants each (one for score users and one for test takers). For a variety of reasons, most importantly convenience of scheduling, only one of the focus groups in the main study ran as intended: the test developer/administrator group (four members). In addition, I conducted one interview with a score user and received written responses from six other score users.

I used this approach, talking directly to stakeholders, because, as Shohamy (2001) and others (Hamp-Lyons, 1997 and Lynch, 2001) note, it is important for test developers to listen to the voices of stakeholders, especially those of test takers. The focus groups allowed me to
obtain rich information fairly easily and inexpensively, and the interview and written responses, while still providing rich information, allowed the participants the security of not having to put themselves at (potential or perceived) risk by voicing their opinions in front of their colleagues.

Through these methods, I got opinions on the ethics and usefulness of entrance/placement testing from test-takers (successful and unsuccessful), test developers/administrators, instructors, and other administrators. I had intended to divide the stakeholders into three separate focus groups: one of successful and unsuccessful test takers, one of instructors and program administrators, and one of test developers and test administrators. Although the focus groups did not all convene as intended, I maintained separate sets of questions for each of these groups. Of course, this does not cover all stakeholders, for as I mentioned earlier, the impact of the tests is far-reaching. It does, however, cover those most immediately affected and those likely to be familiar with the content of the tests.

I originally intended to use focus groups for a variety of reasons; as mentioned above, they allow for the collection of rich data fairly inexpensively. In addition, Fontana and Frey (2000) note that focus groups are useful for exploratory research. In many ways, that was what I did – explore how the tests are perceived and how stakeholders are affected by them. Fontana and Frey (2000) also state that focus groups may serve phenomenological purposes; while I did not intend to do a purely phenomenological study, this aspect was especially valuable for building a picture of what the experience of taking the test is like for the test takers. As well, focus groups provide a manageable forum for stakeholders to share opinions and ideas, and this format can provide a more dynamic and open environment than interviews or questionnaires. As Gall, Gall, and Borg (2003, p238) suggest, “interactions among the participants stimulate them to share feelings, perceptions, and beliefs that they would not express if interviewed individually.” The focus groups that did run certainly exhibited these qualities.

As mentioned earlier, aspects of critical theory were be crucial to my investigation. It may not be practically possible for tests to be entirely emancipatory or to conform completely to the requirements of critical theory, as Lynch (2001) points out. However, I believe it is important to rethink the current testing paradigm, and to expand the paradigm to include practices that better serve test-takers and, therefore, are more beneficial to the test taker, the institution, and society as a whole.
3.2 Participants

The participants in this study were, as previously mentioned, stakeholders in the testing process. They included test takers (both successful and unsuccessful), test score users (instructors and one administrator who deals with students), and test developers/administrators. There were only two test score users who took part in the pilot study: one was successful on the test and taking courses at the institution, and the other was a foreign-trained doctor who was not successful on the test. Most of the test score users who took part were instructors in the Communication department. However, there were a few from other departments (one Language Preparation Course instructor, one instructor from the School of Business, and one Continuing Education Accounting instructor). Only one score user was not an instructor but an administrator who deals with students frequently. Two of the test developers/administrators were also instructors of Language Preparation Courses, one was also an instructor in the Communication department, and one was both an instructor in the Communication department and an administrator. All of the score users and test developers/administrators had many years of experience at the institution.

3.3 Procedures

Before running the focus groups, I conducted a pilot trial with a smaller number of members of the stakeholder populations (except the test administrator group since there are only four members of this group at the institution) to both determine if the questions were clear as well as to elicit relevant information. Once this was done, I made necessary revisions to the questions for the test takers. Test score users were satisfied that the original questions were clear and helpful.

I used stratified purposeful sampling to obtain information from as many different stakeholders in the testing process as possible, representing different aspects of testing. Therefore, I solicited participation from stakeholders in the testing process: test developers and administrators, instructors (whose classes successful test takers enter), program administrators, and test takers (both successful and unsuccessful).

To solicit participation from test-takers, I posted an ad on the website where students check their grades. While this may have somewhat limited the respondents I got (people comfortable using computers), I believe this limitation was minimal – students (and potential students) are expected to register for tests/courses online and to check their results online, so it
is likely that the majority of students and potential students are computer-literate and have access to computers with internet access. To lessen this limitation, test administrators gave flyers out when a test was given so those interested in participating could contact me.

To solicit participants from the remaining populations, I emailed the faculty and staff and posted a notice on the institution’s website asking those who are involved in or impacted by testing to contact me if they wished to participate in a focus group. With these populations, I also used snowball sampling, as there are a few well-placed people who recommended highly-qualified participants. This was intended help ensure stratification of the sample. Both the pilot and the focus groups were held at the institution’s largest campus.

Verification of the data was done by engaging in informal peer debriefing sessions with my advisor at UBC, Dr. Ken Reeder. Since Creswell (1998) suggests that two methods of verification are sufficient, I also used triangulation of data to some extent. The most significant issues that came up were those shared by all three groups. However, since the opinions and perceptions of all three groups are equally important, I did not want to rely too heavily on this kind of triangulation. The fact that there are two groups of employees puts the one group of test-takers in the minority. As demonstrated earlier, it is necessary to take test-takers’ opinions and perceptions into account. Triangulation with the two other groups could potentially defeat this purpose.

In order to help readers outside the institution determine if the results are transferable to other settings, I have contextualized the results with a description of the research setting, as Creswell (1998) recommends.

3.4 Data Collection

The focus groups that did meet, the pilot with test takers, the pilot with score users, and the main study with test developers and administrators, were convened in a meeting room, my office, and a classroom respectively. Light refreshments were served and the discussions were tape recorded. Each focus group lasted roughly an hour.

Many attempts were made to convene focus groups with score users as part of the main study; however, as will be discussed later in chapter five, it was not possible to arrange for the groups to meet. For the most part, score users emailed me their responses, and sent their consent forms to me in the institution’s internal mail. One score user preferred to be
interviewed. This interview was conducted in the score user's office and was tape recorded. The interview lasted about 45 minutes.

I received emails from 13 test takers who expressed interest in participating, and I attempted to arrange a time that was convenient for them to meet. I emailed them requesting their preferred times and locations, and a focus group was convened for the pilot study. Unfortunately, I was not able to arrange a focus group with test takers for the main study. The reasons for this will be discussed in chapter five.

Complete lists of the focus group questions can be found in Appendix A.

3.5 Protection of Human Subjects

Two of my stakeholder groups consisted of colleagues and superiors of mine. Therefore, there was little danger that any power differential would compromise them. However, my position as a colleague and not an outsider and, in the focus group that did meet, the presence of their colleagues may have made people more hesitant to express what they may have thought were 'unpopular' or 'radical' opinions. Therefore, I tried to divide these stakeholders into groups of participants who held positions at similar levels of power. In addition, I ensured that all participants were aware of the potential power differential, and when requested, I let participants know in advance who the other participants would be. This way they would be prepared for any possible differences in power and could decide freely if they were willing to consent to participate anyway. However, with the score user group, some potential participants still did not feel comfortable participating, even one-on-one with assurances of confidentiality. This issue is discussed further in chapter five.

It is also important to obtain the opinions of test takers; however, protecting them was more difficult given that I am an instructor at the institution, and therefore have a certain degree of power. To minimize this, I offered assurances of confidentiality and the right to withdraw at any time. I also offered assurances that their grades and status in their courses would not be affected by their participation.

I informed all participants that information from participants would be kept confidential and materials would be kept in a locked cabinet for at least five years. I also informed them that once the materials were removed from the locked cabinet, they would be destroyed. All participants were given a consent form to sign. Generally, those who take the entrance test have a sufficient level of English to be able to deal with basic written material. However, I was
careful to word the consent form as clearly and simply as possible and was willing provide translation if necessary. Participants were told they were free to withdraw from the discussions at any time without consequence.

3.6 Summary of Research Methods

This qualitative study used focus groups, questionnaires and an interview to obtain stakeholders’ opinions and perceptions of the usefulness and ethical status of language entrance/placement testing at a post-secondary institution. Participants were recruited using stratified purposeful sampling and snowball sampling. The stakeholder groups included test takers, score users, and test developers/administrators. While a limitation of this study is that I did not have first-hand access to the tests, the intention was to see the tests from the stakeholders’ perspectives.

Test developers/administrators mentioned Accuplacer, the CBT that is being piloted to take over from the in-house tests that are currently in use. However, the focus of the study was the existing testing practice, and the questions asked of the participants were about this. Data collection was carried out with 15 participants at the institution’s largest campus from September to November of 2005.

The next chapter will cover the findings of the research project in terms of ethical considerations, psychometric considerations, influences on and of the test, and institutional considerations.
4 Findings

This chapter outlines the method of data analysis and the results of the research. The results are categorized into the different aspects of ethics and usefulness. First, the ethical considerations covered include equal opportunity to succeed, benefit maximization, and respect for persons. Usefulness is covered in terms of psychometric considerations (including reliability and construct validity), influences on and of the test (including authenticity, interactiveness, and impact/consequential validity), and institutional considerations (including practicality and face validity/respect for the authority of the test).

