An Examination of Reliability and Validity Claims of a Foreign Language Proficiency Test

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Walter J. Mircea-Pines
Master of Education
SUNY at Buffalo, 1996
Bachelor of Arts
Babes-Bolyai University, 1985

Director: Anthony E. Kelly, PhD, Professor
College of Education and Human Development

Spring Semester 2009
George Mason University
Fairfax, VA

DEDICATION

To Victor. Rest in Peace.

ACKNOWLEDGEMENTS

Eamonn, thank you for being there and for not allowing me to lose faith. Tolle, Kelly and The Secret came to the Joker’s rescue.

Victor, you exerted a similarly powerful influence, even if in a different vein. I owe you.

Dr. Dimitrov and Dr. Behrmann, thank you for your unwavering support and encouragement. I appreciate your guidance very much.

Mirela, thanks for loving me. I love you too.

Pia, thank you for trying to keep my days bright and yellow.

And, for all others who ever asked “how is the dissertation coming along?” please do take credit. You know who you are.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

AN EXAMINATION OF RELIABILITY AND VALIDITY CLAIMS OF A FOREIGN
LANGUAGE PROFICIENCY TEST

Walter J. Mircea-Pines, PhD

George Mason University, 2009

Dissertation Director: Anthony E. Kelly

This dissertation study examined the reliability and validity claims of a modified version of the Spanish Modern Language Association Foreign Language Proficiency Test for Teachers and Advanced Students administered at George Mason University (GMU). The study used the 1999 computerized GMU version that was administered to 277 test-takers via WebCT, (a web course management tool), and not the original 1961 pencil-and-paper edition. The GMU examination retained five sections from the test's original seven, namely Speaking, Listening, Reading, Writing, and Culture and Civilization. The original response formats, i.e. multiple-choice and true-false, were preserved for the most part. The Writing section was changed from the original error correction and text editing format to a fill-in-the-blanks format. The Culture and Civilization section was re-written due to obsolescence; however it retained the original answer format, to which a matching component was added.

The study's design was framed by the investigative steps proposed by Messick (1989) and Nitko and Brookhart (2007). An *ex post facto* approach was used, which mixed the confirmatory factor analysis and multiple regression analyses with descriptive interpretations. The quantitative analyses examined the impact of nationality, gender and average time spent on completing the tasks on the five subscales of the Spanish GMU MLA examinations. The analyses provided mixed results related to the predictors' ability to explain the subscale scores. The descriptive analyses revealed numerous validity flaws and gaps in the test construction. Such inadequacies overshadowed the positive aspects and eventually led to my questioning of the hypothesis that the judgments based on data from the Spanish GMU MLA examinations are reliable and valid. Recommendations for possible revisions were made, along with suggestions for future research avenues.

1. Introduction and Background

> Validity is an evolving property and validation is a continuing process (Messick, 1989)

All tests are designed to collect data to support an argument for certain policies or decisions. If the tests are used for different goals or purposes, the validity of that usage can and should be questioned. Additionally, the context, purpose, and goals for which a test is designed can change over time, sometimes radically. For that reason, data collected from older tests may not serve the goals, purposes, and decisions of current contexts.

Creators and administrators of tests, have the professional obligation of ensuring that tests are up-to-date, and that they measure, within an acceptable margin of error, what they claim to measure. Due primarily to the social dimension of testing, it is imperative to ensure that assessment instruments do not become the embodiment of inertia based on the simple fact that they are already in place, and that replacing them might entail additional labor. If such time-driven revisions do not occur, interpretations of results could be outdated or erroneous. If a pass or fail result can influence one's career or placement, then the test yielding such an influential result needs to be as accurate as possible. If inferences are to be made about the practical, "real life" applications of test results, whereby subsequent performance is predicted by a one-time testing sample, then those predictions themselves need to be evaluated and re-evaluated periodically, through the test validation process (McNamara, 2000, p. 10).

This dissertation examines the validity of judgments made, using data from a version of the Modern Language Association Foreign Language Proficiency Tests for Teachers and Advanced Students (MLA tests), which were first published in 1961. From this point onward, the designation "MLA tests" refers to the originally developed tests, whereas the designation "GMU MLA examinations" will be used for the 1999 version that was modified by members of the Spanish faculty from the Department of Modern and Classical Languages at George Mason University.

While we are surrounded by facilitating technological advances throughout the educational arena, the testing of foreign languages does not seem to have progressed at a similar rate. Even though the theoretical testing framework has developed through several stages, just as language acquisition theories have, often times with a skewed parallelism between the two, there are many factors that decelerate today's development and application of assessments. One could argue that there is no immediate need for an update in the testing arena, given that language itself has not undergone a fundamental change. However, what has changed is the understanding of how languages are acquired, and subsequently the methods of teaching them. It would thus be a natural expectancy that the testing of such acquisition reflect this progress. Fortunately, as Bachman and Cohen (1998) remarked, research that used to characterize the "sporadic" dialogue between the practitioners of the two fields, namely that of second language acquisition (SLA) and that of language testing (LT), has become more unified, allowing for better alignment of the theories.

So why is it, then, that we can still encounter tests that were developed under the influence of the structuralist linguistic view of the 1960s and 70s, as explained by Bachman and Cohen (1998), when today's accepted LT theories emphasize a communicative model based on contemporary SLA research? Given that SLA and LT are now in theoretical alignment, due primarily to the fact that the former renounced the discrete model proposed by Lado (1957) and LT followed suit, albeit more slowly, it is logical to ponder the absence of the view that language is multi-componential in some current tests.

Is it a question of the language testing theory being left behind once more, not being able to keep up with the acquisition theories? Or, is it that such "antiquated" methods of proficiency testing are still adequate today because they do not take into account when and how the candidates may have acquired their language ability? Such older assessments are still being used despite the stark contrast between viewing language, and subsequently the ability to use it, as consisting of distinct components (e.g. grammar, vocabulary) and skills (speaking, listening, reading, and writing), versus the view that those same components do not occur in isolation, and as such, should not be tested individually. It is a natural question to ask if the inferences that were drawn from use of the MLA tests in the past are still reliable and valid today, taking into consideration current certification goals, revised models of language acquisition and changes in demographics.

Background of the Problem

Perhaps the single most important educational assessment aspect remains the fact that testing theories must be anchored in current language acquisition theories, so that the means for data elicitation parallel the means of acquisition. Such a theoretically anchored testing model was advocated by Bachman and Palmer (1996), in contrast to the empirically based model envisioned by Chalhoub-Deville (1997). Thanks to such differences of opinion, the testing theory arena is making progress: hence the need to revisit such fundamental issues in subsequent pages.

Until recently, the field of LT has not experienced the tumultuous growth that SLA has known. The LT field generally lags behind that of SLA, in some instances by more than a decade. Due to circumstances such as the ones detailed below, one can still encounter tests that are not in tune with contemporary SLA theories.

According to Kunnan (2004), the study of language testing began in the 1930s, but did not get much attention until the *Language Testing* book by Lado and the "Fundamental Considerations in Testing for English Language Proficiency of Foreign Students" article by Carroll, appeared in 1961. Moreover, the first issue of the *Language Assessment Quarterly* journal did not appear until 2004, joining the twenty-year-old *Language Testing* journal: a short history when compared to that of an SLA counterpart like the *Modern Language Journal* with its ninety years of uninterrupted publication (Magnan, 2006). It must be noted here that the Modern Language Journal "has over the years published a respectable body of articles on testing" (Spolsky, 2000, p. 536). However, even when an important test put together by a reputable institution appeared and was administered to thousands, it was not mentioned in the literature: the Modern Language Aptitude Test appeared in 1957, but according to Spolsky (2000), it was not mentioned in the test-use literature until some ten years later.

As can be expected, the perspective has changed in both fields since Lado's (1961) proposed SLA framework, which attributed the learning of a foreign language to the transferring of habits from the native language into the language to be learned (p. 14). Lado saw the need to test the problematic transfer points, focusing on the difference of habits between the native and the non-native language, and to do so at discrete points. Today's reader, the one who is aware of the predominance of the communicative approach in both SLA and LT, is neither shocked by the fact that the second chapter of Lado's book is entitled "Testing the *Elements* [italics added] of Language", nor by the fact that he views validity as being the high correlation between two sets of scores.

Lado's discrete point theory continued to influence the testing arena, in parallel to that of Carroll (1961), which advocated an integrative approach that was meant to reveal the test takers' global proficiency. As Oller (1976) pointed out, Lado's approach was oddly enough based not on the contemporary theory that informed language teaching in his time, in turn influenced by the linguistic trend, but by the much earlier contrastive analysis of Bloomfield's (1933) behaviorist principles, thus creating a time gap of approximately thirty years. Furthermore, the creators of the Test of English as a Foreign Language (TOEFL) perpetuated the use of the discrete-point method promoted by Lado, even though they were not in agreement with all points, and even though Chomsky's generative grammar approach was taking over Bloomfield's structuralism in the early

1960s. Oller (1976) argued that the reason the TOEFL creators did not use the contrastive analysis was because of practical issues, as it would not have been feasible given the variety of the exam-takers' backgrounds. In the same article, Oller advocated the use of integrative tests, stating that "it seems that integrative tests are simply better windows through which to view language proficiency than are discrete-point tests" (p. 295).

Yet, in her 1968 book, Rivers tried to convince SLA practitioners that there is "a place for both points of view" (p. 78). She advocated the use of the audio-lingual method for the early stages of language acquisition, which consists of induction, drill, and analogy, followed by an emphasis on "real communication situations devised in the classroom, rather than in continual drills and exercises" (p.80). Therefore, when it comes time to test whether the objectives have been reached or not, and, as a first step in ensuring test validity, she proposes a combination of test items: those that will address individual skills, and those that will take the various language elements in interrelatedness. However, when constructing objective tests that should be considered statistically more reliable, Rivers chose to promote foremost the testing of linguistic skills in isolation.

As illustrated above, the time gaps between theory and its application can be significant. If learners were taught according to the audio-lingual method in the past, it follows that they should be tested with an instrument developed on the same principles, namely that of examining discrete language elements. The question of what happens when such an instrument is applied to learners that have been exposed mainly to communicative language acquisition methods is what actually triggered the need to further investigate the validity of judgments inferred from the MLA test results.

Statement of the Problem

The pencil-and-paper MLA tests were released after two years of cooperation between the Modern Language Association of America, the Educational Testing Service, and the United States Office of Education, Department of Health, Education and Welfare, under the provisions of the National Defense Education Act of 1958 (Starr, 1962, p. 31). Over the years, these tests have been administered in multiple contexts and for various purposes, seemingly under conditions that often ignored Bachman and Palmer's (1996) cautionary remarks on inappropriate test use, namely that tests should be used exclusively for what they were originally designed. To exemplify, the MLA tests are still used at various institutions in order to garner different types of information, as illustrated by the following partial list:

- The Graduate School of Education (2006) at George Mason University accepts the results of the MLA tests in lieu of thirty hours of coursework towards the foreign language endorsements requirements.
- The Department of Modern and Classical Languages (2006) at George Mason University administers the MLA tests for the Virginia State Department of Education, which accepts the results of these tests as proof of foreign language proficiency (1998).
- The Language Learning Center (1999) of the Department of Foreign Languages and Literatures at Old Dominion University is also a Virginia State

testing center for the MLA tests. In addition, MLA credits may be used to satisfy the Foreign Language Requirement.

- The Department of Languages and Literature (2006) at The University of Utah offers special credits for language ability based on MLA test results, with an equivalency of up to two years of foreign language study.
- The Department of Foreign Languages and Literatures (2005) at the State University of New York at Geneseo administers the MLA tests in French, German, Italian and Spanish, twice per year, and grants up to six credits, which allow students to enroll in 300-level courses.
- The Department of Modern Languages (2003) at Central Connecticut State University uses the MLA tests for placement purposes, determining eligibility to continue with higher level courses that would allow a student to consider opting for a major or minor.
- The College of Education and Human Development (2006) at the University of Minnesota requires native speakers of English to take an MLA test if they are seeking an additional licensure in world languages and cultures.
- The US Army Special Operations Command (2001) deems a candidate for the position of trainer/translator as being natively proficient by virtue of passing the French MLA test.
- The San Diego State University (2006) allows students to waive the foreign language requirement upon successful completion of the MLA test.

- The Truman State University (2006) administers the MLA test to all seniors majoring in Spanish.
- The Department of Languages (2006) at the Houston Baptist University set the minimum overall MLA Spanish test score at the 80th percentile for those students seeking to gain admission to the bilingual education program.
- The Department of Bilingual Education (2006) at the Texas A&M University at Kingsville acknowledges the demonstration of second language proficiency based on the MLA test results for those candidates who are seeking admission to the doctoral bilingual education program.

The above examples show that the MLA tests have a variety of uses, ranging from placement to proficiency demonstration. As such, the tests connote the strong social implication of determining a possible future path for the candidate. Consequently, the MLA test has great power, and therefore belongs in the high-stakes tests category. It is this very aspect that triggers the need for a closer look, especially when considering the fact that it has remained generally unchanged for more than 45 years now. If we consider that the MLA is a test grounded in 1960s principles of assessment based on discrete language elements and separate skills, and if we consider that the prevailing communicative competence model of language assessment views these component parts as interdependent, it is appropriate to investigate whether the MLA model remains reliable and valid today.

Research Questions

The overarching question asks how well and to what degree do the results of the Spanish GMU MLA examinations support the claim that candidates are language proficient and thus qualified for licensure. The following research questions, adapted from the model proposed by Nitko and Brookhart (2007), helped examine the criteria for test validity and reliability, and provided the framework for the study. The nomenclature appearing at the end of each research question belongs to Nitko and Brookhart.

Quantitative

1. Is there evidence that the Listening, Reading, and Writing factors represent the underlying structure of the proficiency examination? (*content and internal structure evidence*)
2. Are the results of the assessment internally consistent or reliable over time? (*reliability evidence*)
3. Is there evidence that the Speaking, Listening, Reading, Writing and Culture factors interact with the variables gender and nationality and with the average time taken to complete the test? Are there significant correlations among these five factors? (*external and generalization evidence*)
4. Are the results predictive of future performance or consistent with the results of similar testing (*external structure evidence*)?

Descriptive

1. Do the hypothesized factors represent constructs in more current theories for assessing language proficiency (*content evidence*)?
2. a. How relevant to the thinking skills and processes of demonstrating language proficiency are the tasks in the Spanish GMU MLA examinations (*substantive evidence*)?

   b. To what extent do any features of the assessment tasks detract from a candidate's ability to demonstrate these thinking and substantive processes?
3. Are there anticipated or unanticipated side effects from interpreting and using the Spanish GMU MLA examinations as tests of proficiency (*consequential evidence*)?
4. How do the factors of cost, efficiency, and practicality justify the use of the Spanish GMU MLA examinations (*practicality evidence*)?

Purpose of the Study

Personal reasons deserve to be mentioned first, even though, in Maxwell's (1996) words, they might "bear little relationship to the 'official' reasons for doing the study" (p. 15). Ever since I began administering the Spanish GMU MLA examinations in the fall of 1998, I have been puzzled by the fact that even though the examinations are old, they are still being offered. Furthermore, I found out that the Modern Language Association (i.e. the organization that was instrumental in the development of the tests) lifted copyrights and ceased involvement some twenty-five years ago because the examinations were deemed antiquated (N. Lusin, personal communication, July 14, 2006). However, after the MLA gave distribution rights to CBT McGraw-Hill at some point in the 1980s, the examinations continued to be offered. Throughout the history of administering this examination, I have seen and read frustration on the candidates' faces. While most of it may have very well been related to the normal stress associated with taking an important and infrequently offered examination, I supposed that the rest could be attributed to factors related to the tests themselves. This inference led to the desire to satisfy practical purposes, which in turn relate to research purposes. I was convinced that the practicality would be evidenced and satisfied, regardless of in which direction the findings were going to point to: tests are indeed obsolete, or, they are as strong today as they were when first created. The former case calls for the dissemination of investigative results to test administrators, pointing out the shortcomings; the latter would most likely lead to studies that focused on possible improvements. According to Pellegrino, Chudowsky and Glaser (2001), problems should be addressed "not by stepping up the amount of testing or abandoning assessments entirely, but rather by refashioning assessments to meet current and future needs for quality information" (p.25). Along similar lines, Spolsky (2000) offered the following argument:

> I believe … that testing is an important but potentially dangerous component of language teaching. It deserves better understanding than most language teachers have of it, and it demands more careful use than most testing experts seem ready to acknowledge. (p. 537)

Ultimately, the answers to the research questions provided the scientific support for both my practical and personal reasons. To my knowledge, no other study of a similar nature has been undertaken. However, there seem to be efforts to find alternative assessment tools, at least in the state of Virginia, as evidenced by the proceedings of the Department of Education (2004), which proposed to use the American Council on the Teaching of Foreign Languages (ACTFL) Oral Proficiency Interview (OPI) test in combination with the ACTFL Writing Proficiency Test (WPT) for demonstrating oral and writing proficiency. I must note here that, at the time of this writing, the OPI and WPT are not intended to replace or eliminate the MLA tests, but rather were provided as possible alternatives. The MLA tests still function in the Commonwealth of Virginia as a high-stakes test that represents a step in the process of granting or denying a teaching license for foreign languages in the K-12 public school system. The test candidates are either native speakers or speakers who have acquired the foreign language by non-academic means and need to prove foreign language proficiency.