4.1 Data Analysis

Data was collected via three focus groups (test administrators, four members; test takers, two members, test score users, two members) and one interview with a score user. Generally, score users preferred to email me their responses to the focus group questions. While the test taker group had only two members, I did receive emails from 13 test takers in total, many of whom provided helpful information in their initial email. However, I was unable to set up a time to meet with the others, and they did not respond to my request to answer the questions on email.

Once I conducted the focus groups and the one interview, I transcribed the discussions. After I received the completed questions via email from participants who preferred that option, I compiled all the emailed responses into one document. Then I coded and categorized the data based on Bachman and Palmer’s (1996) definition of usefulness and Hamp-Lyons’ (1997) principles of ethics. I looked for other emerging categories, but overall, the relevant discussion fit into these categories. I then organized these sub-categories into broader groups: psychometric considerations (reliability, construct validity), ethical considerations (equal treatment, respect for persons, benefit maximization), institutional considerations (practicality, face validity/respect for the authority of the test) and other aspects of usefulness (authenticity, interactiveness, consequential validity/impact).

The goal of this analysis was to determine if there were any common issues among stakeholders, and which issues were most pressing to stakeholders. In addition, I was interested in determining if there were any issues that were perhaps not mentioned by a variety of stakeholders, but would merit future investigation. Hopefully, this information will provide a
more complete picture of the testing process and how the tests are perceived and help guide the development and implementation of new tests.

4.2 Results

4.2.1 Ethical Considerations

The three concepts that make up the ethical qualities of a test (equal treatment, benefit maximization, and respect for persons) are inter-related and difficult to separate from each other. Of course, this is true of all the test characteristics, but if conceptualized as a Venn diagram, these three characteristics would overlap far more than any of the other characteristics. Indeed, they bleed into almost every other category. The divisions I have created for them, therefore, are quite artificial but should be useful for the exploratory purposes of this research.

4.2.1.1 Equal Opportunity to Succeed

There was agreement among all stakeholders that the test takers were all treated equally in the testing situation. As one test administrator pointed out, all test takers experience the same rigorous conditions that are not likely conducive to optimum performance, “We talked about that we should have it on a Saturday in the morning when they are fresher. I think writing a test after a work day on a weeknight until nine o'clock at night is not the optimum testing. I pity them. That would be hard for me to do. For those reading questions you really have to focus, and you got all your life going on and you are really concerned about this test – it can be the difference of your job, your career in a new country. That’s a lot of pressure for nine o'clock at night – that’s when they leave. You know they start at 6 p.m. And they're hungry I think usually. Hunger is a factor in this too because you may have eaten and you may not, and there is only a 10 minute break... So they are all tested in the same conditions – rigorous conditions, but I don't think I would give the best results.”

In addition to the timing of the test, according to the same test administrator, the location of the building can add to a test taker's stress: “I think for some students, even the location is problematic because they can't find the building, they can't find the parking. They maybe take the bus, and it's in the middle of sort of nowhere. There is no map that's put on that [confirmation form]. So they come in and they are late, and they are flustered. And that's a bad way to start a test too – where you are already thinking about everything else.” These issues are also related to construct validity: the setting of the test could privilege some test takers over
others. For example, test takers who are not arriving directly from a full day of work may have an opportunity to perform better on the test.

Another issue a test developer brought up was that, while they are all treated equally in the actual testing situation, some unsuccessful test takers had been allowed entrance by a departmental Dean who overrode the test score. Some find other ways to circumvent the testing process. For example, some unsuccessful test takers enrol in courses at a private language school that has entered into a partnership with a department within the institution. According to a test administrator, this partnership was entered into without consulting the Communication department and is not recognized by the Communication department because, “it's never been articulated, nor are they part of the articulation process with [another reputable local post-secondary institution], or anyone else, to grade 12.”

4.2.1.2 Benefit Maximization

Timing and location were the main concerns mentioned in the test administrators’ focus group when we talked about equal treatment. While the test does meet the requirement that all candidates be treated equally, it could be improved in terms of the ethical quality of benefit maximization; test conditions that are not conducive to optimum performance benefit neither the test taker nor the institution. Interestingly, though, the test takers did not mention the timing and location of the test as problematic.

Among test score users and test administrators, there is a general attitude towards the test as primarily a gatekeeper: its job is to keep people out. This was expressed rather overtly in some cases, “No, the tests don’t allow them to exhibit their strengths, but that is not the purpose of these tests. Yes, their weaknesses are adequately apparent.” In other cases, it was more subtle, “I would say that it works. Wouldn’t you say? The people who do have a good level, they are successful in whatever they’re put in. So it's an effective filter or assurance that people will be successful.” One score user repeatedly referred to the test as a roadblock, “It’s just the difficulty and the roadblock that’s put on the students is my main concern.” Using a test as a gatekeeper does not maximize benefits to the test takers or to the institution. In addition, as will be discussed later, the perceived gatekeeping role impacts the face validity of the test as well.

4.2.1.3 Respect for Persons

The issue of the timing of the test and its location are also of concern when considering respect for persons. These ‘human’ issues are crucial to testing and taking them into account.
shows respect for the individuals taking the test. Starting the test, as mentioned above, hungry and flustered, would not likely make a test taker feel respected. Again, it is somewhat surprising that the test takers did not mention these issues, but I believe that is likely explained by the fact that the power of the test makes the test takers feel that these are things that must be put up with.

Another very ‘human’ consideration is, similarly, the state of mind of the test taker. One test administrator noted that given the short time and the nature of the test, “we can’t tell whether it was just a bad day that they had. One immigrant coming in, who is not used to this test ... maybe if they knew better or whatever or got up to speed with it, they would do better. I don't think we have any feedback on that.” Another administrator agreed, but suggested that there was little that could be done about that aside from ensuring that the test itself did not add to this problem: “I think in any test situation, you're going to have those anomalies and is not much you can do about it. You just try to set the test in such a way that minimizes that.”

Yet another test administrator expressed confidence that Accuplacer, which is being piloted to potentially take over from the current tests, would help deal with some of these issues. Because Accuplacer does not have time limits (the current tests allow test takers 40 minutes to complete the reading passages and answer the multiple choice questions, another 40 minutes to complete the grammar questions, and 60 minutes to complete the writing section), test anxiety will likely, in his opinion, be reduced. In general, the test administrators expressed certainty that they new CBT test would “take care of” test takers. Also, some issues with construct validity (bias) and the lack of transparency could be considered evidence of lack of respect for persons.

4.2.2 Psychometric Considerations

These are the ‘classic’ measures of the quality of a test: reliability and construct validity. A limitation of this study is that these tests were not available to me first-hand. However, my interest was more in the perceptions and experiences of stakeholders in the testing process. While I did not undertake a statistical analysis of these qualities, they were mentioned by participants, directly or indirectly. The test developers/administrators mentioned Accuplacer in these discussions. However, as Accuplacer was introduced late in the stages of this research and was still in the piloting phase at the institution, an in-depth review of Accuplacer is beyond the scope of this research project.
4.2.2.1 Reliability

This was not addressed much in the data I gathered. One issue that did arise, rather indirectly, was that the test administrators did not seem to be concerned about the potential impact that changing from a P&P test to a CBT could have on the test and on test takers’ performance and therefore, on the reliability of the test across formats. It was assumed that if the surface structure of the test (multiple choice grammar and reading sections plus a writing section) is the same, then it will yield similar results: “The format is basically the same it's just on the computer rather than paper. The reading and writing the grammar - it's all the same. So I think we can probably talk about Accuplacer – it's just another format for [Test B].” Research shows that this transference, unfortunately, is not so simple and may result in changes to the construct being tested (Alderson & Banerjee, 2002a).

The test administrators were concerned, however, about more systematically determining the reliability of the scoring on the writing section of the test. Although I cannot assess how comparable the writing prompts are on the current tests and on Accuplacer without access to them, the test administrators do intend to check the inter-rater reliability. Since Accuplacer has a computerized system to mark the writing, this will mean checking inter-rater reliability between the two human raters and the computer.

4.2.2.2 Construct Validity

Initially, the test administrators agreed that the construct they wanted to test was the language proficiency required to handle first-year Communication courses: “I think the goals are specifically to work towards the expected writing and language proficiency required to handle first-year communication courses ... Can they handle the skills necessary to be successful in first-year?” Test A is a bit different. It is not an entrance test but a placement test to determine the appropriate level of language preparation course (LPC) for test takers. The construct being tested with Test A is language proficiency required to complete a given LPC level. While these are quite general definitions of the construct, the test developers/administrators felt they were reasonably accurate descriptions of what they intended to test.

In later discussion, the test administrators clarified that the construct tested on Test B is not simply English language ability for success at the institution, but a grade 12 English equivalency. This means that a successful test taker will have to demonstrate a fairly broad range of knowledge. One test administrator explained, “I think you have to keep in mind, too,
we're trying to place students with proficiency of language, but at the same time we're trying to assess that against the grade 12 education in English. And that's significantly different than just passing the test or being placed at an appropriate level. So I think so many tests just, you do the test, you pass. But going through the background in the reading ... one expects that if one has a grade 12 English language course, they have been exposed to some ideas and background. And the problem, I think, with some of the ESL students is that that's not necessarily true. So is it truly a reasonable thing, you know, grade 12 level?” It is not testing just how they deal with language, but how well-versed they are in cultural, social and other kinds of background knowledge. In order to succeed at the institution, they will need to have more than just ‘classroom’ English.

Score users were also concerned that students should have an understanding of everyday English (slang, idioms, and expressions) and context to succeed. One score user suggested that, from his experience dealing with students who have taken the test, the test must be based on ‘pure, textbook English’ since many of the NNS students were unable to cope with anything else. Among other things, another score user noted that for many of her students, “common weaknesses are in the use of idioms, standard wording and phrasing.”

As an example of background knowledge that might give one test taker an advantage over another, some of the reading passages deal with historical situations. While they are not tested on their knowledge of that situation, it is possible that it might interfere with test performance. Specifically, one test administrator mentioned that there is a reading comprehension passage on black people in the southern United States. This test administrator recognized that a test taker would need more than language knowledge to cope with this. At the same time, the same test administrator says that in general, the test covers, “a broad scope of general knowledge that you should have read something somewhere in your life.’ Therefore, this is considered to be in keeping with the test construct: a grade 12 Canadian education.

There is some evidence that it is not, in fact, measuring a grade 12 Canadian education. As will be seen later, test administrators and one score user make several references to test takers who have achieved the required C+ in grade 12 English being unable to pass Test B.

Once again, it was widely agreed by all stakeholders that including a listening/speaking component would greatly improve the construct validity. As one test taker noted, if there is a lot of difficult reading to do for a course, that is not a big problem since readings can be done on a
student’s own time, at their own pace. However, “you have to listen all the teachers, what they say” in class, at the teacher’s pace. If a student’s listening skills are poor, he/she is not able to compensate by spending more time on his/her own. A test administrator noted adding listening/speaking would improve the decision making process, “if we’ve got these questionable students with [the scores from the three sections of the test], well, did you happen to talk to them? Were they understanding during the test? So that extra piece would have made a better picture.” Another test administrator added that, “sometimes it’s interesting when I talk to people on the phone because there isn’t a good chance to talk them [during the test]. Wow, we can actually have a conversation. Wow, they failed. That’s unusual right? So the oral skills aren’t as bad as the writing might indicate.”

Test administrators also expressed confidence in the switch to Accuplacer to ensure construct validity: “by moving to computerized model, we’re going to address more of these construct-related issues mainly because the course designers are doing that constantly in feeding back to us as users what’s changing and they’re making changes all the time to it to validate these things. Well we’ve given – since I’ve been here – we’ve given five or six thousand tests on [Test B] whereas the computerized one, Accuplacer, has given five million across a wide range of abilities. We feel somewhat more comfortable doing that with what they’re doing. With ours, the only research we’ve done is to track students and see. And we did that across the board once.” This is also an issue of practicality. Developing valid, reliable tests is not easy, so test administrators felt it was more practical to use a test that has been proven to meet some standard of validity and reliability.

4.2.3 Influences on and of the Test

4.2.3.1 Authenticity

Opinions on the authenticity of the test varied among the stakeholders. One test taker said that the types of written assignments that she had to do in her Communication class were quite similar to those that were required on the test, such as letters and memos as opposed to the essay-style questions that are common on writing tests (such as TOEFL). One score user commented that the reading passages on the test were “too academic and should be revised to have a more technical/scientific flavour” in order to better match the types of reading that successful test takers would be required to do in class. While test takers are asked to write documents for a business/technical context, the test developers recognize that most test takers do not have training in business and technical writing, the types of writing they would have to
do in class. Therefore, they do not expect them to produce the specific style that would be required in class. Instead, they look for general proficiency in writing, knowing that “if the person has the proficiency with the language, they can learn and shift into a different mode.”

As previously mentioned, there was resounding agreement among all stakeholders that a listening/speaking section would improve the authenticity of the test (along with construct validity) since listening and speaking are necessary skills in any course, including the more math-based courses. One score user commented, “I think there is a misconception that a student in an accounting class (numbers) doesn’t really need to understand English, or at least not orally. However, it is very difficult for a student to learn all of the material from only the textbook (not from the lessons in class) if they do not understand due to language difficulties.” The test developers were very clear that one of the coming changes would be to add a listening section and that would be facilitated by the implementation of Accuplacer as it has a listening component.

4.2.3.2 Interactiveness

As mentioned previously, the test not only covers English language ability, but also a Canadian grade 12 education; some questions may be easier to answer (reading comprehension for example) if you have the background knowledge, the “broad scope of knowledge” that is expected. Thus, there is some interactiveness, taking into account test takers’ knowledge, personal characteristics and affective schemata. The task does presuppose some topical knowledge from test takers. This is useful to the institution since students are expected to have this education in classes. However, the concern was raised that the test “doesn’t work, I don't think, for non-native speakers, international students.” This relates to interactiveness as well as to construct validity which requires that possible sources of bias be minimized (Bachman & Palmer, 1996).

While test administrators and score users did not often have similar views on the test, there was a general consensus among members of the two groups that one group of NNSs is at an advantage over many other test takers: foreign-trained professionals. They concurred that overall, foreign-trained professionals tend to pick things up more quickly. They tend to do better in classes. As one test administrator stated, “foreign-trained professionals, as opposed to foreign and domestic non-professionals, usually have a better and faster ability to pick up the conceptual material as long as language is not a major issue,” and this is frequently reflected in their success on the test, especially the reading section. They have a broader range of
knowledge to draw on and tend to do better on the reading section than other NNS test takers who are often younger international students. Often, they do better than NS test takers who are just out of high school. Sometimes, the problem arises that although their scores on reading are very high, anecdotally, they often have a great deal of difficulty with listening and speaking tasks. One test administrator noted that, “the reading score will exceed everything, but then they can’t speak to you. [They] can’t understand the instruction in class.”

Also due to the interactiveness of the test, immigrants, whether they are foreign-trained professionals or not, seem to have an advantage over international students simply because they have more experience in Canada and generally more exposure to English and Canadian culture. One test administrator noted that, “the difference between immigrants and international students is that typically immigrants have been here for a longer period of time and that makes a big difference in language skills and cultural knowledge and background as well.”

4.2.3.3 Impact/Consequential Validity

The perception of the impact of the test varied a great deal depending on the stakeholder group that a participant belonged to. Test takers, both successful and unsuccessful, tended to view the impact of the test negatively. Test administrators saw the test as having a positive impact in maintaining standards and predicting success (predictive validity). Score users were not in agreement on the impact of the test, and perceptions ranged from moderately positive to moderately negative for a variety of reasons.

For test takers, the main areas of concern about the impact of the test fall into three areas identified by Bachman and Palmer (1996): 1) “provisions for involving test takers directly, or for collecting and utilizing feedback from test takers” in test design and development; 2) provision of relevant, complete and meaningful feedback; and 3) providing test takers with “information about the procedures and criteria that will be used in making decisions” (pp. 153-4).

Although I was able to meet in person with only two test takers, I received emails from eleven other test takers expressing interest in the research. Many of those emails expressed anger and frustration about testing. This suggests that test takers do not feel that they are involved in the testing process as anything more than subjects. I found no evidence that their feedback is sought or used to influence the testing process except in one instance: the number of phone calls that the test administrators received from students looking for more information
about their test scores did lead to changes in the amount of feedback given. A test administrator explained that, “We were getting so many phone calls, right? They knew nothing about the test, how well they did on the different parts, so that’s when we started to [give scores on each component of the test along with the final score].”

The test takers interviewed said they received feedback only on their overall score. This minimal feedback was a source of frustration. Test takers are graded on the three sections of the test, but, as mentioned above, until recently received only one grade: the lowest score of the three. One score user described how the lack of feedback impacted the students he dealt with. He suggested that “One of the problems with this is that it’s not very useful in the sense where a student knows what their final grade is, but you have to actually contact the marker in order to get a breakdown. There is no way they can just check themselves what the breakdown is. I get from other students where they say, ‘Okay, I got 35% in [Test B], but what does that mean? How did I do in each of the sections?’ And then we have to refer them to the marker for a breakdown. Then the marker will have to meet with them and discuss, and I imagine go over the exam as well. That’s kind of frustrating for students. They get the mark, but they don’t know how they placed them, where they need to improve.” As mentioned above, the test administrators have addressed this problem after receiving a number of calls from test takers wanting more information. Now, test takers can see the three scores.

Still, this was a source of frustration for test takers. The numerical scores did little to help them understand their weaknesses since the criteria for arriving at a certain score were unavailable to them. One test taker believed their grades were based on the whim of the marker. In her view, “it depends on the instructor, and his perception of, I don’t know, factors.” One score user also expressed concern that there was a great deal of room for the marker’s own opinion and mood to influence the scores on the writing section, “One thing – I kind of find this odd, and I don’t mean this as any disrespect to who is marking – it seems like they have the same marker for the whole time of the exam; they haven't rotated it. I kind of view it as kind of a monopoly in that sense. Where maybe they should rotate the exams or have multiple instructors mark the exam to get feedback. To my knowledge I think only one instructor’s marking it.” A double-marking system was suggested.

After I spoke with the test markers, I found that this is not the case; there are two markers who double mark all the tests. However, the important issue here is one that could be considered under the heading of face validity as well: the perception that the test scores really
amount to one person’s opinion. At any rate, there is no published, easily accessible list of the criteria; it appears that the markers’ years of experience and familiarity with the skills required in courses at the institution are what guide the grading. While this may work well for the test administrators, for the test takers, it was important that they understand the criteria so that they did not feel the process was arbitrary or capricious.