Additionally, it must be mentioned that no further actions have been taken by the Virginia Department of Education either to assess validity in the case of the MLA tests, or to determine whether correlations exist between the MLA tests and the proposed combination of OPI with WPT. Such correlations would allow one to prove or disprove with a certain level of authority the hypothesis that the two testing models are equal in their ability to assess a candidate’s qualifications.

All the above-mentioned reasons encouraged my decision to set forth with this project. I hope to have shed additional light on the current situation of the Spanish GMU MLA examinations in particular, and on validity issues in testing in general.

Scope of the Study

This *ex post facto* study intended to determine the degree of reliability and validity of using the Spanish GMU MLA examinations for the purpose of demonstrating Spanish language proficiency. In the Commonwealth of Virginia, the results of these high-stakes tests function as one of the determining factors in the language teaching licensure process for those candidates who are native speakers or for those who have acquired the foreign language(s) by non-academic means. A general statement regarding the judgments of pass or fail was sought, based on results from the Spanish tests. The interpretation of these scores and the structure of examination questions were analyzed in order to validate test use.

Definition of Terms

High Stakes Test: A high stakes test is a test that may have serious consequences for the test taker, such as decisions regarding placement, admission, graduation, or promotion.

Pencil-and-Paper Test: Pencil-and-paper tests are traditionally used to assess discrete items, whereby the test items are presented in a fixed response format, e.g. multiple choice and/or fill-in-the-blanks.

Foreign Language Proficiency: Foreign language proficiency is a method of globally assessing the ability to speak, listen, read and write in a non-native language. Proficiency is defined by scales that usually range from elementary to native-like.

Proficiency Test: Proficiency tests are forms of assessment that aim to predict future language use, with little or no reference to the way the language was acquired.

Constructs: Constructs are variables that cannot be observed directly, such as proficiency, aptitude, intelligence.

Validity:

> Validity is the soundness of your interpretations and uses of students' assessment results. To validate your interpretations and uses of students' assessment results, you must combine evidence from a variety of sources that demonstrates [that] these interpretations and uses are appropriate. You must also demonstrate that students experience no serious negative consequences when results are used as you intend (Nitko & Brookhart, 2007, p. 38).

2. Literature Review

In an attempt to support the feasibility of the research design, the survey of the relevant literature is built around three topics: validity, language testing and test item crafting. Efforts were made to address the timeline that spans from the publication of the MLA tests to current practices. All topics will be expanded throughout the fourth and fifth chapters, given the descriptive component of this dissertation.

For the first topic, the investigation refers to the literature concerning the historical overview of general validity and reliability principles. The focus here is specifically on applications in the case of language testing.

The topic of language testing is explored next, from the pioneering MLA tests to the latest paradigms informed by current SLA theories and teaching methodologies. The framework of this portion of the investigation is based in part on the report by Alderson and Banerjee (2002). The survey of the literature targets the art of crafting test items next, tracing general principles and gradually focusing on specific applications for foreign language testing. The review of the literature concludes by raising issues that cannot be ignored when dealing with language testing in general, since they may define the topic more clearly. However, these aspects must be set aside for future endeavors, as they extend beyond the scope of this dissertation.

Reliability and Validity

The issues related to validity have been topics of debate since their acknowledgment was first noted in the literature, which, according to Anastasi (1986), occurred with the age-differentiation criterion developed by Binet and Simon in 1908. Anastasi further claims that it was not until 1954, the year in which the *Technical Recommendations for Psychological Tests and Diagnostic Techniques* were published through a joint effort between the American Psychological Association, the American Educational Research Association and the National Council on Measurements Used in Education, that "a major effort to introduce some order into the chaotic state of test construction procedures as a whole" (p. 2) was undertaken. The *Recommendations* (1954) identified four types of validity: content, predictive, concurrent, and construct validity. The authoring committee defined and recommended ways of investigating the four validity types as follows:

- Content validity represents the ability of test content to sample the subject matter and should be argued descriptively.
- Predictive validity refers to the anticipatory properties that are drawn from a test and should be checked through correlations with subsequent evidence.
- Concurrent validity is analogous to predictive validity, except that verification should occur through correlations with concomitant evidence.
- Construct validity reveals what psychological qualities are measured by a test. It should be evaluated by investigating the test's underlying theory for possible predictions involving score variation from person to person or occasion to

occasion, based on data. Observation, correlation and factor analysis were the recommended methods of investigation.

This taxonomy has continued to exist until the present, though predictive and concurrent validity have been merged into criterion validity, as proposed by Cronbach and Meehl (1955). Reliability, which is defined in part D of the *Recommendations* (1954), refers to several distinct coefficients:

- Internal consistency, which is obtainable through an analysis of variance or the split-half method.
- Equivalence, analyzable through correlations between results stemming from administering two forms of the test at almost the same time.
- Stability, testable like equivalence, through correlational analysis, but with a time gap between the administrations of the two forms of the test.

According to Messick (1989), the 1974 edition of the Recommendations (renamed as *Standards for Educational and Psychological Tests*), evidenced the beginnings of what he called “critical changes in the theoretical formulation of validity” (p. 18), along with the first recognition of the social consequences that can be triggered by the use of tests. Over time, validity began to mean more than multiple criterion validities. Test researchers made a shift towards an encompassing view, along with focal changes that shifted from predictions to actual explanations of facts that would allow such predictions. The 1985 *Standards for Educational and Psychological Testing* could be considered the passage phase into modern times insofar as validity theory is concerned. Not only did these *Standards* emphasize the concept of validity as being a single if multifaceted model, but they corrected the conceptually and practically inappropriate way of attributing validity to tests rather than to the inferences made from test scores. Furthermore, the three ingrained *types* of validity become *categories of evidence*, to be called content-related, criterion-related, and construct-related evidence of validity. The 1999 *Standards for Educational and Psychological Testing* continue the unified theory of validity and add the need to investigate the consequences of test use.

Messick's (1989) *Validity* chapter in *Educational Measurement* continues to be highly influential today. Nitko and Brookhart (2007) adapted Messick's unified validity concept, with the understanding that evidence from all relevant validity aspects must be collected prior to judging the interpretations and uses of candidates' results.

Not surprisingly, a parallel development of the concept of validity can be found in the case of foreign language testing. Chapelle (1999), who saw language testing researchers coming even closer to educational measurement literature, declared that "the definition of validity affects all language test users because accepted practices of test validation are critical to decisions about what constitutes a good language test for a particular situation" (p. 254). Lado's (1961) view that a test is valid if it measures what it is intended to measure is consistent with the measurement theories of his days. Lado's statement that "validity can be achieved and verified indirectly by correlating the scores on a test with those on another test" (p. 30) demonstrates his belief that validity is an attribute of the test. However, Lado, just like Nitko and Brookhart (2007) 46 years later, recognized that high reliability does not mean high validity, which points to the fact that historically, the understanding of test reliability has not changed. The measures and investigations of reliability mentioned in the 1954 *Recommendations* can still be found in today's studies.

Rivers' (1968) view of validity was also chronologically aligned. She found that if a test measures too many elements in interrelationship, it becomes impossible to validate that particular test. For instance, a test consisting of a dictation may not be a valid measure of listening skills if the language is phonetically uncomplicated. In 1979 Oller stated that correlational methods were the key to establishing validity, thus exemplifying the psychometric thinking of the time. Perhaps the most compelling argument that foreign language validity theory was approaching or even surpassing that of validity in general is the fact that Madsen (1983), Hughes (1989), and Canale (1989) introduced the language specific validity concepts of affect, washback, and ethics respectively, which have a direct correspondence with Messick's (1989) consequential aspects of validity.

The discussion about the definition and scope of validity advocated by Messick (1989) in educational measurement was fine tuned to language testing application by Bachman and Palmer in 1996. While recognizing the inclusion of multiple validity aspects under construct validity, they added the aspect of test usefulness: "the most important consideration in designing a language test is its usefulness, and this can be defined in terms of six test qualities: reliability, validity, authenticity, interactiveness, impact and practicality" (p. 38). Authenticity represents "the degree of correspondence between the characteristics of TLU [target language use] tasks and those of the test task" (p. 23). By "interactiveness" Bachman and Palmer understood a personal type of authenticity, one that activated the candidate's knowledge, metacognitive strategies, and schemata in order to accomplish test tasks. "Impact" referred to consequential validity (what might happen as a result of test use), which included the washback effect through which testing might affect candidates, teachers (e.g. teaching to the test), society, and education systems. "Practicality" referred to an accounting approach to the feasibility of test design, production, and administration. Current trends in validation research will continue to be addressed throughout the subsequent chapters, as validity has become an integral part of language testing.

Language Testing

As mentioned in the "General Statement of the Problem," perhaps the best explanation of why theory did not inform practice when the MLA tests were first released, is the fact that the two research fields were not in alignment at the time. However, now that there are cross research interests being pursued by specialists in both fields, it is worthwhile to investigate the current status of testing theory and practice, as informed by the latest developments in second language acquisition.

Since language tests have traditionally assessed language abilities in isolation, it seems appropriate to address the five skills (speaking, listening, reading, writing, and culture) separately. It must be noted that only four language skills are generally considered, namely the receptive and productive ones, i.e. listening with reading and speaking with writing, respectively. Although languages and the cultures of the respective countries go hand in hand, culture appears to be the most elusive component of the five when the time for assessment comes.

Speaking

It is not surprising that the MLA tests were among the first widespread language tests to address speaking proficiency (Kaulfers, 1965). Standardization implied the need for group testing, a feature that was not practical before the establishment of the language laboratories. In his seminal book published in 1961, Lado mentioned testing the *integrated skill* [italics added] of "Speaking a Foreign Language". Today's reader could easily perceive the integration as bearing the modern connotation of a communicative approach. However, what Lado meant at the time, was to bring together the elements of pronunciation, sound segments, stress, intonation, grammatical structure, and vocabulary in a testing format that would render the test-taker "not aware of what is being tested in each particular item" (p. 243). Rivers (1968) made it clear that an integrative approach to testing the ability to speak the foreign language is not desirable and gave the example of an interview where the test-taker is placed in a "communication situation ...[to see]…how he behaves" (p. 296). While Rivers agreed that this would constitute an assessment of speaking, she contended that it is more than that, because there are listening skills involved, along with other factors that would make it impossible to distinguish the individual skill of speaking. For that reason, she recommended an approach that "can be examined and evaluated apart from an act of communication, and therefore through tests which allow for a more objective assessment" (p. 297). A return to the direct testing of speaking was observed in the 1980s, due to the growing interest in communicative language teaching (Alderson, 2002, p. 92). The publication of the guidelines for assessing language proficiency by the American Council on the Teaching of Foreign Languages

(ACTFL) in 1986 was followed shortly by the release of the ACTFL Oral Proficiency Interview (OPI). History tending to repeat itself, there are those who criticize this popular communicative interview method of assessing speaking, for various reasons, which continue to be debated today.

Listening

Lado (1961) proposed pictures, true-false items, and multiple-choice items as auditory comprehension prompts, in the discrete item testing fashion. Even if the targeted response refers to the "integrated skill" of auditory comprehension, Lado argued that separate scores may be obtained for segmental phonemes, stress, intonation, grammatical structure, and lexical units (p. 221). In order to eliminate the danger of the test-taker knowing what is being tested, Lado proposed an arrangement by which items targeting the different elements appear in cycles of five. If the elements were assigned numbers, then intonation (3) would appear in questions 3, 8, 13, etc. Just as in the case of Speaking, Rivers (1968) proposed to isolate the skill of listening. If in the early stages of acquisition the candidates were presented with pictures or objects in lieu of written choices, they would not have to read. As they progress, they could be presented with multiple-choice answers. Rivers argued that the test becomes valid if skill mastery can be "clearly identified in the mode of answering and that credit can be given for this skill in isolation, quite apart from skill in other areas" (p. 296). The difficulty of isolating the listening skill has not been solved and is continues to be the focus of research today. Alderson (2002) commented that "it is well nigh impossible to construct a 'pure' test of listening that does not require the use of another language skill" (p. 87). History comes to the rescue again, albeit temporarily, in the form of the return to dictation as a tool for assessing listening. Coniam (1998) argued that a dictated text offers more contextual coherence and eliminates the need for additional answer formats. But the dictation method raises the question of what happens when the pronunciation of the language does not differ from its orthography, and so the debate continues.

Reading

According to Lado (1961), the best method to assess reading comprehension is through multiple-choice questions that address crucial problems in the chosen passages. Such problems could be inherent to typical language reading difficulties, or they could be related to the graphic representation of language (e. g. different alphabets). Rivers (1968) considered that the practice of using translation of passages was an invalid method of assessing reading skills because it implied that the test-taker had an ability to write in the native language. She also rejected as invalid the requirement of answering in the target language questions presented in the target language on a passage. Assuming that the test-takers have comprehended the passage and the questions, they would then have to use writing skills in order to answer the questions. Instead, she proposed the "more validly" choice of using multiple-choice questions written in the target language. Bernhardt (1993) considered multiple-choice procedures as being problematic because they may not be passage-dependent in the sense that the test-taker might be able to deduct the correct answer without reading the passage. This shortcoming has two possible causes: first, it could be due to the fact that test-takers rely on prior knowledge and on their schemata, and second, because test makers use the same words, both in the passage and in the multiple-choice questions. The latter is a constraint dictated by the consideration that the test-takers' vocabularies are not vast enough to allow for the use of synonyms. In Bernhardt's (1993) opinion, cloze procedures do not fare off much better, for several reasons. They lack face validity (does the test seem valid) since readers do not normally have a pencil handy to fill in gaps. They lack construct validity because the cloze test does not "accurately and adequately" reflect the process of reading, since the ability to insert the correct word in a gap and the ability to understand a contextual concept re totally different. Bernhardt also rejects direct content questions as being appropriate reading measures, since they might imply the answer, or worse, a shifting of understanding might occur, due to the question itself. Instead, Bernhardt offered the alternative of using the immediate recall protocol, which refers to a qualitative analysis of the test-takers' native language written recall of the target language passage. Alderson (2002) concluded that "it is essential to use more than one test method when attempting to measure a construct like reading comprehension" (p. 86).

Writing

Lado (1961) offered a sample test of writing (p. 255), along with suggestions on how to grade each of the three parts. The test would contain multiple-choice items that deal with spelling, punctuation, grammar, and vocabulary, a long passage with partial production type items that test sequence and transition, three pictures with directions to write a descriptive paragraph for each in order to test mechanics, and two compositions on assigned topics used to grade mechanics, style and content. Hamp-Lyons (1993) posited that a design such as the one proposed by Lado, where indirect testing of writing covers roughly half of the test, had completely disappeared from the field of writing assessment. Its place was taken exclusively by direct tests of writing, which have four components for which validity must be demonstrated: task, writer, scoring procedure, and reader (p. 73). The task's most divisive validation variable is that of topic. Writers, and implicitly their individualities, are often left out of the validation equation, even though they are the ones whose performance we measure. The challenge surrounding the scoring procedure is actually the readers' fault for not sharing the construct of writing quality. Furthermore, the reader as rater needs to be trained, and how that training is achieved will have a strong influence on the validity and reliability of the scoring of essays in a consistent and agreeable way. Alderson (2002) concured with Hamp-Lyons that the field of writing testing had moved from indirect testing to direct testing, thus eliminating error detection practices and the multiple-choice testing of grammar and cohesion. The shift was "encouraged by the communicative movement, resulting in writing tasks being increasingly realistic and communicative – of the sort that a test taker might be expected to do in real life" (p. 95).