Test administrators expressed confidence in the test’s ability to ‘weed out’ those who would be unable to successfully take part in regular courses. In a study done on predictive validity, Test B was shown to be the best indicator of a person’s ability to succeed at the institution: “Two years ago, three years ago, we ran a survey of various streams through which people got through to [the institution] in first year, and the best predictor [of success] was [Test B].” Other test administrators agreed and provided anecdotal evidence of the predictive validity of Test B: “I had a student who wasn’t – somehow she got through Continuing Ed Grade 12, so she got into her first-year Communication. So she couldn’t manage. So she went out and went to prove, by taking the test, and she couldn’t pass the test. She did it twice, and she couldn’t pass. So [the test] was saying she wasn’t able to get in and she demonstrated she wasn't able to be in.” It seems then, that those who do not do well on the test and get in through some sort of ‘back door’ tend to struggle in classes and end up repeating them, often never achieving success. This suggests that the test scores are “relevant and appropriate... to the decisions to be made” (Bachman & Palmer, 1996, p. 154).

To assess the impact of a testing, another question Bachman & Palmer (1996) ask is how consistent the interpretations made of the test scores are with the values of society. One of the test administrators mentioned the impact on the business community served by the institution. According to the test administrator, “the graduates, the business community, the industry community that we serve is starting to really, you know, beginning to suggest that for ESL students particularly, they're not doing well. And so [the business community is] a stakeholder in all of this and that goes back to how [students] into [the institution] and how successful they are when they get out of [the institution].” Of course, by the time they get to the workforce, there will have been a number of other influences on an NNS’s success at the institution besides the entrance test and, by extension, his/her success in the business community. However, as the test developer noted, it still “goes back to how they get into [the institution]” in the first place.

Also, the lack of a listening/speaking skills section on the test, as one test administrator pointed out, can have a negative impact on the business community and the institute’s
reputation. Successful test takers with weak aural/oral skills, may be able to make it through their classes and succeed at the institution. However, when they enter the workplace, employers may find that their aural/oral skills are insufficient: “when they graduate with a certificate or diploma from [this institution] and industry looks at them and says, ‘Well, people from [this institution] can’t speak well. They have errors.’ So that’s affecting them after college.”

Bachman & Palmer’s (1996) usefulness matrix includes several questions that relate to score users. Some of the relevant questions deal with the consistency of the test with a) the language areas included in the teaching materials, b) the teaching and learning activities, and c) the goals of teachers and the instructional program. The score users’ perceptions of the impact of the tests on these areas was less unified than the perceptions of the test takers or the test administrators. Perhaps this is because score users are less familiar with the tests; they are not directly involved in testing and have not necessarily seen the test. Some were quite familiar with the content of the test, some knew about the general structure of the test (the same information that is available to test takers before taking the test), and one said that she was unaware that there was a test at all. However, there were a few areas of general agreement.

As previously mentioned, at least three score users mentioned that the language areas tested were not entirely consistent with the teaching materials and teaching and learning activities. Some mentioned the fact that the language tested was more ‘textbook-like’ than students would actually have to deal with in class. One said that, “Testing ‘pure’ English leaves students at a disadvantage to deal with other varieties of everyday/workplace English.” Also, as mentioned above, some felt that the reading activities on the test were too academic and should be more technical.

Overwhelmingly, as with other stakeholder groups, the consensus was that there needed to be a listening/speaking component to the test. Some specific aspects of oral/aural English included pronunciation, vocabulary, and listening comprehension. Students lacking in these areas were not only unable to participate and profit from the classes themselves, but were seen to slow down the entire class. It is likely the lack of this component that led one score user to assume that there was no testing process in place. As her course does not involve a great deal of writing, she would not be aware that her students may have acceptable writing skills. She would, however, notice that some students “did not understand verbal commands such as ‘Sign your name’ or ‘Leave the room’ or ‘Did you write the exam?’”
4.2.4 Institutional Considerations

4.2.4.1 Practicality

These tests are very practical to administer. It appears that practicality weighs quite heavily in the balancing act that is required to deliver a useful test. This is evidenced by the fact that there is no listening/speaking section, despite the overwhelming agreement that it is necessary. With Accuplacer, it will be easier to administer a listening test, but it is unlikely that a speaking test will become a standard part of the test in the near future.

Aside from the writing assignment on the test, everything is easily scorable. The grammar and reading comprehension questions are all multiple choice, so numerical grades are easily assigned and given to the test takers as their only feedback. One of the limitations on feedback given to test takers has been practicality. According to one test administrator, the feedback is limited by what the institution’s computer system will allow. Test takers are given only numerical grades, not qualitative feedback on what was done well and what was not because the computer system will only accept numerical data. Again, test administrators felt that Accuplacer would improve on this. Ideally, it will be able to quickly provide test takers with more than just the numbers. Perhaps it would be able to identify areas of weakness and strength, if any were apparent, so that test takers would know where to focus their energy for future tests.

As previously mentioned, test developers/administrators felt that monitoring psychometric criteria would also be easier and more practical using Accuplacer because the College Entrance Examination Board (CEEB, the maker of Accuplacer) take care of ensuring that these criteria are met.

4.2.4.2 Face Validity/ Respect for the Authority of the Test

While the tests may have redeeming qualities in many areas, face validity is an area that is clearly problematic. Test takers found some aspects of the test to be valid, but generally, they were unhappy with it. Many of the test takers were quite experienced in taking these types of tests and could be considered ‘experts’ in test taking. Several who emailed in response to my ad for participants compared it unfavourably to TOEFL. One test taker said that she had been taking language tests seemingly continuously for the last year, and she felt that she had a good sense of where her strengths and weaknesses were. However, when her test scores on the institution’s test were significantly different, she felt that she had not been given the opportunity
to show her abilities in English, "The last year of my life, I've been doing many tests: the TOEFL, the TSE, for all the school boards. The score that I got at [the institution] is the lowest of all the tests. I expected I would have high mark in grammar and lower mark in reading. It was vice versa. I got a low mark in grammar and a high mark in reading. The results that I got were unexpected. I did the [institution's] test at the end. I did the TOEFL test, I did the school board tests. I was really well-prepared for the [institution's] test. But I got lower marks than I got on TOEFL."

Another issue that impacted face validity was the lack of preparation materials. Test takers had minimal knowledge of what the test would cover and felt that they didn't have the chance to give their best performance: "I wrote [Test B]. I wanted to find information about the tests, and I asked about it, but all there was was a description that there are three parts: reading, writing and grammar. And the time, so I knew how much time there was. There were no samples available. It's not available, so I can't do it." While the test administrators have good intentions in keeping the content secret, this secrecy also serves to make the test appear to be as much a tool for wielding power as it is a bona fide form of assessment.

Test takers were also frustrated by the calculation of their final score. It seemed unfair to them that the lowest score of the three should be the one that is entered as their final grade, and therefore, they felt the scores were not valid. While test takers recognized that taking an average of all three scores might artificially inflate the grades, they felt that a different scoring procedure could benefit both test takers and the institution. For example, one test taker suggested the following procedure: a test taker must get at least 60% on each section to pass, and must have an overall average of 75%. Whether or not this specific formula would work for this institution is beside the point. This test taker had clearly given a lot of thought to the scoring of the test and was not, as might be expected, just interested in a scoring system that would allow her easy access to the institution. She was, instead, aware that standards for language proficiency must be maintained, but also aware that some test takers were being punished for low scores in one area even if their other scores were high. This is a good example of an area where benefit maximization, for the test takers and the institution, could be greater.

Some score users agreed that this appeared to be nothing more than an attempt to keep people out. Many students whose skills may be weak in one area are generally able to compensate for those weaknesses by drawing on their strengths. This is especially true, it was suggested, of foreign-trained professionals who have good study skills and high motivation.
These qualities, along with their strong technical knowledge, could be put to use to make up for weaknesses in, for example, the grammar section. One score user noted that the grammar section itself was problematic for test takers "whose expertise is embedded in their workplace knowledge who can't transfer it to multiple-choice grammar test." This possibility was reinforced by one test taker, a foreign-trained doctor, who said that although she got 80% on the writing and reading sections, her score of 50% on the grammar section was entered as her final score and she was denied access to the institution.

There is also the issue of face validity for score users and other non-test takers. To differentiate this from the face validity from the test takers' point of view, this is better stated as respect for the authority of the test. It is apparent that the test is not widely accepted to be a valuable tool that should be respected. There are cases, as noted earlier, of departmental Deans overriding a test taker's score to allow him/her entry into a program. One test administrator described a recent incident in which, "this particular student had failed [Test B] across the board in July, went to [a particular local private language school], took five language courses there, got As and Bs, and then the Dean at [our institution] sends a letter to Admissions saying he deems the student has met the entrance requirements on the basis of the performance at [that private language school]. We do not recognize [that private language school], and their language requirements, it's never been articulated, nor are they part of the articulation process with [a large, reputable, local community college], or anyone else, to grade 12. So that's what some students are doing. As test takers, they're blocked here; they're going anywhere to get the requirements to get in if they can't pass this test or make the placement or they will go and do TOEFL and somehow take a test somewhere else if they can or take grade 12 extension services and get a B still fail our test afterwards. The program areas can do anything they want and admissions are not fighting the Deans. So, the Communication department is fighting the Deans."