Culture

In the realm of foreign languages, culture was understood as meaning the art, literature and civilization of a people and was taught and tested as such. Lado brought forth another dimension, namely that of "cross cultural understanding", which, according to him, had "not even been touched in testing" (1961, p. 276). In order to assess what people of a culture do and what it means, he proposes utilizing the multiple-choice format. Historically, variations and combinations of the two definitions have been accepted and they have appeared in the literature as an important element of language proficiency. However, as Schulz (2007) pointed out, the challenge of assessing cultural understanding continues today, mainly due to the fact that there are no clear guidelines for implementation and assessment, despite the recognition of its importance.

Test Item Crafting

When crafting multiple-choice items, Rivers (1968), advised the use of a "table of random numbers" that will allow random arrangement of correct choices. She did not recommend the use of true-false items, because students would be too tempted to take a guess at the answer, even if a third choice (True, False, Do not know) were to be added. However, Rivers identified an easy fix for eliminating the guesswork from matching exercises, by using unequal lists or by "providing several items which may be matched with more than one item, and others which do not match at all" (p. 313). In addition, she encouraged humor in test item creation along with item elimination if multiple administrations and subsequent revisions were to reveal such need.

As seen in the Language Testing section above, many criticisms are raised against the most common testing items, such as multiple-choice or true-false questions. However, as Nitko and Brookhart (2007) pointed out, if the test maker is careful in crafting the assessments, most problems can be eliminated with ease. In their conception, true-false items must not seek trivial information, must be clearly either true or false, and are not obviously one or the other. In addition, the authors offered an exhaustive checklist compiled from multiple sources (p. 142), which enables the test creator to reduce the number of correct answers based on guessing alone. For multiple-choice items, Nitko and Brookhart proposed adherence to five principles:

- Items need to address specific targets.
- The stem should be initially prepared in the form of a question.
- The correct alternative (keyed alternative) should be to the point and of a length that will not give it away as being the correct answer.
- The distractors should be probable.
- Items should be review before being released in test form, as the first iteration is always “unfit for human consumption”.

The checklist for reviewing the quality of multiple-choice items on page 166 is an additional tool for the test creator. In order to create quality matching exercises, Nitko and Brookhart suggested the following procedures: eliminate items that are not important or that do not fit the assessment plan, items on both sides of the list should belong to the same category, directions for test completion should be clear, the 10 elements or fewer contained in the response list should be plausible alternatives to every item in the premise list and should not be equal in number, the logically ordered longer statements should be in the premise list – the shorter in the response list, and finally, the premises should be numbered and the responses should be lettered.

Future Topics

One major topic that warrants further investigation is that of authenticity. The need to define it occurred together with the beginning of the communicative language testing approach (Alderson, 2002), which stipulated the need for authentic tasks if the predictions of the test-taker’s ability to function in the real world were to be valid. Bachman and Palmer (1996) placed test authenticity on the same level with validity and reliability.

Perhaps the most important and elucidatory aspect of language testing is the understanding of how the test tasks interact with the test takers and their characteristics. Solving this puzzle would offer clear guidelines to the challenge of creating feasible language tests.

Ethics, politics and testing standards should be part of the investigation, but their coverage is too broad for the task at hand. I would be remiss if I did not mention the issue of computerized assessment and the implications of validity when taking into account the medium of test administration.

3. Methodology

In order to answer the overarching research question of whether the judgments based on data from the Spanish GMU MLA examinations are reliable and valid, an *ex post facto* approach was used, which blended a quantitative approach with a descriptive component. For both the quantitative and the descriptive investigation, the test was examined using an adaptation of the criteria for test reliability and validity posited by Nitko and Brookhart (2007, p. 46). The main reason for employing a mixed design was my quest for objectivity. I am aware of how my own bias towards the Spanish GMU MLA examinations, due to having administered them for the past eight years, might have had an undue influence upon the findings. However, while one might argue that descriptive is synonymous with subjective, I am confident that having followed the model for developing test items as proposed by Nitko and Brookhart enabled me to remain impartial and allowed me to achieve statistical precision even within the descriptive framework.

It should be noted here that although all of Nitko and Brookhart's recommended procedures for establishing reliability and validity were considered, not all were followed. The reasons for this divergence, which stemmed mostly from data and follow-up issues, are described in the Limitations section and in subsequent chapters.

The research questions were:

Quantitative

1. Is there evidence that the Listening, Reading, and Writing factors represent the underlying structure of the proficiency examination? (*content and internal structure evidence*)
2. Are the results of the assessment internally consistent or reliable over time? (*reliability evidence*)
3. Is there evidence that the Speaking, Listening, Reading, Writing and Culture factors interact with the variables gender and nationality and with the average time taken to complete the test? Are there significant correlations among these five factors? (*external and generalization evidence*)
4. Are the results predictive of future performance or consistent with the results of similar testing (*external structure evidence*)?

Descriptive

1. Do the hypothesized factors represent constructs in more current theories for assessing language proficiency (*content evidence*)?
2. a. How relevant to the thinking skills and processes of demonstrating language proficiency are the tasks in the Spanish GMU MLA examinations (*substantive evidence*)?

   b. To what extent do any features of the assessment tasks detract from a candidate's ability to demonstrate these thinking and substantive processes?

3. Are there anticipated or unanticipated side effects from interpreting and using the Spanish GMU MLA examinations as tests of proficiency (*consequential evidence*)?
4. How do the factors of cost, efficiency, and practicality justify the use of the Spanish GMU MLA examinations (*practicality evidence*)?

Following is a description of the approach for answering each research question referring to reliability and validity.

- Content and internal structure evidence. In order to address whether there is evidence that the Listening, Reading and Writing factors represent the underlying structure of the Spanish GMU MLA proficiency examinations, I employed the framework of Structural Equation Modeling and investigated the factors via a Confirmatory Analysis Procedure, utilizing the computer program for latent trait analysis Mplus (Muthén & Muthén, 2003). In addition, I resorted to a descriptive approach to determine to what extent the analyses of the factors sample from constructs in more current theories for assessing language proficiency.
- Substantive evidence. In order to identify how relevant to the thinking skills and processes of demonstrating language proficiency are the tasks in the Spanish GMU MLA proficiency examinations, I compared these skills with the types of thinking elicited by current tests. In addition, I paid attention to whether candidates may be detracted from demonstrating these thinking and substantive processes by the features of the assessment.

- External structure evidence. Since I do not have data from similar assessments for this population, I had to restrict my comments to the fact that lacking such evidence constitutes a serious validity issue for the use of the Spanish GMU MLA examinations as proficiency tests.
- Reliability evidence. In order to assess the reliability of the Spanish GMU MLA proficiency examinations, Cronbach's α statistic was used to determine the degree to which the items intercorrelate within the test.
- Generalization evidence. In order to investigate whether there is evidence that the Speaking, Listening, Reading, Writing and Culture factors interact with the variables Gender and Nationality and with the Average Time taken to complete the test, I ran five multiple regressions, using the Statistical Package for the Social Sciences.
- Consequential evidence. I used a descriptive approach to address the question of the presence of anticipated or unanticipated side effects from interpreting and using the Spanish GMU MLA examinations as tests of proficiency, from the perspectives of both administrator and content expert. My main point referred to the pass/fail limit and its potential to affect test takers.
- Practicality evidence. A descriptive section that takes into account my experience as both the proctor and administrator for the Spanish GMU MLA examinations, from which I have been able to infer analyses of cost, efficiency, and practicality.

In addition to answering questions regarding the ability of the Spanish GMU MLA examinations to sample the criterion (Spanish language and Culture), the crafting of the test items was analyzed, based on the model proposed by Nitko and Brookhart (2007). The investigation addressed general principles for constructing assessments, such as the elicitation of only relevant language points while discouraging testwiseness, or the accuracy of directions and appropriateness of vocabulary. Then, the true-false, multiple-choice, and matching items in the Spanish listening, reading, writing and culture and civilization sections were discussed, taking into consideration both the advantages and the criticisms for each format. For instance:

- In the case of true-false items, while they are easy to write and score, candidates might randomly guess the correct answers, or, if the items are poorly constructed, they will only assess unimportant details/facts (e.g. a true-false item is poorly constructed when the question is a word-for-word reflection of the statement).
- For the multiple-choice items, investigations will be made to determine the clarity of the stem, the arrangement of the distractors and keyed alternative, as well as their degree of difficulty, considering the thinking level required by the targeted level of proficiency.
- Matching items will be screened for their construction (e.g. total number of items, exact number of premise and response elements, premises that give away the answer due to grammatical clues), and for the types of

thinking skills they elicit: rote memorization, classification or comprehension.

Setting

The available records show that the Spanish MLA test has been administered at George Mason University (GMU) by the Department of Modern and Classical Languages since 1990, initially in the audio laboratory. Beginning with 1999, the Spanish GMU MLA examinations were administered in the computer/teaching laboratory, since it offered optimal test-taking conditions, both from ergonomic and privacy standpoints. The computers were equipped with high-speed processors, high-resolution flat screen monitors and high-end audio components.

Participants

Two hundred seventy-seven subjects who registered to take the examination to demonstrate Spanish proficiency for foreign language teaching licensure purposes have been identified for this study from the archival data. The subjects vary in age and gender and include both native and non-native speakers of the language tested. The common feature of all candidates, be they native (n = 207) or non-native (n = 70) speakers, is that they do not have degrees in Spanish language and literature. Although all candidates have degrees, perhaps from universities where the teaching language is Spanish, they did not complete study programs that lead to certification in teaching Spanish. Most non-native speakers have acquired the language without formal academic credit.

Descriptive statistics were run for the variables in this study. Examination of the descriptive statistics indicates that, as expected, the native speakers readily outperformed non-native speakers with the exception of the dependent variable “Culture.” These data are presented in Table 1.

Table 1

Descriptive Statistics

Variables	Nationality	Gender	Mean	SD	n
Total listening score	non-native	Male	22.42	8.248	19
		Female	25.67	7.202	51
		Total	24.79	7.579	70
	native	Male	29.44	5.175	39
		Female	30.16	4.698	168
		Total	30.02	4.787	207
	Total	Male	27.14	7.097	58
		Female	29.11	5.695	219
		Total	28.70	6.055	277
Total speaking score	non-native	Male	71.11	15.989	19
		Female	73.39	14.319	51
		Total	72.77	14.708	70
	native	Male	100.44	3.885	39
		Female	99.42	6.428	168
		Total	99.61	6.037	207
	Total	Male	90.83	16.841	58
		Female	93.36	14.152	219
		Total	92.83	14.759	277
Total reading score	non-native	Male	24.16	10.813	19
		Female	24.37	7.489	51
		Total	24.31	8.435	70
	native	Male	38.56	6.863	39
		Female	37.92	6.204	168
		Total	38.04	6.321	207
	Total	Male	33.84	10.716	58
		Female	34.76	8.676	219
		Total	34.57	9.127	277
Total writing score	non-native	Male	26.89	13.055	19
		Female	27.98	10.654	51
		Total	27.69	11.267	70
	native	Male	43.44	9.805	39
		Female	44.87	8.184	168
		Total	44.60	8.506	207

	Total	Male	38.02	13.388	58
		Female	40.94	11.337	219
		Total	40.32	11.830	277
Total culture score	non-native	Male	44.47	7.633	19
		Female	37.49	8.967	51
		Total	39.39	9.124	70
	native	Male	42.69	8.715	39
		Female	37.61	8.961	168
		Total	38.57	9.115	207
	Total	Male	43.28	8.351	58
		Female	37.58	8.942	219
		Total	38.77	9.108	277

Out of the total 277 candidates, 51 were non-native females, 19 were non-native males, while 168 were native females and 39 were native males. Their mean raw Listening scores were 25.67 (s = 7.20), 22.42 (s = 8.25), 30.16 (s = 4.70) and 29.44 (s = 5.17) respectively. Their mean raw Speaking scores were 73.39 (s = 14.32), 71.11 (s = 15.99), 99.42 (s = 6.43) and 100.44 (s = 3.88) respectively. Their mean raw Reading scores were 24.37 (s = 7.49), 24.16 (s = 10.81), 37.92 (s = 6.20) and 38.56 (s = 6.86) respectively. Their mean raw Writing scores were 27.98 (s = 10.65), 26.89 (s = 13.05), 44.87 (s = 8.18) and 43.44 (s = 9.80) respectively. Their mean raw Culture scores were 37.49 (s = 8.97), 44.47 (s = 7.63), 37.61 (s = 8.96) and 42.69 (s = 8.71) respectively.

Instrumentation

While the MLA tests are referenced abundantly from their publication up until the 1980s, very little information remains available today. Searches of the *Mental Measurements Yearbook* (online version going back to 1989) and of the PsycINFO database yield no results. Given the age of the test, it is not surprising that no one still possesses the test manual. This is perhaps also explainable by the fact that the exams have been abandoned by their producers and distributors due to their antiquated nature, as revealed by the personal communication mentioned in Chapter One with MLA's assistant director for information services, Dr. N. Lusin. However, the print format of the sixth and seventh editions of the *Buros Mental Measurement Yearbook* reveals two reviews by Kaulfers (1965) and Probst (1972), respectively. Both reviews refer to Form B of the examination, the one dedicated to state and local certification programs, which is also the one that was the base for the Spanish GMU MLA examinations.

The original MLA tests were presented in a pencil-and-paper format and consisted of seven sections: speaking, listening, reading, writing, culture and civilization, professional preparation, and applied linguistics. The professional preparation and applied linguistics sections were never administered at George Mason University (GMU) for reasons that remain unknown. In 1999, several professors in the Spanish section of the Department of Modern and Classical Languages at GMU revised the test, in preparation for its online delivery via WebCT, a web course management tool. The Spanish version of the MLA test underwent the process of digitization first, since it drew more test-takers than the tests offered for other languages. Besides attempting to keep pace with the reality that computers had become an integral part of education, and implicitly part of assessment, the decision to produce and use a digital version of the test was prompted by reasons similar to those mentioned by Rocklin (1999), namely the aspects concerning timesaving, wide-distribution attributes, item banking, and ease-of-correction convenience. Also taken into account was the fact that if the current generation of test-takers still has a choice of delivery formats, future test-takers will undoubtedly not have the option of choosing tests in a pencil-and-paper format. Following, is a description of the Spanish GMU MLA examination as it exists today. The modifications that were made in 1999 are indicated in the appendix, whenever applicable.

The speaking test consists of two parts (see Appendix A for modification information). The first part presents the candidates with a printed text that they need to read and record. The second part consists of two separate situations presented in picture format. The candidates are asked to record their description of the illustrations and to make some comment about what the drawings might suggest. The recordings are stored digitally onto a dedicated server and onto the local hard drive as backup. The total time allocated for this section is ten minutes.

The listening comprehension skill is tested by means of 36 spoken questions. The first 20 remarks/questions have multiple-choice answers with three distractors and one keyed alternative (the construction of the multiple-choice questions is the same throughout the entire test). Items 21 through 28 are based on two spoken dialogues, followed by their respective spoken questions. Items 29 through 36 are based on one dialogue followed by eight spoken questions, with true-false format choices. The total allotted time is 25 minutes. Candidates can navigate between questions (items are clearly marked as answered or unanswered) and can change responses at will. This is a feature that remains consistent throughout the reading, writing and culture and civilization sections as well.

The fifty questions of the reading test are divided into three parts. The first fifteen consist of single sentences with blank spaces followed by multiple-choice answers. Items 16 through 45 consist of multiple-choice questions that refer to four passages which contain anywhere from 128 to 243 words each. The remaining five items ask candidates to read short passages and demonstrate their understanding via incomplete statements that offer multiple-choice answers for completion. The total time allocated for this section is 50 minutes.

The writing test consists of four fill-in-the-blanks passages (see Appendix A for modification information) in which numbered lines represent omitted words. Each passage consists of fifteen blanks with corresponding answer boxes in which candidates are asked to type their answers, each consisting of a single Spanish word. The default input language for the testing machines' keyboards is set on United States International, which allows for typing accent marks. The total allotted time for the writing section of the test is 60 minutes.

The culture and civilization test consists of two parts with 30 questions each (see Appendix A for modification information). The first part comprises 30 true-false items that refer to single statements, and the second part contains 30 matching pairs. This test is the only section that has a recommended study guide: Carlos Fuentes' *The Buried Mirror* (1992). The time allocated for this section is 60 minutes.