Another example is the business management preparation course. One test administrator related a situation in which "we found out that for the [business management preparation course], students get in with a mark from a private school. They come in and are tested here and no matter what they get in Test A they go into [second highest LPC]. I found out, to my surprise at the last meeting, they don't even need to pass [that LPC] to get their certificate." Students who see the test as a barrier look for ways to get around it, such as attending other language schools, or they give up on attending the institution altogether: "we have numbers of students
who are very qualified, and ... if they were to get that English 12 C+ — and a lot of them do — it would be sufficient enough. But the problem in some of them just can't get by [Test B]. There's a lot of negativity about this exam, so it's affecting students even on I guess a mental level where they're coming in thinking, 'Oh no I'm not going to pass. I'm struggling.' And when they don't pass it once, their confidence just dips right down. And a lot of them they don't want take it again. I talked to students saying, 'you know what I've taken it twice. I don't want to take it again, but then I can't do anything until I pass this.' So then they make the choice: if they don't get in, they go someplace else.”

Some score users expressed their lack of respect for the test as well. One said she was surprised to hear there even was a test for admission to part-time studies, “I don’t know if what I heard about the absence of tests for part-time studies students is true, but if it is, the quality of tests is not of much concern to me...I would just like to see there be some tests and some minimum level of English comprehension.” Generally, score users agreed that the test did a good job of testing reading and writing skills, but that its lack of a listening/speaking component meant that it was not as effective as it could/should be.

4.3 Summary of Results

The main areas related to usefulness and ethics that came up fell into several categories including ethical considerations, psychometric considerations, influences on and of the test, and institutional considerations. While the testing process met certain criteria for ethics and usefulness, there were a variety of areas that could be improved. There was overwhelming agreement among all stakeholders that a listening/speaking component be added. Overall, a lack of communication between test developers/administrators and other stakeholder groups, as well as a general lack of transparency in the testing process, led to widespread misunderstandings of the tests’ content and purposes. This, in turn, created a lack of respect for the authority of testing, and a lack of face validity.
Discussion and Conclusions

This final chapter will first discuss how the research questions were answered by the data gathered. Next, it will go on to cover lessons learned for the researcher, for testing in general and for the institution. This will include limitations to this research project and implications for future research. Finally, recommendations for improving the balance between the various aspects of usefulness and ethics will be offered.

5.1 Answers to Research Questions

Returning to the original research questions, it is clear that some were answered more fully than others. The overarching question, how useful and ethical are the language entrance/placement tests at the institution, was addressed through the sub-questions outlined below:

5.1.1 How is the language ability construct defined? In other words, are the conditions for construct validity adequately met?

The construct is defined primarily as the language proficiency required to handle first-year Communication courses. However, the test developers/administrators went on to mention grade 12 English equivalency, including the broad knowledge that one is assumed to have upon successful completion of grade 12 English, as part of the construct. According to one test administrator, the test was developed about sixty years ago. This construct definition is too broad to meet Bachman and Palmer's (1996) requirements for the development of a construct definition. They repeatedly use words like "specific" and "precise" to describe the elements required of a well-defined construct (pp. 88-89). Also, since the construct definition would have been developed at this time as well, the most important thing that should be done to ensure construct validity is to re-examine the construct definition in light of the changes, especially demographically, that have taken place in the classroom since that time.

5.1.2 Are the tests currently being used seen to be free from bias that might unfairly disadvantage certain people?

This question is closely related to the first question. The fact that the test was developed for a very different and more homogeneous audience points to the potential for a great deal of bias in the content. It is quite likely that test takers with more knowledge of American history, as in the example given by a test administrator, will have a greater opportunity to do well.
While it might appear that this could potentially disadvantage foreign-trained professionals, the test administrators agreed that this group of test takers tended to do better on these sections than many other test taker groups. Also, given that the test takes place at 6pm on a weeknight, test takers who did not have to rush from work and who are not, therefore, likely tired and hungry, have a better opportunity to do well.

5.1.3 Are the skills tested relevant to the skills required to succeed at the institution?

For the most part, stakeholders found that the test generally tested relevant reading, writing, and grammar skills. Some did mention, however, that testing more everyday and business English, as opposed to ‘textbook’ English would be helpful. Without question, the most important change that needs to be made to make the test more relevant is to add a listening/speaking component. This was an issue that came up repeatedly with all stakeholders. Listening and speaking are clearly crucial for success at the institution, yet are not tested at all.

5.1.4 Are the scores obtained adequate reflections of the construct being tested, and, therefore, useful in determining whether a candidate should be able to enrol at the institution?

Given the studies done on predictive validity that a test administrator mentioned, it appears that the scores are useful for this purpose, but the lack of a listening/speaking component certainly detracts from this. The issues with construct validity and bias outlined above also impinge on the usefulness of the scores. Also, that the lowest score on the three sections is what is recorded as the test taker’s final grade suggests that the test taker’s true abilities may not be reflected in the scores.

5.1.5 Are the scores useful beyond simply determining if a candidate is ‘in’ or ‘out’? For example, can test takers use the feedback they receive to target areas of language they need to improve?

In the past, test takers received only one numerical score or a ‘U’ for ‘unsatisfactory’. This provided, obviously, very little guidance to test takers. Some improvements have been made to the type and quality of feedback the test takers receive. However, the scores they receive are still simply numerical scores, albeit with reference to the section of the test they received that score on (reading, writing, and grammar). This provides them with more guidance as to what language areas they need to work on, but it is still far from the ideal of providing diagnostic feedback that will help test takers work towards achieving a sufficient level of
language to enter the institution. Hopefully, in the future, the new test that is being piloted will allow for a greater quality and quantity of feedback.

5.1.6 Can the institution use the scores to more accurately place students in programs where they will receive the support they need?

Providing language support does not appear to be a priority at the institution. There are five language preparation courses (LPCs) offered, but once a student has been accepted to a regular program, there is very little support available. In addition, if a test taker is unsuccessful on a test for entry into a regular program, they must take another very similar test (which they have to pay for again) to be placed in an LPC. Test administrators hoped that the new test being piloted could be used for both functions: determining if the candidate is ready for the regular stream, and if not, placing the test taker in the appropriate LPC.

5.2 Lessons Learned and Implications for Further Research

5.2.1 Lessons for the Researcher

If I were to do this again, I would probably send out questionnaires widely and interview a few key participants. For all the reasons stated earlier, focus groups can be very valuable. In this case, however, focus groups were not workable, with the exception of the test administrators group. Since that group is small (four people) and fairly cohesive, it was easier to arrange a meeting. It also proved to be valuable for the test administrators themselves: they expressed gratitude for the opportunity to get together and talk about issues. In the end, the richest data I obtained was from this group, validating the previous assertion that focus groups would provide a dynamic format in which to obtain rich information.

For the other groups, it was not a successful methodology. The test takers expressed a willingness to meet in groups, but given the varied course and work schedules they had, I was only able to meet with two test takers. This was quite surprising to me given the positive response and willingness, even eagerness, to talk about their experiences that many expressed in their emails responding to my advertisement for participants. In the future, were I to attempt to convene focus groups from the student population, I would offer to provide lunch or enter the participants in a draw for a prize to provide more incentive to participate.

The lack of success in conducting focus groups with score users was the result of a variety of factors. Of course, their varied schedules played a big part; even though there were several people who were willing to meet in a focus group, it was not possible to find a time that
suited even two or three score users. Beyond the practical logistics of convening a focus group, the very contentious and political nature of testing meant that many score users were unwilling to talk in groups, and a few answered my questions only after confirming that their identities would be kept confidential. Others refused, despite those assurances of confidentiality, because they felt that they might face negative consequences for voicing their opinions.

Also because of the contentious and political nature of testing, I felt that I was limited in the questions I could ask and, therefore, the amount of information I could get. I often felt that I was 'sticking my nose in'. For example, my requests to get copies of the tests to show to score users during the focus groups (which never materialized anyway), were responded to positively, but the test administrators were not actually forthcoming with copies of the tests, despite my promises to keep them secure.

5.2.2 Limitations of the Research Design

Given the scope of this research project, I had to limit the number of stakeholder groups involved. It would be valuable to conduct a similar, larger study that included more stakeholder groups. Some of these groups include non-test takers and NS students in classes with NNS and other test takers, for they are impacted by the test as well. As one test administrator noted, the business community and employers would also likely have useful insights. However, since the tests in question are entrance tests, not exit tests, the impact they have on the business community is not nearly as direct.

Until a code of ethics is established, and indeed to help determine what it must include, an analysis of current testing procedures based on the ILTA code of ethics would help determine in what ways testing at the institution is successful ethically and the areas that need to be investigated further to make improvements.

Of course, an analysis of the reliability and validity (especially predictive validity) of both the current entrance/placement tests and Accuplacer is crucial. I was unable to obtain a copy of Test B, so I was unable to analyze it. Given that Accuplacer was introduced in the later stages of this project and is in only the piloting stages at the institution, an analysis of it is beyond the scope of this study. However, it is important that the psychometric characteristics of the tests are examined. While much of the literature on ethical testing and test usefulness involves more qualitative methodologies, the classic measures of reliability and validity must not be forgotten.
5.2.3 Lessons Regarding Testing in General

Before I began this research project, I was aware I would not be looking at simply the institution’s testing process and people’s experiences with it but at people’s attitudes towards testing in general. Even among instructors with a great deal of experience, who would necessarily be doing a great deal of testing as part of their courses, and people directly involved with test administration, testing tends to be seen as a means of deciding who is in or out, who passes or fails, and those who are out or fail are clearly not ready for the program. This view of tests is neither ethical nor useful. I believe research into the types of tests most often used, the perceptions of these types of tests, and the perceptions of the value of different assessment methods would yield helpful information that could be used towards professional development at the institution. Since testing plays such a large role at the institution, it would be useful to investigate how it is carried out and how instructors who develop and administer their own tests view this activity.

The test developers expressed a great deal of faith in the work that others had done to prepare the language entrance/placement tests. It was stressed that people’s best efforts were made to make the current test as valid and reliable as possible. In addition, as mentioned earlier, Accuplacer is seen as a kind of panacea for entrance and placement testing with all the work of ensuring reliability and validity done by professional test developers. While this is indeed practical, it is not without problems. Accuplacer, for all of its positive attributes, is a one-size-fits-all, off-the-shelf test. It is important for future research to investigate how well it fits and how useful it is for the institution.

5.2.4 Lessons for the Institution

5.2.4.1 Implications for Policy

It is interesting that the institution chose Accuplacer and expressed such faith in it when the test developers and many others at the institution believed that their Communication courses were unique and, thus, the language requirements were unique. This is demonstrated by one test administrator’s comment that, “I suspect we are rather unique too at [this institution] because – I will focus is on what we do here in the communication department – we really don’t care what the rest the world is doing. You know, because universities, colleges and that, have a different, well their English departments are looking for different things than we are. So I rather suspect that the language testing community, we [referring to the four focus group members] are here.” There was a general attitude that the institute was such an anomaly, filling such a specific niche,
that there was little or nothing to be learned from research on language testing or the experiences of other institutions, that they somehow exist outside of the typical concerns around language testing. Yet at the same time, the test developers were quite happy to rely on the CEEB to provide them with a test that is considered reliable and valid for the institution's purposes. If the test administrators have not already studied available reviews and done an analysis of the usefulness of Accuplacer, using Bachman and Palmer's (1996) usefulness matrix, for example, I would recommend that reviews be studied and usefulness be analyzed in future research projects (Brown, 1996).

It is quite common for testers, and often the test takers and general public, to have a great deal of faith in tests to do what they want them to do. Because they were designed by professionals, ready-made tests are often accepted as valid and reliable with little thought to what they were designed to do or if they were designed well. This is part of the symbolic power of tests that Shohamy (2001) describes. Tests like this also provide testers with the appearance of neutrality and objectivity: the “burden of proof” of the validity of each score is placed on the test itself, “while the tester remains a neutral observer, shrugging off all types of responsibilities” (Shohamy, 2001, p. 40). This endangers both the ethical status and usefulness of a test. As Hamp-Lyons (1997) cautions, those involved in the testing process at any stage must be accountable for the consequences of testing. It is, therefore, important to guard against the adoption of an attitude of distance and neutrality. If too much faith is placed in the CEEB to provide a ready-to-use test, then the test administrators would potentially be shrugging off their responsibilities not only to the test takers, but to the institution itself. This opportunity to maximize the benefits to as many parties as possible must not be lost.

5.2.4.2 Importance of Managing Perceptions of the Test

One of the most striking conclusions I came to as I gathered data was that the purpose and content of the language entrance/placement testing used at the institution was widely misunderstood by stakeholders. This may be in part due to the lack of transparency in the testing process. What is critical, however, is that because of this widespread misunderstanding about testing in general and the language entrance/placement testing process in particular, testing became a scapegoat.

For example, when instructors had students in their courses who do not have a sufficient level of English to participate successfully, testing was blamed. As mentioned earlier, there was one score user who assumed that there was no testing at all based on the insufficient English
skills of some of her students. In fact, as we have seen, it is possible that these students could have found ways around taking the test.

In other instances, when test takers were unsuccessful, or it was recommended that they take LPC courses at the institution, testing is seen as the problem. They see testing as just an attempt to make more money or to provide a justification for keeping people out. One potential participant expressed these ideas in this email (I’ve kept the writer’s all-caps style because it highlights the writer’s anger and frustration): “THIS ENGLISH EXAM WHICH [the institution] CONDUCTING IS LOOKS LIKE FRAUD.BECAUSE I WHEN I APPERARED FOR THE TEST I GOT VERY LOW SCORE REPORT WITH SUGGESTING TO TAKE SOME MORE ENGLISH COURSES IN [the institution] IN ORDER TO IMPROVE MYSELF. NEXT WEEK I HAVE SCORED 240 IN MY TOEFL. SO, IT IS RIDICULOUS A PERSON GETS O ON HIS EVALUATION ONE WEEK AFTER HE GETS A MORE THAN 75% IN TOEFL. QUITE INTERESTING>???. THIS SEEMS THAT THEY WANT TO DO SOME BUSSINESS FOR THE STUDY>???.” Test takers are not alone in this point of view; some score users felt that the scoring procedures (taking the lowest of the three scores) was evidence of this as well.

5.2.4.3 Setting Standards Through Testing

Of course, it is important that language entrance/placement testing keep people who have not yet acquired a sufficient level of English out until they have reached the level of language required to succeed in the institution’s programs. This is for the sake of both maintaining standards in the classroom and preventing people who have little chance of succeeding from wasting their money. At the same time, it is unethical to use testing as simply a gatekeeper. Shohamy (2001) often refers to the unethical use of language tests to dictate knowledge. However, this is generally in reference to liberal education institutions. Because they have a more limited mandate, technical education institutions such as the one in this study are, therefore, well within their rights to use tests to set standards and dictate knowledge which is largely based on industry requirements (Shohamy, 2001). However, to be ethical, we must examine those standards: Why are they set as they are? Are there other important measures that have been neglected?

The goal should not be to simply keep people out, but to assess whether or not someone has the skills to succeed. This is never easy to do and the best that is possible is to arrive at an estimate. As McNamara (2001) points out, “the relationship between performance and
competence in language testing remains obscure." Therefore, simply setting a cut score high enough, or making a test hard enough that only the very top test takers can pass is not only not ethical, but does not ensure that those most able to succeed at the institution are admitted.

Certainly, not every test taker who wants to attend the institution can be accepted. However, we must be confident that a broad enough range of language is tested and that a variety of measures that are as authentic as practically possible are used to help ensure the most qualified candidates (and not just those who are more testwise) are admitted. This requires re-examining a variety of aspects of the test in order to ensure that benefit maximization is achieved.

Making the marking/cut score criteria available and giving test takers more feedback, quantitative and qualitative, will be an important step in demystifying what the test means and what is it used for. These changes would likely defuse some of the anger and frustration felt among test takers. In addition, test takers should be offered the opportunity to anonymously give feedback on the testing experience, both at the time of the test, and after they have received their scores. The test administrators claimed little knowledge of the anger and frustration felt among test takers. However, the majority of the emails that I received from potential participants expressed these feelings quite openly. This fact speaks to the disconnect between these two groups and the lack of opportunity for dialogue.

### 5.3 Recommendations

These recommendations are not intended as criticisms of the existing test, nor are they intended to make the test ‘perfect’. As discussed in chapter two, developing a ‘good’ test, requires a careful balancing of all of the measurable test qualities; test developers must always make trade-offs between a variety of different criteria for test usefulness and ethics. However, I believe implementing these recommendations will not upset the balance but rather enhance the testing process’ ability to meet several of the criteria. At any rate, they are an attempt at the “thoughtful engagement” that Lynch (2001, p. 368) describes.

Many of these recommendations focus on making the testing process more transparent. As mentioned previously, one of the main conclusions I have drawn from this study is that the tests lack face validity and/or are misunderstood. Improving the transparency of the process would have a variety of positive effects: improving face validity, making testers more aware of
their role and position, helping stakeholders understand the complex role of testing, developing greater respect for the job that testing does.

5.3.1 Re-examine Construct Validity

As outlined in the findings, several issues with construct validity (essentially the extent to which the language ability being tested is relevant to the purpose of the test) became apparent through this research: the inclusion of irrelevant material (like texts on American history), the lack of listening and speaking sub-tests, and the fact that the construct has not been revisited, according to one test administrator, for sixty years to ensure that it is relevant to current realities. To deal with these issues, I would suggest that test administrators gather samples of reading and writing assignments from classes; survey the types of listening and speaking activities that go on in classes (not just Communication classes); and talk to stakeholders, including test takers and current students, about the language skills they see as being important. Indeed, Bachman and Palmer (1996) emphasize that this type of investigation is the first crucial step in test development.

One major issue with the construct definition is that it seems to have been arrived at rather intuitively rather than through a systematic examination of the different language functions required at these levels. Perhaps when the language entrance/placement tests in question were developed approximately sixty years ago, a careful examination did take place. However, accepting that this construct definition is still valid after so much time has passed seems problematic. Recently, a survey of communication skills among graduates of the institution was carried out “to provide information for designing future communication courses across the wide range of programs within [the institution], in the light of [the institution’s] evolving role and the changing ethnic mix of its student population” (Hamilton, 2005, p. 2). This information is equally relevant to the development and use of the institution’s language entrance and placement tests.