The total administration time for the Spanish GMU MLA examination is estimated at four hours, which includes the registration process, the preparation of the equipment (logging in, adjusting the headphones/microphones) and the breaks that are given at will.

Data Collection

All data come from archival sources at GMU, as obtained from the electronic administration of the five sections of the Spanish GMU MLA examinations (speaking, listening, reading, writing, culture and civilization). No new data was collected. The data consist of gender and origin demographics (native or non-native speaker) of test results from the five exam sections and of the duration of time spent on answering each question. The five sub-tests under investigation, namely speaking, listening, reading, writing and culture and civilization are scored with 0s and 1s, with a maximum possible of 105, 36, 50, 60 and 60, respectively. The time spent answering each question is calculated in seconds, based on the server recorded times of submission. The demographic data were collected at the time of exam registration, and were further proofed by a specialist in linguistics, who assessed the candidates' native versus non-native status based on individual speech samples. All data was collected with the approval of the George Mason University's Human Subjects Review Board and complete anonymity is insured for the 277 candidates whose scores were analyzed. All ancillary data describing the subjects was coded to ensure anonymity. No identifying codes, such as Social Security numbers, GMU G – numbers, or student identification numbers were used. All reports generated from the data set are described at group level and any individual case descriptions are completely anonymous.

Independent and Dependent Variables

The independent variables in this study are the gender and origin of candidates and the time they required for completing the tasks. The measured dependent variables are the test scores. Reliability statistics were run for the variables used in this study. Cronbach's alpha for the combined dependent variables minus the culture subscale was .87. When the culture subscale was added, the resulting Cronbach's alpha was .81, which is still within acceptable limits. Based on the success of these findings, the decision was reached to continue.

Data Analysis Procedures

The analyses conducted in this study were guided by the model proposed by Nitko and Brookhart (2007). The quantitative analyses employed two statistical computer programs: SPSS for Windows 16.0 for analyses of variance and Mplus (Muthén & Muthén, 2003) for identifying possible latent traits as part of a confirmatory factor analysis (CFA) of the five exam sections. Descriptive statistics were used for describing the sample in terms of measures of central tendency (mean, median, mode), and in terms of the measures of dispersion (minimum/maximum values, range, variance, standard deviation, and quartile ranges). In addition, the test was examined qualitatively, using the criteria for test validity and reliability. Several analytical procedures were used in order to address the construction of the multiple-choice, true-false and matching questions that appear throughout the listening, reading, culture and civilization sections, the construction of the cloze sentences for the writing section and the read-out-loud and describe-the-illustrations speaking section.

Limitations

Even though I do not believe that construct validity can ever be unquestionably demonstrated, I support Messick's (1989) proposition that concurrent validity is an integral part of such an attempted demonstration. Thus, I recognize that the central paucity of this study is the missing external structure check, i.e. the fact that I do not have other similar assessment instruments' results with which to compare the Spanish GMU MLA examination data for this population. Another limitation is the fact that even though the sample size is appropriate, the investigation is limited to the Spanish GMU MLA examinations, which limits possible claims to generalization evidence. Perhaps future studies will involve the addition of the French, German, Italian and Russian language and culture proficiency examinations.

4. Results

The results and interpretations presented in this chapter address the research questions that represent the framework for the study's design, data collection and analysis. The goal was to evaluate the reliability and validity evidence for the Spanish GMU MLA proficiency examination, using quantitative and descriptive methods and structured according to the model proposed by Nitko and Brookhart, 2007. While the quantitative questions addressed both the measurement part and the structural part, the descriptive questions sought to provide answers referring to content, substantive, consequential, and practicality evidence types.

Quantitative

1. Is there evidence that the Listening, Reading, and Writing factors represent the underlying structure of the proficiency examination? (*content and internal structure evidence*)
2. Are the results of the assessment internally consistent or reliable over time? (*reliability evidence*)
3. Is there evidence that the Speaking, Listening, Reading, Writing and Culture factors interact with the variables gender and nationality and with the average time taken to complete the test? Are there significant correlations among these five factors? (*external and generalization evidence*)

4. Are the results predictive of future performance or consistent with the results of similar testing (*external structure evidence*)?

Descriptive

To the extent that the quantitative analysis affirms that the five factors describe and explain performance on the test, the following questions will be examined descriptively:

1. Do the hypothesized factors represent constructs in more current theories for assessing language proficiency (*content evidence*)?
2. a. How relevant to the thinking skills and processes of demonstrating language proficiency are the tasks in the Spanish GMU MLA examinations (*substantive evidence*)?

   b. To what extent do any features of the assessment tasks detract from a candidate's ability to demonstrate these thinking and substantive processes?
3. Are there anticipated or unanticipated side effects from interpreting and using the Spanish GMU MLA examinations as tests of proficiency (*consequential evidence*)?
4. How do the factors of cost, efficiency, and practicality justify the use of the Spanish GMU MLA examinations (*practicality evidence*)?

Quantitative Results

In an effort to provide a better understanding of the scales, and possible enhancement of the construct validation, the following analyses were completed: factorial validation, investigation of the resulting scales' reliability, and the investigation of important factors' behavior on those scales.

Question 1: Is there evidence that Listening, Reading, and Writing represent factors that underlie the structure of the proficiency test?

The results related to the hypothesis that Listening (L), Reading (R) and Writing (W) represent underlying factors were addressed in the framework of Structural Equation Modeling (SEM). In order to validate the scales, three factors were investigated, namely Listening (L), Reading (R) and Writing (W), via a confirmatory analysis procedure. It is important to investigate both whether the hypothesized structure is the same for the whole population of candidates in general, and whether it is the same for females/males/native/non-native speakers in particular. The assumption was that the 36 items under the listening construct do indeed measure Listening comprehension, that the 50 items under the reading construct do indeed measure Reading comprehension, and that the four writing items do indeed measure Writing ability. The Speaking and Culture constructs were eliminated from the analysis due to the fact that they only have one score, whereas at least three items are needed for capturing.

The Listening factor consists of 36 items represented by aural questions, which have either written multiple-choice answers with three distractors and one keyed alternative, or a true-false choice. Each of items 1 through 20 is based on an individual, fifteen to twenty second long, spoken questions. The multiple-choice answers are provided in written format. Items 21 through 28 are based on two three-minute spoken dialogues, followed by their respective spoken questions. Items 29 through 36 are based on one three-minute spoken dialogue followed by eight spoken questions, with true-false answer choices.

The Reading factor consists of 50 items. The first 15 consist of single sentences with blank spaces followed by multiple-choice answers. Items 16 through 45 consist of multiple-choice answers that refer to four passages which contain from 128 to 243 words each. Each underlined word in a passage has four possible substitutions. The remaining five items ask candidates to read short passages and then demonstrate their understanding via incomplete statements that offer multiple-choice answers for completion.

The Writing factor consists of four passages with one word missing fill-in-the-blanks format. Each passage has fifteen blanks with corresponding answer boxes in which candidates may type their answers.

For illustration reasons, the hypothesized model for these three factors is provided in Table 2, with three example items for each factor, which are representative for the different formats. Please note that subsequent to the analysis, parts of the questions have been redacted by me for purposes of publication.

Table 2

Model of Three Hypothesized Constructs (Listening, Reading and Writing)

Construct	Items	
Listening		
	Items (Spanish - Original)	Items (English - Translated)
Example for items 1-20.	L1: ¿Mire usted la bandera ahí en el centro de la […] Crees que es de […]?	L1: Look at the flag there, in the center of the [...] Do you think that it is the […] one?
Example for items 21-28.	L21: ¿Cuál es el […] de esta […] con la señorita […]?	L21: What is the […] for this meeting with Ms. […]?
Example for items 29-36.	L29: Esta … tiene lugar […] de empezar la […].	L29: This […] takes place […] the beginning of the […].
Reading		
Example for items 1-15	R1: La […] de volar se . . . con rapidez y los […] serán tales que apenas se […] prever.	R1: The […] of flying is. . . quickly and the […] will be such that they […] hardly be expected.
Example for items 16-45	R16: Se ha […] de a. Ha sido inútil b. Ha habido que pensar en c. Han intentado d. Ha sido cuestión de	R16: It was […] to a. Has been futile b. One has had to think of c. They have tried d. It has been a matter of
Example for items 46-50	R46: El […] su a. soledad b. vejez c. ignorancia d. serenidad	R46: The […] his a. solitude b. old age c. ignorance d. serenity
Writing		
Passages 1, 2, 3, and 4 have the same format	W1: __1__ probable que las naciones de […] W2: __2__ el […] un primer paso importante _ W3: _3__ la creación de un vasto […], libre de restricciones […].	W1: __1__ probable that the […] nations W2: _2__ […] a first important step W3: __3__ the creation of a large […], free of […] restrictions.

The testing for the validity of the factors Listening, Reading and Writing was conducted through the use of a confirmatory factor analysis (CFA), using the computer program for latent trait analysis Mplus (Muthén & Muthén, 2003). The following four goodness-of-fit indices were used with the CFA in this analysis - *chi-square fit statistic*, (χ_2), c*omparative fit index*, *CFI*, *standardized root mean square residual*, *SRMR*, and *root mean square error of approximation*, *RMSEA*, with a 90 percent confidence interval. A relatively good fit is indicated with *CFI* > .90, SRMR <.08, and RMSEA < .06 as recommended by Browne and Cudeck (1993) and by Hu and Bentler (1999). Given the sensitivity of the chi-square statistic to sample size, its role in CFA testing for model data fit is more descriptive than inferential (Dimitrov, 2006).

The results presented in Table 3 show that CFI (.616) and the Tucker-Lewis Index TLI (.607) do not provide evidence for data fit of the model, while the most frequently recommended indices do provide such evidence: RMSEA = 0.56, with 90% CI = (.054; .058) and SRMR = .075.

Table 3[1]

Factor Loadings Estimates and Their Standard Errors: Confirmatory Model for Listening, Reading, and Writing

Item	Estimates	S.E.	Est./S.E.
Listening			
L1	1.000	0.000	0.000
.	.	.	.
L36	4.743	1.284	3.694
Reading			
R1	1.000	0.000	0.000
.	.	.	.
R50	1.129	0.249	4.530
Writing			
W1	1.000	0.000	0.000
.	.	.	.
W4	1.201	0.053	22.719

Therefore, the decision was made to consider the model acceptable and to proceed with the analysis involving the constructs of Listening, Reading and Writing, meaning that the hypothesized three-factor structure holds fairly well with the sample data.

Also, the factors Listening and Reading are practically not correlated, as indicated by the very low correlation among them (r = .005, p < .01). The confirmatory model is depicted in Figure 1.

[1] The complete table is presented in Appendix B

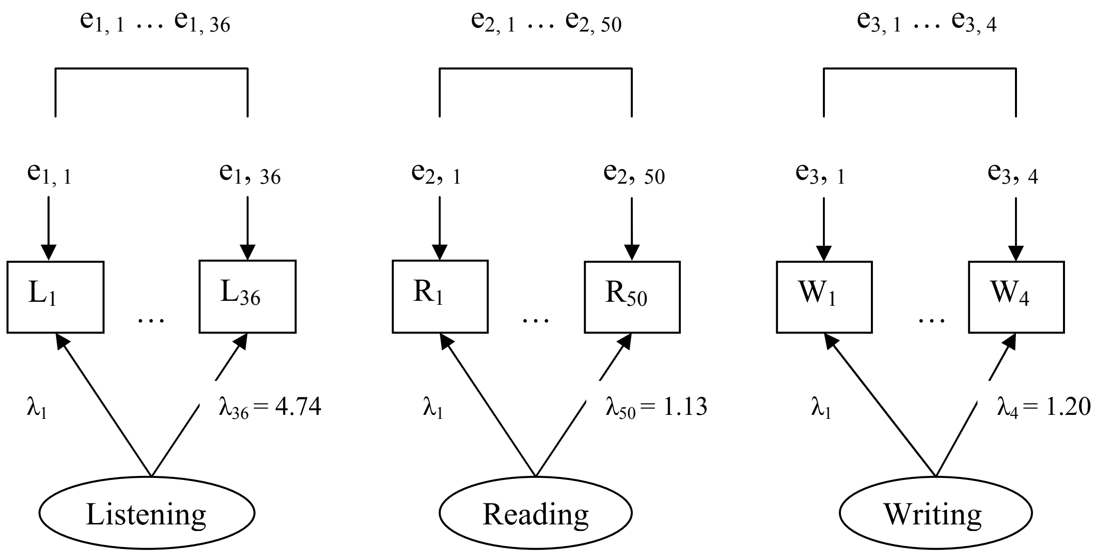

Figure 1: Confirmatory model – measurement section[2]

However, the statistically significant correlation between writing and reading ($r_{WR} = .295$, $p < .01$) deserves attention. There is evidence to support the hypothesis that a Listening factor is measured by the test, as all questions relate well to this factor, except for item L26, which is only slightly below the acceptable level of two for the critical ratio (CR = 1.558). All other items are greater than 2.0, indicating statistical significant relations (loading) of the factor to the respective item, at the .05 level. The factor Reading is also confirmed, with all questions relating well to the factor, except for item R16 (CR = 1.562). Even though the two items mentioned above were relatively weak, they were kept, for the integrity of the test. The items for the factor Writing are highly related, with critical ratios above 20. In terms of the three factors' variance (VAR), there is a relatively

[2] "e " stands for error unexplained, wherein $e_{1,1}$ to $e_{1,36}$ represent errors for listening items 1 through 36.

heterogeneous performance on Writing ($\text{VAR}_{\text{Writing}}$ = 6.413), but homogeneous on Listening ($\text{VAR}_{\text{Listening}}$ = .002) and Reading ($\text{VAR}_{\text{Reading}}$ = .016) as indicated by the small variances.

Question 2: Are the results of the assessment internally consistent or reliable over time?

There are no previous studies on the reliability of scores on the Spanish GMU MLA examinations; therefore, the estimation of the reliability of the Spanish GMU MLA subscales is an important task in this study. As stated in Chapter 3, the entire original culture subscale was considered obsolete by members of the Spanish faculty in the department of Modern and Classical Languages at George Mason University, mainly because keyed alternatives were no longer historically correct, or because the formulation was unacceptable. This prompted the subsequent re-writing of the subscale in 2000, by those same faculty members mentioned above (Dick Gerdes, personal communication, September 9, 2000). The updated Culture section consists of two parts with thirty questions each. The first part comprises 30 true-false items that refer to single statements, and the second part contains 30 matching pairs. The Cronbach's alpha for the combined dependent variables minus the culture subscale was .87. When the culture subscale was added, the resulting Cronbach's alpha was .81, which is still within acceptable limits, with the note that coefficients might be slightly inflated, given that some questions refer to the same prompts throughout the examination. Based on the success of these findings, the decision was reached to continue the analyses using all five subscales of the Spanish GMU MLA examinations. The individual results that indicate the degree to which the items intercorrelate within the test are presented in Table 4.

Table 4

Reliability

	Listening	Reading	Writing
Cronbach's α	.88	.91	.93

Even though the Speaking and Culture subscales were omitted due to the availability of a single item, they were taken into consideration for the five dependent variables model. The omission occurred because the scoring procedure only reports an aggregate score which cannot be decomposed into subscales or individual items. Even though the assessors have available to them the items and the subscales, they report only a single combined score. In this archival analysis, only the final composite scores were available. Reliability over time could not be established as no data was available.

Question 3: Is there evidence that the Speaking, Listening, Reading, Writing and Culture factors interact with the variables gender and nationality and with the average time taken to complete the test? Are there significant correlations among these five factors?

Five multiple regressions (Model 1-5) were run in order to determine the answers to question 3, using the Statistical Package for the Social Sciences (SPSS) for Windows, version 14.0. Table 5 presents the results of these analyses.

Table 5

Multiple Regression Results

Dependent Variable Predictors	*B*	*SE (B)*	β	*F*	df_1	df_2	*p*
Total listening score				19.61	3	273	.000
Nationality	4.73	0.78	.340**				
Gender	1.21	0.82	.082				
ALT	-0.24	0.81	-.163**				
Total speaking score				157.09	3	273	.000
Nationality	26.71	1.25	.788**				
Gender	-.121	1.33	-.003				
AST	-.016	.007	-.08**				
Total reading score				71.51	3	273	.000
Nationality	13.16	.99	.63**				
Gender	-.35	1.02	-.016				
ART	-.11	.05	-.110**				
Total writing score				59.20	3	273	.000
Nationality	16.47	1.31	.606**				
Gender	1.24	1.37	.043				
AWT	-.005	.003	.070				
Total culture score				11.81	3	273	.000
Nationality	.902	1.24	.043				
Gender	-5.31	1.28	-.238**				
ACT	-.32	.082	-.232**				

** $p < .01$.

Figure 2 depicts the SEM model for differences in Gender and Nationality on the constructs of Listening, Reading and Writing.

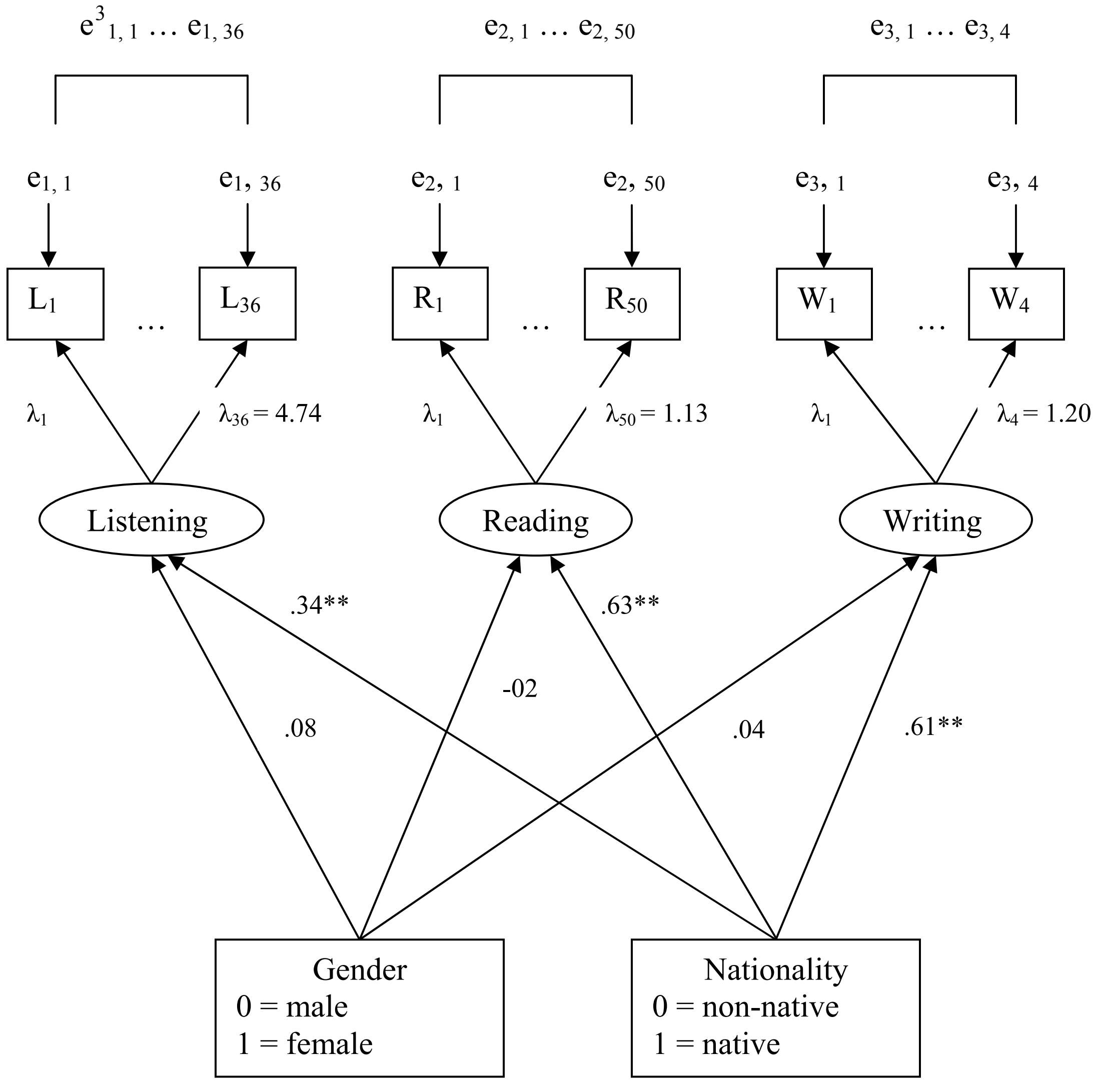

Figure 2: SEM for differences in Gender and Nationality on the constructs of Listening, Reading and Writing.

** $p < .01$.

[3] "e " stands for error unexplained, wherein $e_{1,1}$ to $e_{1,36}$ represent errors for listening items 1 through 36.

In Model 1, the Dependent Variable is the Total Listening Score (TLS), predicted from nationality, gender and the average time spent on the task. The prediction of the TLS from gender, nationality and time is statistically significant, *F* (3, 273) = 19.61, $p <$ .001. In addition, $R^2 = .177$ shows that 17.7% of the variance in the dependent variable Total Listening Score is explained by the variance in all three predictors. However, only two predictors were statistically significant: Nationality ($p < .001$) and Average Listening Time, or ALT ($p =.004$). Also, as indicated by the standardized coefficients (Beta), Nationality ($\beta = .34$, ALT ($\beta = -.163$) is relatively more important than ALT. Further, given the positive coding for nationality (0 = non-native, 1 = native) and the regression coefficient ($b = 4.73$), we can say that native candidates performed better than non-native candidates. Specifically, with everything else being equal, native candidates have a 4.726 higher predicted score than non-native candidates. In other words, if candidates were of the same gender and performed the same time-wise, the native speakers would have a higher predicted score than non-native candidates, by almost five points, noting that the range is from zero to a maximum of 36. The fact that the B coefficient for the Average Listening Time is negative (B = -.24), indicates the fact that the longer the time spent on answering the question, the lower the score.

In Model 2, the Dependent Variable is the Total Speaking Score (TSS), predicted from nationality, gender and the average time spent on the task. The prediction of the TSS from gender, nationality and time is statistically significant, *F* (3, 273) = 157.09, $p <$ .001. Also, $R^2 = .633$ shows that 63.3% of the variance in the dependent variable that is TSS is explained by the variance in all three predictors. However, only two predictors were statistically significant: Nationality (p < .001) and Average Speaking Time, or AST (p =.031). Also, as indicated by the standardized coefficients Nationality (β = .79), is relatively more important than AST (β = -.08). Further, given the positive coding for nationality (0 = non-native, 1 = native) and the regression coefficient (b = 26.71), we can say that native candidates performed better than non-native candidates. Specifically, with everything else being equal, native candidates have a 26.71 higher predicted score than non-native candidates. In other words, if candidates were of the same gender and performed the same time-wise, the native speakers would have a higher predicted score than non-native candidates, by almost twenty-seven points, noting that the range is from zero to a maximum of 105. The fact that the B coefficient for the Average Speaking Time is negative (B = -.016), indicates the fact that the longer the time spent on answering the question, the lower the score.

In Model 3, the Dependent Variable is the Total Reading Score (TRS), predicted from nationality, gender and the average time spent on the task. The prediction of the TRS from gender, nationality and time is statistically significant, F (3, 273) = 71.51, p =.000. In addition, R^2 = .44 shows that 44% of the variance in the dependent variable TRS is explained by the variance in all three predictors. However, only two predictors were statistically significant: Nationality (p < .001) and Average Reading Time, or ART (p =.02). Also, as indicated by the standardized coefficients, Nationality (β = .63) is relatively more important than ART (β = -.11). Further, given the positive coding for nationality (0 = non-native, 1 = native) and the regression coefficient (b = 13.16), we can say that native candidates performed better than non-native candidates. Specifically, with everything else being equal, native candidates have a 13.16 higher predicted score than non-native candidates. In other words, if candidates were of the same gender and performed the same time-wise, the native speakers would have a higher predicted score than non-native candidates, by more than thirteen points, noting that the range is from zero to a maximum of 50. The fact that the B coefficient for the Average Reading Time is negative (B = -.11), indicates the fact that the longer the time spent on answering the question, the lower the score.

In Model 4, the Dependent Variable is the Total Writing Score (TWS), predicted from nationality, gender and the average time spent on the task. The prediction of TWS from gender, nationality and time is statistically significant, *F* (3, 273) = 59.20, $p < .001$. In addition, $R^2 = .394$ shows that 39.4% of the variance in the dependent variable TWS is explained by the variance in all three predictors. However, in this instance, only one predictor was statistically significant, i.e., Nationality ($p <.001$). The Average Writing Time (AWT) was not statistically significant ($p =.147$). Further, given the positive coding for nationality (0 = non-native, 1 = native) and the regression coefficient ($b = 16.47$), we can say that native candidates performed better than non-native candidates. Specifically, with everything else being equal, native candidates have a 16.47 higher predicted score than non-native candidates, noting that the range is from zero to a maximum of 60.

In Model 5, the Dependent Variable is the Total Culture Score (TCS), predicted from nationality, gender and the average time spent on the task. The prediction of the TCS from gender, nationality and time is statistically significant, *F* (3, 273) = 11.81, $p < .001$. In addition, $R^2 = .115$ shows that 11.5% of the variance in the dependent variable

TCS is explained by the variance in all three predictors. Unlike the previous models, Nationality was not statistically significant (*p* =.467). The two other predictors were both statistically significant at the same level (*p* < .001). Also, as indicated by the standardized coefficients, Gender (β = -.238) and Average Culture Time, or ACT (β = -.232) were almost identical. Further, given the positive coding for nationality (0 = non-native, 1 = native) and the regression coefficient (*b* = .902), we can say that native candidates performed marginally better than non-native candidates. Specifically, with everything else being equal, native candidates have a .902 higher predicted score than non-native candidates. In other words, if candidates were of the same gender and performed the same time-wise, the native speakers would have a higher predicted score than non-native candidates, by almost one point, noting that the range is from zero to a maximum of 60. The fact that the B coefficient for ACT is negative (B = -.320), indicates the fact that the longer the time spent on answering the question, the lower the score.

The differences across models 1 through 5 are discussed in Chapter 5. The practical implications of these findings for foreign language proficiency testing are also presented therein.

Question 4: Are the results predictive of future performance or consistent with the results of similar testing?

In Chapter 3, I made a reference to the fact that although all of Nitko and Brookhart's (2007) recommended procedures were considered, not all were followed. Although some predictive data are available for the 1961 MLA tests, no data are available for the 1999 Spanish GMU MLA examinations, which represents a serious validity issue for the use of the examination as a proficiency test.

Descriptive Interpretations

Following are interpretations in which the author provides the expert based judgment called for in the methodology. The inferences drawn from these findings will be discussed in Chapter 5.

Question 1: Do the hypothesized factors represent constructs in more current theories for assessing language proficiency (content evidence)?

The following section analyzes how the factors of Listening, Reading and Writing compare to current assessments. The above quantitative analysis confirmed these factors for the Spanish GMU MLA examinations[4].

The American Council on the Teaching of Foreign Languages (ACTFL) represents today's authority in assessing language proficiency in the United States. They offer, through their testing arm, Language Testing International, two main examinations: one to determine oral proficiency, the Oral Proficiency Interview (OPI), and one for writing proficiency, the Writing Proficiency Test (WPT). The candidates are rated on a scale that is based on the Interagency Language Roundtable, which ranges from "superior" to "novice low", with eight levels in-between. The OPI is a standardized procedure for the global assessment of functional speaking ability, administered via a face-to-face or telephonic interview (ACTFL, 1999). The WPT is a standardized

[4] As mentioned earlier, the Speaking and Culture factors were not taken into consideration for this analysis.

procedure for the global assessment of functional writing ability, which “requires the examinee to read prompts in English and compose written responses in the target language without the aid of dictionaries or grammar references” (ACTFL, 2001). In addition, the document mentions that the remaining two skills (i.e. Listening and Reading) are next in line to be “reworked.” To date, such a revision process has not been undertaken, or, if currently underway, has not yet appeared in print.

In brief, the ACTFL instruments omit two major skills: Listening and Culture. However, given the fact that the OPI is a dialogue, one can assume that the interviewee must also employ listening skills in addition to speaking ones. In a similar vein, the writing process necessarily involves some reading comprehension skills, and thus the two missing skills are implicitly present. Indeed, the quantitative analysis in the current study demonstrates that there is a strong correlation between Reading and Writing factors.

The Culture factor is only mentioned in the OPI document to describe several proficiency levels that address the writers’ familiarity with the target culture. In the WPT document, Culture is mentioned once, under the rubric “Superior” level, for candidates who employ “developmental principles such as cause and effect, comparison, chronology, or other orderings appropriate to the target language culture” (p. 3).

Pearson, Fonseca-Greber and Foell (2006) stated that the OPI and WPT are used as tests to measure the language proficiency of foreign language teacher candidates, which is precisely the target population for which the Spanish GMU MLA examinations were used. However, it must be noted that the two tests differ in more than one respect. The fundamental difference consists of the theoretical foundations that lead to contrasting ways of assessment, namely task-based versus discrete skills. According to Mislevy, Steinberg and Almond (2002), interest in task-based assessment "can be attributed to such factors as the alignment of task-based assessment with task-based instruction, positive 'washback' effects of assessment practices on instruction, and the limitations of discrete-skills assessments" (p. 477) . In this sense, the OPI is a conversation that gravitates around the candidates' interests and experience, while the Speaking GMU MLA examination is based on a series of visual prompts.

Prior to the updated version of the MLA test, the Listening section had an additional component, namely the recorded repetition of sentences produced by a native speaker. This component was deemed weak since it was not considered to be a strong indicator of oral proficiency because of the relative ease with which one can imitate a short phrase uttered by a native speaker. Consequently, the component was eliminated by members of the Spanish faculty in the Department of Modern and Classical Languages at George Mason University.

Bachman, Lynch and Mason (1995) called for the need for authenticity in their conceptualization of the Language Ability Assessment System, which involved a role-play speaking task. Neither GMU version, pre- or post- update, takes authenticity into consideration (as defined by Bachman, Lynch and Mason). Furthermore, Long and Norris (2000) critiqued the discrete point assessment, which is the method of choice throughout the Spanish GMU MLA examination, by stating that "assessment associated with conventional linguistic syllabuses typically ask examinees to demonstrate knowledge about, rather than actual use of, the L2 [second language]" (p. 600).

Question 2a: How relevant to the thinking skills and processes of demonstrating language proficiency are the tasks in the Spanish GMU MLA examinations?

According to Storey (1997) task design is crucial for collecting evidence of thinking processes:

> What is more crucially important is the manner in which test tasks are performed. Attention to test-taking process is especially crucial and problematic in the testing of cognitive processing skills. The sample will only be representative if the cognitive processes involved in task performance are representative of those involved in performance of the domain of tasks in the real world. Thus the validation of tests involving cognitive processes will not be complete unless it includes some examination of the processes by which solutions to test tasks are reached. (p. 214)

The thinking skills required to complete the Spanish GMU MLA examinations appear to be different from those that would be required for completing a more current test. Since Spanish GMU MLA respondents were not asked to produce think-aloud protocols, it is not possible to describe which types of cognition they employed. Nevertheless, the theoretical difference in thinking skills is supported by the previously mentioned fundamental difference of task-based versus general discrete skills.
Unless thinking skills are explicitly measured, a test-taker's ability to answer questions correctly may be masked by guessing or other test-taking strategies.

Thinking skills and the Spanish GMU MLA Examinations

Speaking. In its initial 1961 version, this section of the examination matched Shohamy's (1994) definition of a "precommunicative" test. The test was administered in a language laboratory, where candidates are asked to repeat words and sentences. It is difficult to conjecture what "critical" thinking skills would have been employed in such instances. Upon elimination of that section, the current Spanish GMU MLA examination format resembles a "semi-direct" oral test, where test-takers respond to visual tasks. Another example of such a test is the Semi-direct Oral Proficiency Interview Test (SOPI), which, according to Shohamy, was developed as an alternative to the OPI, in the absence of a trained rater/interviewer.