It is not uncommon among those who have a great deal of experience in testing to trust their intuition, but after such a long period of time, it is ethical to revisit the construct definition in more detail. Apparently this intuition is working to some degree since Test B was, according to one test administrator, shown to be the most reliable predictor of success at the institution. However, in order to maximize benefits to all stakeholders, and indeed to improve the face validity of the test, it is important that the construct definition be re-examined.
5.3.2 Consider the Changing Demographics in the Design Statement

Bachman and Palmer (1996) suggest that personal characteristics of the test takers should be included in the design statement. The design statement acts as a blueprint for the test and decisions about the types of tasks that will be included is based on this. In addition to outlining the personal characteristics of the test takers, it should outline the purpose of the test, the target language use, the task types, the construct definition, a plan for evaluating the test’s usefulness, and the resources available for developing and administering the test (Bachman & Palmer, 1996, p. 88).

When Test B first developed, the test was not intended for international students. The reality today is much different. Not only are there many international students who want to take classes at the institution, but there are many recent immigrants who have a great deal in common with those international students. If the test demands a certain degree of interactiveness, it must be conceivable that the test takers, including international students and recent immigrants, will have the kind of knowledge demanded. Otherwise, it might appear that the institution prefers not to admit students from these demographics. As Shohamy (2001) states, what often passes as raising educational standards through testing may be viewed as an attempt to restrict minority access. Using a test that was not designed for international students could certainly be seen as such an attempt.

If the intention is to test for a grade 12 Canadian education, it should be made more explicit. In the focus group meeting, we only came around to the conclusion that this was the purpose of the test when prompted by my questions about the kind of background knowledge test takers would need and what might be some areas of knowledge that could be considered construct irrelevant. While some argue that you can never test language without testing culture (Duran, 1989), if the intent of the test is to test social and cultural knowledge and test taking abilities, then it must be made clear to test takers.

While the test administrators feel that they are (and should be) testing knowledge beyond purely language knowledge, it seems that some score users do not believe this is being done well enough. Given that some score users found that their students did not have a strong understanding of everyday English (as opposed to textbook English), it would appear that the test covers either the wrong type of background knowledge or not enough of a variety. It may be justifiable to have a broad scope of knowledge as part of the construct definition, but again, as mentioned above, perhaps it is necessary to revisit exactly what types of knowledge are
important to a test takers ability to succeed in a program at the institution. This will make the test both more ethical and more useful.

It is true that, as mentioned above, some general cultural knowledge is beneficial to students in courses at the institution. However, some items tested do not have clear relevance to courses. For example, knowledge of specific aspects of American history, as in the example given by a test administrator about the reading passage on blacks in the South, would not likely contribute to success at this institution. Shohamy (2001) suggests that it can be considered a form of deception if the test taker cannot clearly see the activity they are taking part in as being related to the ability being tested.

When considering what kinds of background information we can expect test takers to have, we must take into account the rapidly changing demographics in Vancouver and at the institution that were noted previously. At the same time as the institution is courting international students, and a great number of its students are recent immigrants or generation 1.5 (immigrants who arrived in Canada as children and are, therefore, similar to both recent immigrants and second generation immigrants), there is still a tendency to put the onus on the NNS students to ‘catch up’ rather than making it part of their teaching practice to make their classes more accessible to students with different backgrounds.

As Cumming (1994) states, the majority populations and societal institutions have responsibilities to these minority groups, and they must act on those responsibilities. We can no longer assume homogeneity in the classroom. Changing our practices to recognize that the classroom today is not the same as it was 20 years ago (Hamilton, 2005) is an excellent way to maximize benefits: NNS students can benefit from enriched content, NS students benefit from increased cultural sensitivity and awareness (clear assets in a rapidly globalizing business world) which in turn benefit the business community that the institution serves. Indeed, Hamilton’s (2005) study is evidence of interest in making such changes. It is vital that the placement and entrance tests reflect this as well.

5.3.3 Educate Stakeholders About the Test

The purpose, constructs and marking criteria of the test should be clear to more than just the test administrators. It is important that test takers and test score users understand these as well. For example, an information session at the annual campus-wide event featuring workshops and presentations on a variety of topics, a pamphlet for new instructors, a pamphlet
sent with the application package for test takers or available on the website where they register for the test should be available. These would not only demonstrate how testing works and what is tested, but generate more of an atmosphere of openness around testing as opposed to the current atmosphere of secrecy.

I suspect that part of the lack of respect for the authority of the test and the lack of face validity that this study discovered stems from the lack of awareness among stakeholders about the content of the test. The types of questions on the test are kept secret with good intentions. According to one test administrator, “What we are trying to avoid, which I think is contrary to a lot of theory on this, I personally would not want students practicing taking the test because you can practice taking the test enough to pass the test, which is not indicative at all of how well they can integrate language skills. And if we’re talking about grade 12 equivalent, I mean, we've had this discussion before – it's that broad knowledge to apply not ‘what did you learn last night?’; have you isolated the language enough to be at a level that would be useful to you – not useful to just pass the test.” The intention is to minimize the practice effect and ensure that test takers do not receive inflated scores, and to protect the security of the tests since there are no alternate, equivalent forms of the tests. Test administrators also mentioned wanting to make sure that teaching preparation courses for these tests do not become a money-making activity for outside agencies.

Based on anecdotal evidence, I would say this secrecy is not working. Over my past few years at the institution, several students have told me that since they couldn’t practice, they took the test over and over until they had practically memorized the content. Their friends had done the same. Therefore, the theory that keeping the test secret would eliminate the practice effect is clearly mistaken. Those who can afford the time and money will take the test until they get the score they need. Also, the test administrators felt that it would be unethical to create a system in which it was likely for language schools or tutors to profit from the test takers’ desperation to gain admission to the institution. The unintended consequence of people taking the test repeatedly to get the practice they feel they need is no more ethical (arguably less ethical) however. It then appears to test takers that the institution is trying to make more money from them by leaving them what they see to be no other option than to pay to take the test over and over.
5.3.4 Provide Opportunities for Test Takers to Practice Test-Taking Skills

Related to the need to educate stakeholders is the need to be clear about the test-taking skills required. Brown (1996), Bachman and Palmer (1996), and Shohamy (2001) agree that it is important to make provisions for test takers to have prior orientation to the test. If a test taker does not have much information about the format of the test, he or she cannot hone his or her test taking skills. For someone whose language skills are sufficient but whose test taking skills are not, the tests results will not reflect the test takers language ability.

Testing language, as mentioned earlier, cannot be divorced from other influences on language knowledge and testing situations. Stakeholders, especially test administrators, must be aware of the multitude of factors that can influence test performance and of the fact that written tests like this are indirect measures of a test taker’s knowledge and ability. Most people would probably concede that test taking skills are indeed special skills in and of themselves. However, there is a general attitude that if a test taker doesn’t do well, it is because they lacked language proficiency, not because they lacked test taking skills. As Shohamy notes, it is not uncommon to hear things like, “‘The test demonstrated that you are a failure.’ or ‘The test showed that you did not study hard enough.’” (2001, p. 40).

Granted, test taking skills are important for academic success. However, if testing these skills is not part of the construct definition, and if test takers are unaware that this is being tested specifically, it is hardly ethical to test them. Test-taking skills are often, however inadvertently, being tested as well. In a test that is not highly authentic, such as the reading and grammar sections on the language entrance and placement tests at this institution, the testing of test-taking skills is increased. It is, therefore, unethical to not provide some background on the types of test-taking skills that will be covered. As one score user pointed out, “poor test takers/those whose experience is embedded in their workplace knowledge who can’t transfer it to multiple-choice grammar test” may receive scores that do not adequately reflect their language skills. It is possible that foreign-trained professionals, who may be unused to taking multiple-choice language tests (compared to younger test takers and language learners who are more accustomed to the format), may be unduly affected by this, especially because there is so little information about the test available to help test takers prepare.

5.3.5 Make the Purpose of the Test More Open and Transparent

Because of the secrecy around testing, many score users are unfamiliar with how the testing process works and what exactly is tested. As mentioned previously, this leads to testing
becoming the scapegoat for all the language problems that the score users encounter. If the stakeholders had a better idea of the construct being tested and the purposes of the test, instructors would probably be less likely to blame the test for every student who does not seem to meet language requirements. At the same time, they might be more critical of the test for not doing what it should be doing. Deans may be less likely to grant admission to those who don’t meet the test requirements because the value of the test would be clear. Departments may be less likely to make side deals with private language schools to develop ‘pipeline’ programs that allow students direct entry into a program at the institution; if these departments were aware of the institution’s language requirements as dictated by the test, they would be more likely to recognize that many of the private programs do not meet these requirements.

Increasing the transparency of the testing process would improve the face validity of the testing process among test takers. If they were more aware of why they were being tested and how they were being graded, they would likely feel that the test was more valid, even if their scores were not as high as their scores on other tests. Boyd and Davies (2002) in their call for ethical codes in language testing, refer to the adage that justice must not only be done, but be seen to be done. Test takers must not only be given appropriate scores, but they must understand and accept that the scores are appropriate.