The Speaking section of the Spanish GMU MLA examination consists of three parts in which the candidates are asked: 1) to read out loud and record an eight-line dialogue; 2) create a narrative based on three sequential drawings; and 3) assume the role of a person seeking a favor, also based on an image. While the first part does not require thinking skills attributable to speaking, the second and third parts are more involved: describing, narrating in the present and future, expressing opinion, hypothesizing and apologizing, asking for permission, expressing courtesy and thanks -- all functions that require complex thinking skills. These skills translate into producing language that varies at the linguistic, lexical and communicative level of complexity, and as such, seem to parallel or even match the skills required for the completion of the OPI and SOPI tests, as described by Shohamy (1994).

Listening. Alderson and Banerjee (2002) stated that relevant features of Listening are difficult to describe, because it “is essentially an invisible cognitive operation” (p. 87). In this section of the Spanish GMU MLA examination, the candidates listen to oral statements or dialogues that vary in length anywhere from seven seconds to three minutes and then choose answers from 28 written multiple-choices and eight true-false written statements. Such a format would appear to require more than just listening skills, since the candidates have to read the distractors and keyed alternatives. Yet, as Alderson and Banerjee pointed out, constructing a “pure” test of listening may prove impossible.

Although far from complete, the understanding of how Listening is processed has changed drastically over time, as first described by Buck (1994). In the 1950s, Listening was considered to be a “bottom-up”, sequential cognitive process that pieced together the various elements of language in order to construct meaning. Today’s understanding is that Listening represents a multidimensional cognitive process that depends on a variety of factors, as Buck (2001) pointed out: “the cognitive environment […] is the context which includes all the others, and that is the one that has the strongest influence on comprehension.” (p. 21) Despite the theoretical understanding that language comprehension is a “top-down” process, traditional listening tests that use multiple-choice or true-false response formats tend to dominate the testing of this skill. Such a process may be influenced by nonlinguistic factors, such as background knowledge. In other words, the language acquisition theory may have advanced, but not its application. In short, it appears that the thinking skills and processes involved in completing the

Listening section of the Spanish GMU MLA examination match those needed to demonstrate Spanish language proficiency as described by Alderson and Banerjee (2002).

Reading. Not much has changed in the assessment of Reading since Lado (1961) promoted and encouraged the use of the multiple-choice response format. Anderson, Bachman, Perkins and Cohen (1991) claim that although criticized, “the most common of these testing methods used in standardized reading tests is the multiple-choice format” (p. 42), which the Spanish GMU MLA examination also utilizes. Although the section is divided into three parts, with slight variations in what is being targeted (e.g. single words in independent sentences, phrases in paragraphs, or comprehension interpretations based on short passages), the response format remains multiple-choice with three distractors and one keyed alternative throughout. Cohen and Upton (2007) separated the types of thinking skills and processes that occur during test-taking into actual reading ones and test-taking ones. Only recently has there been an impetus in the assessment field to account for and incorporate test-strategies into tests and not solely testing Reading by itself. Unless this is achieved, what is actually being tested remains unclear: Is it the strategy or the comprehension? If we could understand the full complexity of what Bernhardt defines as an “intrapersonal problem-solving task” (1991, p. 6), the Spanish GMU MLA Reading task would seemingly emulate thinking skills associated with Reading proficiency.

Writing. In contrast to Reading, theoretical and practical approaches to Writing have changed fundamentally since Lado (1961). Gone is the multiple-choice testing approach that tested spelling, punctuation, grammar, vocabulary and error-detection skills. Alderson and Banerjee (2002) offered the following explanation as to why the approach existed in the first place: “one answer to the problem of the subjectivity of essay marking was to seek to test the ability by indirect means” (p. 95). Today’s methods are characterized by open-ended tasks, as encountered in the ACTFL Writing Proficiency Test (2001). The WPT contains four separate prompts for writing, each with three or more tasks that gradually increase in complexity, which are then double rated by trained specialists. While the need for objective marking may still be an issue, what changed is the perception that writing is primarily about text discourse rather than exclusively about grammar. In the framework of Bloom’s taxonomy, or better yet, in that of the revised taxonomy (Krathwohl, 2002), open-ended tests encourage the test-taker to utilize integrative skills that require critical thinking. In contrast, a cloze paragraph mainly asks about grammar (Farhady & Keramati, 1996) and thus employs one-dimensional, lower order thinking skills. Unfortunately, the Spanish GMU MLA Writing section follows the latter approach. The section consists of four fill-in-the-blanks paragraphs, which are then scored based on the appropriateness in meaning and form of the inserted words. Fill-in-the-blanks exercises are more appropriate for puzzles or vocabulary exercises. Since this fundamental difference exists, no comparison is possible between thinking skills employed in today’s approaches to testing writing and those present in the Spanish GMU MLA Writing examination.

Culture. According to Quinn Allen, “portfolios have been used in foreign language instruction to assess both culture and language learning” (2004, p. 233). While the portfolio approach represents the preferred way to assess culture and civilization, it cannot be applied to a standardized test without complicating its format of delivery and ability to employ simple objective scoring algorithms. Even though ACTFL (2006) emphasizes Culture as one of the five components of foreign language learning in the 21st century, the only standardized Spanish culture test in use today is the PRAXIS II (ETS, 2008). Both the PRAXIS and the Spanish GMU MLA Culture examination use the multiple-choice format, which is dictated by the convenience of simple scoring rather than by current opinions of how culture should be assessed. Schulz captured the missing clear directions of teaching and assessing culture:

> Despite a vast body of literature devoted to the teaching of culture, however, there is no agreement on how culture can or should be defined operationally in the context of foreign language learning in terms of concrete instructional objectives, and there is even less consensus on whether or how it should be formally assessed. (2007, p. 10)

Taking into account this scarcity and the fact that the PRAXIS is the only other standardized test that assesses Culture, it would appear that the thinking skills and processes employed by the GMU MLA test-takers are relevant to demonstrating Spanish language proficiency. However, when considering the trends of non-standardized assessments for the Culture component, via portfolios, or even blogs (Elola & Oskoz, 2008), the sole use of the multiple-choice approach may not be justified (this issue is discussed further below).

Question 2b: To what extent do any features of the assessment tasks detract from a candidate's ability to demonstrate these thinking and substantive processes?

The ability to assess the extent to which candidates are negatively influenced, or in some cases bolstered, to exhibit Spanish proficiency supporting thinking and substantive processes depends on two interconnected features of the GMU MLA examination: mode of administration and actual content. In order to capture the most relevant aspects of both features, the findings are presented in the order of the sections, with general comments first, as they apply to all sections.

General

Difficulties associated with the computer interface. As mentioned in previous chapters, the mode of administration was via WebCT, an internet web management tool. Throughout the history of administering this exam, I have witnessed the frustration of the candidates, which may have been related to many factors; for example, the "normal" stress associated with taking an important and infrequently offered exam. I suspect that some of the frustration is related to having to manipulate a mouse in order to demonstrate language proficiency. I have seen panic when digital microphones would not work, when networks were slow and prevented instant delivery of the next question or the submission of results before the internal clock of the machine lapsed. I have seen frustration when having to waste precious time scrolling up and down either because the screen was too small or the text too large, when Java generated pop-up windows blocked the screen and had to either be closed repeatedly, or read once and disabled. Since traditional tests are timed as well, comparisons with traditional pencil-and-paper testing formats are unavoidable. My inference is that the issue of insufficient time can only be accepted as an impediment insofar as allotted time is unduly spent on computer-related manipulations. Exasperation was also common when participants would "mistakenly" click on a different window and were led to believe they had lost the entire exam, or even when they had to input diacritic marks and were unfamiliar with the international keyboard setup of the testing center. To account for this it would have been ideal to test/survey the computer proficiency of the candidates in order to have an unbiased picture of their language proficiency levels.

Use of English. Another general limitation may include Spanish native speakers' deficiency with directions written or spoken in English, a circumstance which would potentially hinder their performance in the completion of tasks. After all, the test claims to investigate Spanish proficiency and not English.

Eurocentric approach. The GMU MLA examination appears to have a Eurocentric (Peninsular Spain) point of view with respect to cultural references and thus does not reflect the variety of Hispanic cultures and lived experience. This factor points to the issue identified by Messick's (1989) and Nitko and Brookhart's (2007) of "unintended consequence." The subgroup of test-takers who are not from Spain might be biased against.

Spelling errors. Further hindering a candidate's performance are spelling mistakes (espanñoles, Alphoonso) which occur both in text and in the English test directions. This flaw is attributed to the process of digitization and should have been corrected prior to the Spanish GMU MLA examinations' release.

Practice tests. Finally, in contrast to most standardized tests (SAT, GRE, MCAT, LSAT), no practice test was made available to the test-takers. This issue was mentioned repeatedly by candidates who were interested in taking the test, especially upon being informed that the pencil-and-paper version was no longer an option.

Speaking

Raters, following the original scoring manual, were directed to assign a numerical value to discreet linguistic, semantic and grammatical components (vocabulary, pronunciation, structure). Although the test-taker may have read and pronounced a grammatically and linguistically correct phrase, such performance does not necessarily speak to the candidate's ability to communicate in a real-life context. In this way, the Speaking test does not represent an "authentic or communicative language assessment" (Bachman, 2000, p. 24).

Unclear directions. For example, the test directions for the Speaking section of the Spanish GMU MLA examination are: “provide your response aloud at normal speed with the expression appropriate to the style and subject matter.” At best, the test-taker might conjecture a subjective interpretation as to what “normal speed” or “appropriate expression” could mean, which may possibly differ from that of the rater.

Choice of pictures. The pictures chosen to elicit speaking for the second and third part of the Speaking section have no relevance or reflection to real-life situations appropriate to teacher candidates’ experience. Again, since the Speaking part of the Spanish GMU MLA examination is still in use, these pictures are not described in detail here. The first picture depicts a scene that does not represent a plausible scenario in the life of a teacher. The 1961 test manual asked the rater to identify primarily pronunciation and grammatical mistakes, along with use of inappropriate vocabulary. Since the manual has not been revised, the rater might subtract valuable points simply because the test-taker is possibly suffering a lapse of memory and cannot recall a specific term called for by the prompt, regardless of the fact that such a term has little relevance in the task of teaching Spanish. The second picture, while depicting a more credible scenario, may have an inherent male bias towards a female test-taker who is asked to impersonate a male speaker, and thus needs to make necessary but unnatural grammatical adjustments.

Listening

Poorly designed items. The Listening section format is still in use today, meaning that most tests will have multiple-choice responses; however, what differs is the way in which the written stems, distractors and keyed alternatives are crafted and presented. For example, let us go back to the quantitative CFA analysis, which pointed out that only the 26th item of the Listening section, L26, does not relate well to the Listening factor. The aural stem asks (in translation): Why is Aurelio Suárez currently considered the most important Spanish surrealist painter? ["*¿Por qué le llaman a Aurelio Suárez el primer pintor surrealista español actual?*"]. The distractors are: a) Because he created surrealism ["*Porque originó el surrealism*"]; b) Because he painted exactly what he saw ["*Porque pinta exactamente lo que ve*"]; c) Because realism is a characteristic of Spanish art ["*Porque el realismo es un rasgo del arte español*"]. The keyed alternative reads: d) Because he is considered to be the best Spanish surrealist ["*Porque se le considera el mejor de los surrealistas espanñoles*"]. The question is poorly constructed from more than one point of view. Firstly, the usage of the word *primer* if incorrectly interpreted by the listener according to its primary definition of "first" rather than "foremost," may elicit an incorrect response. In this case, the test-taker would naturally choose distractor a) Because he created surrealism. Further still, the false cognate *actual* ("present day") is commonly misinterpreted by native as well as non-native Spanish speakers as the equivalent of the English "actual". This again would cause the listener to disregard the most important information: Suárez is the most important <u>contemporary</u> artist not the <u>first</u>, originating artist. Background information and the fact that the test is outdated might also elicit confusion on the part of the listener, as Suárez is not currently considered the foremost Spanish surrealist painter, a title disputed by Salvador Dalí, Joan Miró or even Antoni Tàpies. Again, the test-taker would in all likelihood mistakenly choose answer a) Because he created surrealism.

Although the CFA confirms all the other factors as relating well to Listening, most items do not follow what Nitko and Brookhart (2007) consider legitimate ways of constructing multiple-choice questions. For example, the 15th Listening item, L15, which has the third highest critical ratio (CR = 3.927) of the 36 questions, has the following stem (in translation): I don't understand my brother-in-law Miguel, even if I use an inappropriate word, he never takes it to heart ["*No puedo entender a mi cuñado Miguel. Aunque le diga una grosería, nunca lo toma a pecho*"]. The distractors are: a) He must not know to whom he should take it. ["*No sabrá a quién llevarlo*"], b) Your brother-in-law should not say such things. ["*Tu cuñado no debiera decir tales cosas*"] and c) It's that he's never been interested in drinking ["*Es que nunca le han interesado las bebidas*"]. The keyed alternative is: He must not be a very sensitive person ["*Será una persona poco sensible*"] Distractors (a) and (c) are not functional alternatives of multiple-choice items, as defined by Nitko and Brookhart (p. 158), because they are not appropriate to the stem, and are not arranged in a logical or meaningful order.

Audio quality, accents and speech rate. Further still, despite the adequate computer equipment, the quality of the audio recordings was poor due to the original reel-to-reel analog format that could not be enhanced during the digital conversion process. Brindley and Slayer (2002) pointed out that speech rate and accents are among factors that might affect task difficulty. In the case of the Spanish GMU MLA examinations, the speed with which stems are delivered differs from question to question and a variety of regional accents (Castilian Spanish, Argentinean Spanish, Puerto Rican Spanish, Mexican Spanish, etc.) are used, often within the same question thus potentially hindering the test-takers' ability to demonstrate Spanish comprehension.

Reading

Poorly designed items. According to Nitko and Brookhart (2007), the construction of multiple-choice distractors and keyed alternatives should not be framed by personal opinion, which is possibly the case for the 16th item (R_{16}) of the Reading section. The quantitative CFA analysis indicates that only R_{16} does not relate well to the reading factor. The directions state that the candidate should choose as the keyed alternative the one phrase that could be substituted for the underlined words in the passage, without changing the meaning of the sentence. The specific paragraph, in translation, is: "Anti-Semitism also existed outside Germany, especially in Central and Eastern Europe, but under Hitler's regime reached a maximum level. There has been an attempt to force the states that formed the Axis to establish a fund to compensate Jews for assets which were seized from them." ["*El antisemitismo existía también fuera de Alemania, especialmente en la Europa Central y Oriental, pero bajo el régimen hitlerista alcanzó el grado máximo. Se ha tratado de obligar a los estados que formaron el Eje a constituir un fondo de indemnización para resarcir a los judíos de los bienes que les fueron arrebatados*"]. The following alternatives are given as possible substitutes for the first underlined phrase (in translation): a) It has been futile, b) It has been necessary to think of, c) They have tried to, and d) It has been a matter of [a) "*Ha sido inútil*"; b) "*Ha habido que pensar en*"; c) "*Han intentado*", and d) "*Ha sido cuestión de*"]. One possible explanation for the poor performance on this question may be the test-takers' value judgment regarding whether or not the compensation plan has been successful.

Lack of context. The first fifteen items on the Spanish GMU MLA examinations' Reading section are presented as stand-alone sentences. Lacking context, these items become tests of vocabulary. The debate in testing Reading is no longer about whether the test-taker should be provided with context or not, but rather about how long the text should be and how varied its sampling fields (Cohen & Upton, 2007). Rupp, Ferne and Choi (2006), stated that the reasoning process a test-taker goes through is largely dictated by the multiple-question itself. If the stem lacks context, the test-taker is potentially encouraged to apply test taking strategies associated with deriving an answer based on the distractors rather than on the stem.

Writing

Antiquated language. The first passage in the Writing section of the Spanish GMU MLA examination contains an antiquated reference. It uses the old title for the European Organization for Economic Cooperation (EOEC) instead of the current name which is Organization for Economic Cooperation and Development (OECD). In Spanish, this translates to using the outdated [Organización Europea para la Cooperación Económica (OECE)] instead of the current [Organización para la Cooperación y el Desarrollo Económicos (OCDE)], but then inserts the future tense for the verb to meet ["reunirán"], thus confusing the test-taker into believing that the text is indeed talking about the past, which in turn might lead to them using incorrect verb tenses from that point onward.

Computerized delivery and design related issues. The test directions state: “Complete the text by writing a single Spanish word which is appropriate both in meaning and form in the corresponding answer box,” which is a restriction dictated by the convenience of having the test scored by the computer and the fact that it does not currently have the ability to recognize phrases that might fit better than a single word. Further, candidates are frustrated when they have to type accents and are not familiar with non-Spanish keyboard layouts. Perhaps the most problematic issue with the Writing section of the Spanish GMU MLA examinations is the design of the test items. While the computer could easily handle open-ended answer formats, it could not grade them. Thus, the fill-in-the-blanks format was dictated by convenience, i.e. the computerized delivery method, and not by theoretical guidelines.

Culture

Design and item construction. The revised Culture section of the Spanish GMU MLA examination is presented in English. It could be argued that since it is testing culture and not language it does not necessarily need to be in the target language. However, that brings us back to the issue of Spanish native speakers who might have difficulty with English, a circumstance which could potentially hinder their performance in the completion of tasks. Native speakers are not the only ones who cannot properly understand and interpret questions, when the language is vague. For example, “Today, there are many statues of Cortes [Cortés] in Mexico.” This decontextualized true-false question begs the interpretation of how many means “many” when taking into account that Cortés was not the most popular conquistador. Or, further, the simplified and taken out of context self-disqualifying true-false question: “The Creoles were happy with their situation in America.” Indeed, Nitko and Brookhart (2007) warned against poorly constructed true-false items: “[they] assess only specific, frequently trivial facts; are ambiguously worded; are answered correctly by random guessing” (p. 139).

The authors also approached the topic of matching items, and stated that there are three components to a matching exercise: directions for matching, list of premises and list of responses. This section is confusing right from the beginning. The only directions are referring to what the test-taker should do in order to overcome the shortcomings of the digitized format: “click on the down arrow of each box to select your answer. To see all possible answers, you will have to scroll down in the opened box.” Nitko and Brookhart (2007) also provided crafting suggestions (p.177), most of which are not met by the matching items in the Culture section. For example, the premises are incomplete sentences, e.g. “member of the Araucanian tribe in Chile,” there are more than ten responses, thirty to be exact, which are not clearly numbered for both premise and response and are equal in number, thus potentially giving away at least one answer.

In conclusion, it appears that features of the assessment tasks do indeed detract from a candidate’s ability to demonstrate thinking and substantive processes associated with Spanish language proficiency. All unrevised sections are weighed down by the issues presented; even the revised Culture section has problems of its own.

Question 3: Are there anticipated or unanticipated side effects from interpreting and using the Spanish GMU MLA examinations as tests of proficiency?

When discussing consequential validity studies, Parke, Lane and Stone (2006) stated that: “this aspect of validity evidence examines the impact that the standards and assessments have on curriculum, classroom instruction, and assessment practices, beliefs and attitudes, professional development support, preparation for the assessment, and decision-making” (p. 241). Considering the above aspects, the one positive consequence when interpreting and using the Spanish GMU MLA examinations as tests of proficiency is the fact that a validation effort such as the current study has been attempted. Even if such a consequence was unintended, it should prove beneficial in terms of crafting and analyzing future tests of proficiency. Regrettably, there are more negative unintended consequences than positive ones. For example, the fact that the picture-prompts for the Speaking section were never changed, they became well-known (the picture-prompts are the same for all languages in which the GMU MLA examinations are offered), could possibly allow test-takers to prepare and thus mask understanding by learning rote responses. Further, the Speaking section might lead test-takers into believing that to be proficient in Spanish means to be able to read a text out loud versus demonstrating communicative competence, while to be proficient in Writing means to be able to insert vocabulary items in predetermined spaces, rather than to create text based on open-ended questions. The fact that the Culture section is a poorly constructed true-false and matching exercise reduces its impact and value as a significant construct in language proficiency by leading test-takers to believe that it is no longer or less important than other components. Encompassing all sections, could be the belief that a condition for demonstrating Spanish proficiency is to be more than simply computer literate.

The fact the administrators purport that the Spanish GMU MLA examination is a proficiency test, but have conducted no predictive validity studies, weakens the case that the test is a proficiency one. For the population in this study, the fact that the vast majority of candidates passed the test may not necessarily be a clearly defined function of their Spanish proficiency, but rather a combination of the passing bar being set too low in conjunction with a certain level of command of Spanish.

Question 4: How do the factors of cost, efficiency, and practicality justify the use of the Spanish GMU MLA examinations?

Given that the Spanish GMU MLA examinations are delivered via WebCT, they are practically self-administering. The proctor's role is more that of a technical support person. The computer offers built-in timers, carrels eliminate the danger of plagiarism, and, perhaps most conveniently, the computer scores four of the five sections: Listening, Reading, Writing, and Culture. The single rater spends on an average ten minutes on each Speaking sample and receives minimal compensation. Therefore, considering that the infrastructure is already in place at GMU, there are minimal costs involved in both administration and analyses. Consequently, the profit for the administering unit is high.

The Spanish GMU MLA examinations use the multiple-choice response format for the Listening, Reading and Culture sections. Multiple-choice questions, like all response formats, have five types of costs associated with them: development, administration, scoring, analysis, and reporting. Superficial or poorly designed multiple-choice items are easy to develop, implement, test and score, but the analysis might be suspect. In contrast, more cognitively challenging multiple-choice items are more expensive to develop, equally cheap to administer and score, however, potentially richer analytically. The design of some of the Spanish GMU MLA examinations' multiple-choice items is questionable, as mentioned in the answer to Question 2(b) above.

Non-objective assessment methods, e.g. performance assessments, portfolios, essays, open-ended questions, are potentially selecting from more authentic aspects of language proficiency but at the same time tend to be more expensive in development, administration, scoring, and analysis. Whether the greater expenditures are justifiable is a matter of professional judgment.

In conclusion, the use of the Spanish GMU MLA examination is justified both in terms of cost and practicality. Furthermore, the examinations are efficient, since the test-takers are processed in a timely manner. Also, the scoring reports are available immediately for all sections but the one that has to be hand-scored (Speaking).

5. Findings and Discussion

This study sought to answer the overarching research question of whether the judgments based on data from the Spanish GMU MLA examinations are reliable and valid indicators of Spanish proficiency for teachers seeking licensure. An *ex post facto* approach was used, which blended a quantitative approach with a descriptive component. For both the quantitative and the descriptive investigation, the test was examined using an adaptation of the criteria for test validity and reliability proposed by Nitko and Brookhart (2007, p. 46). The quantitative data was collected from the modified Spanish Modern Language Association Proficiency Test administered at George Mason University. The framework for the accumulated evidence used in the validation process of the scores' interpretation was based on the Spanish language proficiency concept that was to be measured by the Spanish GMU MLA examination. The research questions were:

Quantitative

1. Is there evidence that the Listening, Reading, and Writing factors represent the underlying structure of the proficiency examination? (*content and internal structure evidence*)
2. Are the results of the assessment internally consistent or reliable over time? (*reliability evidence*)

3. Is there evidence that the Speaking, Listening, Reading, Writing and Culture factors interact with the variables gender and nationality and with the average time taken to complete the test? Are there significant correlations among these five factors? (*external and generalization evidence*)
4. Are the results predictive of future performance or consistent with the results of similar testing (*external structure evidence*)?

Descriptive

1. Do the hypothesized factors represent constructs in more current theories for assessing language proficiency (*content evidence*)?
2. a. How relevant to the thinking skills and processes of demonstrating language proficiency are the tasks in the Spanish GMU MLA examinations (*substantive evidence*)?

   b. To what extent do any features of the assessment tasks detract from a candidate's ability to demonstrate these thinking and substantive processes?
3. Are there anticipated or unanticipated side effects from interpreting and using the Spanish GMU MLA examinations as tests of proficiency (*consequential evidence*)?
4. How do the factors of cost, efficiency, and practicality justify the use of the Spanish GMU MLA examinations (*practicality evidence*)?

Discussion and Conclusions

Is there evidence that the Listening, Reading, and Writing factors represent the underlying structure of the proficiency examination?

The testing for the validity of the factors Listening, Reading and Writing was conducted through the use of a confirmatory factor analysis (CFA), using the computer program for latent trait analysis Mplus (Muthén & Muthén, 2003). The following four goodness-of-fit indices were used with the CFA in this analysis - *chi-square fit statistic*, (χ2), c*omparative fit index*, *CFI*, *standardized root mean square residual*, *SRMR*, and *root mean square error of approximation*, *RMSEA*, with a 90 percent confidence interval. A relatively good fit is indicated with *CFI* > .90, SRMR <.08, and RMSEA < .06 as recommended by Browne and Cudeck (1993) and by Hu and Bentler (1999). Given the sensitivity of the chi-square statistic to sample size, its role in CFA testing for model data fit is more descriptive than inferential (Dimitrov, 2006).

The results presented in Table 3 show that CFI (.616) and the Tucker-Lewis Index TLI (.607) do not provide evidence for data fit of the model, while the most frequently recommended indices do provide such evidence: RMSEA = 0.56, with 90% CI = (.054; .058) and SRMR = .075. Therefore, after applying the four goodness-of-fit indices, evidence for the three factors on the underlying structure of the proficiency test was mixed. Although the *CFI* score was less than the recommended value for acceptable data-model fit, the RMSEA and SRMR values were within acceptable limits. The model provides a pilot explanation of the data but it is not one that explains the data across all four of the recommended fit indices. The reasonable conclusion is that there are additional factors that explain the data, but they are not known at this time.

Are the results of the assessment internally consistent or reliable over time?

When the Cronbach's alpha for the combined dependent variables was calculated, each of the three subscales (Listening, Reading and Writing) was reliable. However, it is not known whether the scores are reliable over repeated administrations or with alternate forms, since data was not available.

Is there evidence that the Speaking, Listening, Reading, Writing and Culture factors interact with the variables Gender and Nationality and with the Average Time taken to complete the test? Are there significant correlations among these five factors?

After running the analyses described in Chapter 4, there was some evidence that the factors used in the model correlated with the predictor variables gender, nationality, and average time taken to complete the test (see Table 5). For the Total Listening subscale score, both Nationality and Average Listening Time were statistically significant predictors. Not surprisingly, native speakers performed better and on average in less time than non-native speakers on this subscale. For the Total Speaking subscale score, only Nationality was a statistically significant predictor. Again, not surprisingly, native speakers performed better than non-native speakers on this subscale. However, there was no difference between native and non-native speakers in the time taken to complete the subscale and the total score. For the remaining subscale scores, Reading, Writing and Culture, none of the predictor variables had a statistically significant relationship. The result for the Culture subscale is easier to explain than for Reading and Writing, because it is entirely in English, which means that native speakers may have actually been at a disadvantage when compared to non-native speakers. However, for the Reading and Writing subscale scores, the lack of statistically significant relationship with any of the predictor variables is harder to explain. Native language ability was not an advantage, which could be attributed to any of the factors mentioned by Bachman (2000), namely characteristics of the testing procedure, test-takers' strategies and test-takers' own characteristics, such as academic background.

Are the results predictive of future performance or consistent with the results of similar testing?

While it would have certainly been advisable and interesting to cross-reference data from the Spanish GMU MLA examinations with data from the PRAXIS and/or OPI tests, this possibility did not exist for the current study. In brief, the lack of such data represents a serious validity issue for the use of the examination as a proficiency test.

To what extent do the analyses of the factors sample from constructs in more current theories for assessing language proficiency?

At a macro level, the constructs in more current theories for assessing language proficiency do not appear to differ drastically from the Spanish GMU MLA examination. Listening, speaking, reading and writing are all accounted for in one form or another. However, what does differ is the way in which the constructs are built, presented and assessed. Only parts of the Reading section could be encountered in more current theories, as the other four sections are apparently targeting knowledge about Spanish rather than its use in communicative, real-life situations.

How relevant to the thinking skills and processes of demonstrating language proficiency are the tasks in the GMU MLA examination?

Speaking. While the Spanish GMU MLA Speaking test employs prompts unrelated to real-world tasks, the OPI achieves elicitation by means of a meaningful dialogue with the rater. In the case of the Speaking section, when compared against the criterion, the required thinking skills would be relevant only in the condition that the real-world situation were experienced as portrayed in the test. This is highly unlikely, considering that teacher-candidates are not often placed in the context that corresponds to tasks in the test. Only the last, and shortest, part of the Speaking section is more involved. It prompts the test-taker to produce language that varies at the linguistic, lexical and communicative level of complexity, an activity which most likely requires complex thinking skills. Such thinking skills are part of demonstrating language proficiency.

Listening. Listening tests that use multiple-choice or true-false response formats tend to dominate the testing of this skill. From this perspective, the Listening section of the Spanish GMU MLA examination does evidence thinking skills and processes necessary to demonstrate Spanish proficiency. However, a comparison to ACTFL's OPI would not be favorable, since the latter does take into consideration that language comprehension is a dynamic top-down approach that occurs between interlocutors.

Reading. The Reading section seems to be similar in structure to current reading tests: omitted words or phrases in sentences, paragraphs and passages, with multiple-choice answer formats. It would thus appear that the thinking skills utilized to demonstrate Spanish reading proficiency are present in this section of the test, minus the first fifteen items that are decontextualized and should belong in a vocabulary retention test rather than in a reading one.

Writing. The Writing section, although confirmed by the CFA analysis with critical ratios above 20, appears to be the least sound measure of all. The findings section revealed that there was no comparison possible with other measures in use today. However, what can be concluded is the fact that the thinking skills involved are at the lower level of Bloom's taxonomy, since the section asks the test-taker to demonstrate knowledge of grammar and vocabulary rather than the more cognitively demanding exercise of producing text.

Culture. Considering that the PRAXIS II (ETS, 2008) test and the Spanish GMU MLA Culture examination both use the multiple-choice format, it would appear that the thinking skills and processes employed by the GMU MLA test-takers are relevant to demonstrating Spanish culture proficiency. However, since no direct correlation with the PRAXIS examination exists, this is statement needs to be considered speculative.

To what extent do any features of the assessment tasks detract from a candidate's ability to demonstrate these thinking and substantive processes?

The mode of administration and the actual content of the test items in the Spanish GMU MLA examinations were the most controversial findings. The numerous perceived shortcomings ranged all the way from: use of English, Eurocentric approach, spelling errors, lack of practice tests, unclear directions, choice of prompts, poorly designed items, poor audio quality, speech rate, accents, lack of context, antiquated language, design and item construction, to the all encompassing computerized mode of administration. All these point to the fact that candidates would indeed have to be skilled navigators in order to avoid the potential pitfalls. In conclusion, features of the assessment tasks do indeed detract from a candidate’s ability to demonstrate thinking and substantive processes associated with Spanish language proficiency.

Are there anticipated or unanticipated side effects from interpreting and using the GMU MLA examinations as tests of proficiency?

The findings here were that unanticipated negative consequences abounded. Positive side effects might be restricted to entries in the last section of this study, namely recommendations that stemmed from the identified problem-areas.

How do the factors of cost, efficiency, and practicality justify the use of the GMU MLA exam?

The use of the Spanish GMU MLA examinations appears to be justified from all points of view. The cost for development was minimal, and the one for administration is not high either. The return is high for the administrative unit, but since test-takers are processed efficiently and their scoring reports are made available immediately for all sections but Speaking, all stakeholders are satisfied.

In conclusion, the quantitative analyses provided mixed results related to the predictors’ ability to explain the subscale scores. The descriptive analyses revealed numerous validity flaws and gaps in the test construction. Such inadequacies overshadowed the positive aspects of the test and eventually led to my questioning of the hypothesis that the judgments based on data from the Spanish GMU MLA examinations are reliable and valid for the population in this study.

Recommendations and Directions for Future Research

The example of what happened to the Spanish GMU MLA examinations should be impetus enough to not let such an occurrence be repeated.

Considering the limitations of this study and those of the Spanish GMU MLA examinations I could most likely formulate directions for future research along the same lines present here. However, instead of an exhaustive but not very productive research agenda, I propose to gather what I have learned and proceed onward, in a new direction: that of constructing a new test.

I understand that cost, efficiency and practicality considerations are often the triggers of inaction, but perhaps these could be offset by a better test design that is anchored in current theories of both language acquisition and testing. In order to succeed, such a framework needs to receive support from other disciplines ranging from brain research to artificial intelligence. The key to overcome most of the hurdles described throughout this study is the synergetic inclusion and co-opting of those specialists who can offer solutions to such issues as creating an artificial interlocutor for a speaking test, understanding the cognitive processes that the reader of a second language goes through, providing an automated way to score open-ended questions and essays. While some of these tendencies already exist or are beginning to emerge, more research is needed at the scientific intersection of a multidisciplinary approach to test-creation and investigation of varied types of evidence that support the validity of an assessment.