5.3.6 Involve Stakeholders in the Testing Process

Beyond simply making information available to stakeholders, they should be involved in the testing process. This would also improve the transparency of the test. The more stakeholders are involved, the greater the transparency and face validity. This could be done by providing feedback forms or an ombudsperson, perhaps a student acting as a peer advisor, for test takers. These would be a cheap and easy ways to continually receive feedback, and would allow the test takers to feel more involved and, hopefully, less frustrated. While I recognize that many test takers may express negatives view of the test simply because their level of English is not sufficient to get the score they desire, their feedback may still provide useful observations that would help improve the test. Giving them the opportunity to provide feedback may also make them more willing to accept their scores.

Since testing is often a test taker’s first experience with an institution, making sure that it is a generally positive experience should be of some importance. Even if a test taker is unsuccessful the first time, the goal should be to give them a positive impression of the institution so they will want to try again later and so they will tell others of their positive
experience. One test taker said she would, under no circumstances, recommend that any of her friends take Test A or B, and that it would be better to go to another institution or gain entry to the institution another way.

To involve score users more, an email address could be given to them so they could ask for more information and give suggestions/offer opinions. They are, after all, very clearly impacted by the test. They deal with the students personally and know what kinds of language skills are necessary for success in classes and what kind of skills tend to be lacking (and, therefore, not likely tested sufficiently). Perhaps on a regular basis (every year or two) questionnaires could be sent to score users to gather more information. While I found it difficult to convene focus groups, perhaps a different approach to a focus group would work. Rather than asking for their opinions on the existing test, tell score users that their input is needed in order to develop an appropriate test that meets their needs as score users. This way, score users may feel that they are being given an opportunity to contribute to the test rather than feeling that they are being asked to criticize the test (as they may have felt in my research situation). Another option might be to run an online discussion with those who participate being entered into a random draw for a prize.

It is crucial for this not to be done simply for show. The information should be used, or at least considered, to improve the tests and address the issues that are most commonly raised. “Evaluation has a dialectical character... It’s essential that members of the evaluating organization deeply believe that they have as much to learn from educators directly linked to popular bases as those who study at popular bases” (Friere, 1985, pp. 23-25). Testers have to believe that they can learn from other stakeholders. They have to believe that testing is complex and that they do not hold all the answers.

5.3.7 Make the Criteria for Success on the Test Available to Test Takers

Once again, this is an area that can be more easily addressed by Accuplacer. There is a lot of preparation material for Accuplacer available online through the CEEB, but it would be preferable to have the information available through the institution as well. This does not mean that students will be, in effect, given the answers. As noted above, test developers and administrators expressed concern that giving out information on the test would mean that student would be studying for the test and narrowing their knowledge to fit the demands of the test. They feared the test takers would become too testwise, and the practice effect would cause
their scores to be inflated. These are indeed valid concerns. However, the current solution of revealing next to nothing about the test is hardly ideal.

While giving out information about the criteria may have the potential to cause test takers to take preparation courses or buy textbooks to prepare for the test, this is not necessarily a bad thing. Bachman and Palmer (1996) include impact on learning (in preparation for the test) in their matrix. If the test is a ‘good’ test, this type of preparation is not ‘cheating’ as the test administrators believed, but a valuable form of learning. This is also in line with Shohamy’s (2001) argument that good testing has a positive impact on teaching and learning. If the test covers a valuable scope of language knowledge, then preparing for the test does not mean ‘cramming’ to prepare to answer specific questions, but will instead improve the test taker’s language abilities in valuable ways.

5.3.8 Establish/Adopt a Code of Ethics (and Perhaps a Code of Practice) for Testing/Evaluation

As Boyd and Davies (2002) point out, Codes of Ethics are often used as simply good public relations. Organizations hide behind them, but do not change their practices. Instead, they merely invoke the Code when their practices are called into question. Of course, this isn’t useful. A Code of Ethics can ideally help “create a profession that has high standards, [make] its members conscious of their responsibilities... and [be] open to the public” (p. 312). Since testing is intended to maintain high standards of English language ability of students at the institution, it follows that the testing practice should also be held to high standards.

Finally, a Code of Ethics would make the purposes of the test and the responsibilities of the testers clear to the public. This would likely improve the face validity of the tests and make the processes more transparent to the public. Shohamy describes how testing is seen as a rarefied, secret “club”: people don’t understand how they work, but the language of science makes them appear to be trustworthy and authoritative (Shohamy, 2001). This veil of secrecy allows testers to maintain power and promote their own agendas. In a publicly funded institution, it is crucial that the public be allowed access to the purposes and processes of testing while, of course, at the same time, maintaining test security.

5.3.9 Provide More Extensive Feedback to Test Takers

To be more ethical, a feedback should help a test taker improve their skills and provide guidance for teaching (Shohamy, 2001). Test takers should not be left to wonder what they
need to do to improve their skills in order to pass the test and gain access to the institution: “If
most tests do not provide a test taker with diagnostic information about his or her performance,
there are no clear benefits for the test taker” (Shohamy, 2001, p. 141). Although diagnostic
testing takes more time and money than simple entrance or placement tests, it should be the goal
test developers work towards. This is especially true given that CBT will make the collection
and dissemination of diagnostic data easier. Simply telling test takers that their language skills
are insufficient is not enough. If the institution is interested in recruiting more students rather
than leaving potential students to go elsewhere when their efforts to gain admission are
frustrated, it is important that test takers feel there are clear actions they can take in order to gain
admission. Test administrators agreed that Accuplacer would allow them to give more
feedback. It was not clear if this feedback would simply consist of more numbers, or if it would
be in the form of specific advice about the areas of weakness found through testing. Hopefully,
it will be the latter.

5.3.10 Make the testing conditions more conducive to optimum performance

As a test developer/administrator noted, having the test at 6:00 on a weeknight is not
likely the best time to get the test takers’ best performance. Those who had to rush from work
are hungry, tired, often late and flustered because the building is not easy to find. Saturday
morning was suggested as a better time to run the test, but perhaps having a variety of times
available would be ideal. It was also suggested that test takers be provided with a map with the
location of the test clearly marked. These are simple, practical changes that would lead to
greater benefit maximization and, therefore, a more ethical and useful test.

5.4 Summary

The fact that language testing is a complex undertaking may seem obvious. However,
just how complex it is is difficult to accurately ascertain; this became progressively clearer to
me during this research project as more and more complexities were revealed. While I have
outlined here some of the aspects of testing that require attention in order for a test to be useful
and ethical, there are many more that have only been briefly touched on or alluded to. This
underscores the fact that language testing cannot be entered into lightly. This is especially true
of large-scale, high-stakes language testing, or indeed, any large-scale, high-stakes testing.
While achieving the optimum balance of the criteria for usefulness and ethics may not always be
possible, we must remain in a state of “constant scepticism” and “thoughtful engagement”
(Lynch, 2001, p. 368) about our practices. Otherwise, it will be very difficult to claim that our tests are useful or ethical.
References


Appendix A

Focus group questions – test takers

1. Was the placement test what you expected it to be like? Do you think it gave you the opportunity to really show your abilities in English?

2. Do you think the test was fair? In other words, do you think that everyone has an equal opportunity to do well on the test?

3. Were there any questions you couldn’t answer because you didn’t have the background knowledge (not the English language knowledge)?

4. What changes would you like to see to the test to make it more fair? to better show your abilities?

5. Did your score/performance on the test help you understand what areas of English you need to work on?

If you are taking (have taken) classes at the institution

6. Do you think that your score on the test showed your ability to keep up with language tasks required?

7. Do you think you were placed in the right level based on your test score?

8. Were the tasks that you had to do on the test similar to things that you might have to do in class?

9. Are there any other issues that we should have discussed today?

Focus group questions – program heads, deans, and instructors

1. What language skills do your students need in class? In the workplace?

2. What kinds of background knowledge (not language knowledge) do you expect your students to have?

3. What are some common areas of weakness you find in the language skills of your students?

4. Do you think that the current placement test adequately tests these skills?

5. How do you think testing could more adequately address these issues?

6. What do the tests do well?
7. Do you think there are any special needs that foreign-trained professionals have in your courses? Any typical areas of strength?

8. Do you think the current placement tests allow foreign-trained professionals to exhibit these strengths? Are weaknesses adequately apparent?

9. Another thing to consider (although harder to conceptualize) is who is NOT getting into the classes who perhaps should. Who may be unnecessarily excluded by the placement tests?

10. Ideally, tests should do more than just determine in or out— they should be able to give test takers meaningful information they can use to improve their skills, they should help administrators decide how to best place the students. How useful do you think the scores are to test takers? To other score users?

11. Are there any other issues that we should have discussed today?

**Focus group questions – test developers**

1. In developing the language placement tests, what are your goals? What, specifically, do you want to test?

2. How do you ensure that the test has some degree of construct validity?

3. What kind of background knowledge (not language knowledge) do you test?

4. Are you aware of any sources of bias? What measures do you take to eliminate sources of bias? What else can be done to eliminate possible sources of bias on the test?

5. How much information do test takers currently get about their scores? How can we make their scores more useful to them and the institution?

6. What changes would you, ideally, make to the placement tests?

7. More specifically, how do you think placement tests could better tap into the knowledge and skills of foreign-trained professionals?

8. How do you see technology (CBT) playing a role in placement testing at the institution?

9. Do you believe the tests meet the technical standards set out by the language testing community?

10. Are there any other issues that we should have discussed today?