APPENDIX A:

Modifications to the Spanish MLA Test

The following modifications were made by members of the Spanish faculty in the Department of Modern and Classical Languages at George Mason University:

Speaking

- The original test had three parts: the first consisted of an exercise that asked the test-taker to repeat model sentences spoken by a native speaker. This part was deleted because the ability to imitate the speaker's intonation and pronunciation was judged to be a weak measure of the candidates' ability to speak.

Writing

- The original test consisted of two parts: Part A with two fill-in-the-blanks texts and Part B with two poorly written passages that the test-taker needed to edit and/or revise. Part B was deleted and replaced with two new fill-in-the-blanks texts. The substitution was deemed necessary because the initial format was considered to be a measure of rewriting and not one of writing, this despite the fact that it was considered authentic in comparison with other tasks.

Culture and Civilization

- The whole original test was considered obsolete, mainly because most keyed alternatives were no longer historically correct, or because the formulation was deemed unacceptable (e.g. "Which of the following best expresses the general attitude of Spaniards toward the consumption of alcoholic beverages? (A) The drinking of wine and liquor is contrary to the precepts of the Catholic Church. (B)

The drinking of wine is acceptable, but the drinking of liquor is reprehensible. (C) The drinking of wine at meals and of liquor after meals is a matter of personal taste not involving morals. (D) The drinking of wine and liquor is required for social acceptance and even drunkenness is not frowned upon.”).

APENDIX B

Table 3

Factor Loadings Estimates and Their Standard Errors: Confirmatory Model for Listening, Reading, and Writing

```
DATA:

   FILE IS "C:\MY FILES\WALTER\FACTORS.dat";
   FORMAT IS 92F2.0;

 VARIABLE:

   NAMES ARE G Origin L1-L36 R1-R50 W1-W4;
   USEVARIABLES ARE L1-L36 R1-R50 W1-W4;

 ANALYSIS:

   TYPE IS GENERAL;
   ESTIMATOR IS ML;
   ITERATIONS = 1000;
   CONVERGENCE = 0.00005;

 MODEL: LISTEN BY L1-L36;
          READ BY R1-R50;
          WRITE BY W1-W4;
```

```
TESTS OF MODEL FIT

Chi-Square Test of Model Fit

       Value                      7364.594
       Degrees of Freedom             3912
       P-Value                      0.0000

Chi-Square Test of Model Fit for the Baseline Model

       Value                     12990.086
       Degrees of Freedom             4005
       P-Value                      0.0000
```

```
CFI/TLI

          CFI                          0.616
          TLI                          0.607

Loglikelihood

          H0 Value                   -10792.940
          H1 Value                    -7110.643

RMSEA (Root Mean Square Error Of Approximation)

          Estimate                       0.056
          90 Percent C.I.                  0.054  0.058
         Probability RMSEA <= .05            0.000

SRMR (Standardized Root Mean Square Residual)

          Value                        0.075




LISTEN BY

    L2               4.015     1.031      3.894
    L3               4.142     1.166      3.551
    L4               2.142     0.677      3.163
    L5               2.185     0.659      3.313
    L6               2.828     0.821      3.444
    L7               2.259     0.710      3.180
    L8               2.856     0.749      3.816
    L9               3.155     0.830      3.799
    L10               1.227     0.336      3.651
    L11               1.594     0.433      3.683
    L12               3.766     1.097      3.432
    L13               2.845     0.727      3.912
    L14               2.325     0.666      3.491
    L15               4.519     1.151      3.927
    L16               5.278     1.374      3.842
    L17               4.610     1.160      3.974
    L18               5.390     1.337      4.032
    L19               4.240     1.166      3.636
    L20               4.975     1.312      3.792
    L21               3.015     0.778      3.875
```

L22	3.978	1.091	3.644
L23	2.721	0.766	3.551
L24	3.382	0.943	3.586
L25	2.397	0.765	3.132
L26	1.110	0.712	1.558
L27	2.904	0.912	3.185
L28	1.761	0.630	2.797
L29	3.921	1.102	3.558
L30	2.292	0.679	3.376
L31	4.549	1.222	3.724
L32	3.445	1.019	3.381
L33	4.054	1.109	3.654
L34	3.046	0.904	3.369
L35	4.641	1.243	3.733

READ BY

R2	1.159	0.157	7.393
R3	1.292	0.187	6.920
R4	1.703	0.225	7.573
R5	2.301	0.264	8.726
R6	1.202	0.180	6.691
R7	2.275	0.280	8.110
R8	1.644	0.200	8.214
R9	1.541	0.268	5.742
R10	1.893	0.284	6.660
R11	0.730	0.237	3.079
R12	1.496	0.219	6.832
R13	1.944	0.286	6.791
R14	2.096	0.276	7.583
R15	2.068	0.273	7.584
R16	0.257	0.165	1.562
R17	1.150	0.259	4.435
R18	2.329	0.302	7.709
R19	1.716	0.239	7.169
R20	1.632	0.275	5.931
R21	0.898	0.127	7.059
R22	1.140	0.261	4.374
R23	1.714	0.273	6.274
R24	0.557	0.122	4.555
R25	1.006	0.186	5.422
R26	1.327	0.259	5.134

R27	0.932	0.143	6.526
R28	0.819	0.180	4.542
R29	1.885	0.258	7.297
R30	0.881	0.138	6.371
R31	1.129	0.177	6.380
R32	1.133	0.167	6.787
R33	1.141	0.257	4.438
R34	1.504	0.205	7.328
R35	1.906	0.283	6.741
R36	1.804	0.260	6.946
R37	1.651	0.274	6.029
R38	1.463	0.214	6.846
R39	0.482	0.166	2.908
R40	1.197	0.262	4.568
R41	2.156	0.282	7.634
R42	2.022	0.281	7.203
R43	1.899	0.285	6.675
R44	0.690	0.251	2.751
R45	1.055	0.229	4.603
R46	1.347	0.213	6.309
R47	1.130	0.260	4.339
R48	0.699	0.250	2.801
R49	0.931	0.248	3.760

WRITE BY

W2	1.142	0.054	21.060
W3	1.163	0.057	20.429
W4	1.201	0.053	22.719

READ WITH

LISTEN	0.005	0.001	3.599

WRITE WITH

LISTEN	0.090	0.024	3.793
READ	0.295	0.040	7.386

```
Variances

   LISTEN           0.002    0.001      2.045

   READ            0.016    0.003      4.719

   WRITE            6.413    0.688     9.320
```

REFERENCES

REFERENCES

ACTFL. (1999). *Proficiency guidelines-speaking*. Yonkers, NY: American Council on the Teaching of Foreign Languages.

ACTFL. (2001). *Revised preliminary proficiency guidelines - writing*. Retrieved July 14, 2008 from http://www.actfl.org/files/public/writingguidelines.pdf

ACTFL. (2006). *Standards for foreign language learning in the 21st century*. Retrieved July 14, 2008 from http://www.actfl.org/files/public/StandardsforFLLexecsumm_rev.pdf

Alderson, C. J., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*, 79-113.

American Council on the Teaching of Foreign Languages. (1986). *ACTFL Proficiency Guidelines*. Yonkers, NY: American Council on the Teaching of Foreign Languages.

American Educational Research Association, American Psychological Association, & National Council on Measurements in Education. (1999). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association.

American Psychological Association, American Educational Research Association, & National Council on Measurements in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, D.C.: American Psychological Association.

American Psychological Association, American Educational Research Association, & National Council on Measurements Used in Education. (1954). *Technical Recommendations for Psychological Tests and Diagnostic Techniques*. Washington, D.C.: American Psychological Association.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Anderson, N., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*, 41 - 66.

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

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

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

Bernhardt, E. B. (1991). *Reading development in a second language: Theoretical research and classroom perspectives*. Norwood, NJ: Ablex.

Bernhardt, E. B. (1993). *Reading development in a second language: Theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex.

Bloomfield, L. (1933). *Language*. New York: H. Holt and Company.

Brindley, G., & Slatyer, H. (2002). Exploring task difficulty in ESL listening assessment. *Language Testing, 19*, 369 -394.

Browne, M. W., & Cudek, R. (1992). Alternative ways of assessing model fit. *Sociological Methods Research, 21*, 230 - 258.

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11*, 145 -170.

Buck, G. (Ed.). (2001). *Assessing Listening*. Cambridge: Cambridge University Press.

Calhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*, 3-22.

Canale, M. (Ed.). (1987). *The measurement of communicative competence* (Vol. 8). New York: Cambridge University Press.

Carroll, J. B. (Ed.). (1972). *Fundamental considerations in testing for English language proficiency of foreign students.* (2nd ed.). New York: McGraw Hill.

Cohen, A. D., & Upton, T. A. (2007). `I want to go back to the text': Response strategies on the reading subtest of the new TOEFL®. *Language Testing, 24*, 209 - 250.

College of Education and Human Development University of Minnesota. (2006). *Language proficiency requirements for endorsement in world languages and cultures*. Retrieved June 18, 2006, from http://education.umn.edu/SPS/programs/addlic/WLC-lang.html

Coniam, D. (1998). Interactive evaluation of listening comprehension: how the context may help. *Computer Assisted Language Learning, 11*, 35-53.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Daily, C. B. (2001). *SOL: U -- French language trainer/translator*. Retrieved June 18, 2006, from http://www.fbodaily.com/cbd/archive/2001/05(May)/01-May-2001/usol003.htm

Department of Bilingual Education Texas A&M University Kingsville. (2006). *Doctoral studies in bilingual education*. Retrieved June 19, 2006, from http://bilingualed.tamuk.edu/?page=programs/doctoral

Department of Languages and Literature College of Humanities The University of Utah. (2006). *Frequently asked questions*. Retrieved June 17, 2006, from http://hum.utah.edu/display.php?pageId=951

Department of Languages Houston Baptist University. (2006). *Bilingual education*. Retrieved June 19, 2006, from http://fc.hbu.edu/arts&human/language/bilingualeducation.htm

Department of Modern Languages Central Connecticut State University. (2003). *The MLA exam*. Retrieved June 18, 2006, from http://www.modlang.ccsu.edu/Spanish/H_Files/Q4yes.html

Dimitrov, D.M. (2006). Comparing groups on latent variables: A structural equation modeling approach**.** *WORK: A Journal of Prevention, Assessment, & Rehabilitation, 26*, 429-436.

Elola, I., & Oskoz, A. (2008). Blogging: Fostering intercultural competence development in foreign language and study abroad contexts. *Foreign Language Annals, 41*, 454 - 477.

ETS. (2008). *Spanish: Content knowledge (0191)*. Retrieved August 25, 2008, from http://www.ets.org/Media/Tests/PRAXIS/pdf/0191.pdf

Farhady, H., & Keramati, M. N. (1996). A text-driven method for the deletion procedure in cloze passages. *Language Testing, 13*, 191 - 207.

George Mason University College of Education and Human Development Graduate School of Education. (2006). *Foreign language and Latin (PK-12) initial licensure with Master of Education in curriculum and instruction option, endorsement requirements*. Retrieved May 14, 2006, from http://gse.gmu.edu/programs/descriptions/foreignlang.htm

George Mason University Department of Modern and Classical Languages. (2006). *Modern Language Association proficiency exam*. Retrieved May 14, 2006, from http://mcl.gmu.edu/mla.html

Hamp-Lyons, L. (1993). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second Language Writing* (pp. 69 - 88). New York, NY: Cambridge University Press.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1 - 55.

Hughes, A. (1989). *Testing for language teachers*. New York: Cambridge University Press.

Kaulfers, W. V. (Ed.). (1965). *Review of MLA foreign language proficiency test for teachers and advanced students: Spanish*. Highland Park, NJ: The Gryphon Press.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*, 212 -218.

Kunnan, A. J. (2004). Regarding language assessment. *Language Assessment Quarterly, 1*, 1-4.

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers.* Ann Arbor: University of Michigan Press.

Lado, R. (1961). *Language testing*. New York: McGraw-Hill.

Language Learning Center Department of Foreign Languages and Literatures Old Dominion University. (1999). *The MLA cooperative foreign language proficiency exams*. Retrieved June 17, 2006, from http://www.odu.edu/al/llc/mla/

Long, M. H., & Norris, J. M. (2000). *Task-based language teaching and assessment*. In M. Bryam (Ed.), (pp. 597 - 603). London: Routledge.

Madsen, H. S. (1983). *Techniques in testing*. New York: Oxford University Press.

Magnan, S. S. (2006). From the editor: The *MLJ t*urns 90 in a digital age. *Modern Language Journal, 90*, 1-5.

Maxwell, J. A. (1996). *Qualitative research design: An interactive approach* (Vol. 41). Thousand Oaks, CA: Sage Publications.

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13 - 103). New York: Macmillan.

Mislevy, R., Steinberg, S., & Almond, R. (2002). Design and analysis in task-based language assessment. *Language Testing, 19*, 477 - 496.

Muthén, L. K., & Muthén, B.O. (2003). *Mplus: Statistical analysis with latent variables. User's guide*. Los Angeles, CA: Muthén and Muthén

.

Nitko, A. J., Brookhart, Susan M. (2007). *Educational assessment of students* (5th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Oller, J. W. J. (Ed.). (1976). *Language Testing* (2nd ed.). Ann Arbor: The University of Michigan Press.

Oller, J. W. J. (1979). *Language tests at school: a pragmatic approach*. London: Longman.

Parke, C. S., Lane, S., & Stone, C. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation, 12*, 239 - 269.

Pearson, L., Fonseca-Greber, B., & Foell, K. (2006). Advanced proficiency for foreign language teacher candidates: What can we do to help them achieve this goal? *Foreign Language Annals, 39*, 507-519.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press.

Probst, G. W. (Ed.). (1972). *Review of MLA foreign language proficiency test for teachers and advanced students: Spanish* (Vol. 1). Highland Park, NJ: The Gryphon Press.

Quinn Allen, L. (2004). Implementing a culture portfolio project within a constructivist paradigm. *Foreign Language Annals, 37*, 232 - 240.

Rivers, W. (1968). *Teaching foreign-language skills*. Chicago: The University of Chicago Press.

Rocklin, T. (1999). Computers and Testing. *The National Teaching and Learning Forum, 8*. Retrieved July 9, 2006, from http://www.ntlf.com/html/pi/9909/v8n5smpl.pdf

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*, 441 - 474.

San Diego State University Graduate Division. (2006). *Foreign language requirement.* Retrieved June 18, 2006, from http://gra.sdsu.edu/index.php?areaid=1&sectionid=51&subsectionid=13

Schulz, R. A. (2007). The challenge of assessing cultural understanding in the context of foreign language instruction. *Foreign Language Annals, 40*, 9 - 27.

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

Spolsky, B. (2000). Language testing in *The Modern Language Journal*. *The Modern Language Journal, 84*, 536-552.

Starr, W. (1962). MLA foreign language proficiency tests for teachers and advanced students. *PMLA, 77*, 31-42.

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing, 14*, 214 - 231.

The Department of Foreign Languages and Literatures State University of New York at Geneseo. (2005). Retrieved June 18, 2006, from http://www.geneseo.edu/academic_depts/dept/?pg=Fren/degree.html

Truman State University. (2006). *Assessment*. Retrieved June 18, 2006, from http://assessment.truman.edu/components/SeniorTests.htm

Virginia Department of Education Licensure Division. (1998). *Licensure regulations for school personnel*. Retrieved May 14, 2006, from http://www.doe.virginia.gov/VDOE/Compliance/TeacherED/nulicvr.pdf

CURRICULUM VITAE

Walter J. Mircea-Pines received his Bachelor of Arts from the Babes-Bolyai University in 1985, with a specialization in Germanic languages. In 1996, he obtained the Master of Education degree from the State University of New York at Buffalo, with a specialization in foreign language education. He taught both K-12 and Adult Education foreign language classes while in Western New York. Upon relocating to Northern Virginia, he continued teaching foreign language courses to U.S. government employees who were about to work abroad. Throughout this whole period, he continued to cultivate his computing skills and appreciation for technology in general.

From 1998 to 2009, Mr. Mircea-Pines held various positions at George Mason University (Mason): instructor of German, instructional technologist, information technology coordinator and language laboratory director. He is currently an administrative faculty member at Mason, in the Department of Modern and Classical Languages, College of Humanities and Social Sciences.