Final Report
on the Pilot of a Certification Process for Spanish-English Interpreters in Health Care

Submitted To:

Office of Minority Health
U.S. Department of Health and Human Services

Contract Number: 02T02502101D
Project Officer: Guadalupe Pacheco

The National Council on Interpreting in Health Care

November 2003

Acknowledgements
This certification pilot was the work of a group of dedicated volunteers within the Massachusetts Medical Interpreters Association (MMIA), the California Health Care Interpreting Association (CHIA) in partnership with Healthy House of Merced, and the Standards, Training, and Certification Committee of the National Council on Interpreting in Health Care. Without their selfless commitment, this project would have been impossible.

MMIA certification pilot team
Jane Kontrimas (Chair)
Maria-Paz Beltran Avery, Ph.D
Eduardo Berinstein
Frank Geoffrion
Julie Burns, M.Ed.
Grace Peters
Lisa Morris
Joy Connell (Ex-officio member, President of MMIA)

CHIA / Healthy House certification pilot team
Julie Burns, M.Ed (CHIA)
Betty Moore, M.L.S. (CHIA)
Marilyn Mochel, R.N. (Healthy House)
Tatiana Vizcaino-Stewart (Healthy House)

We are grateful to the Office of Minority Health, U.S. Department of Health and Human Services, and Mr. Guadalupe Pacheco, MSW, Project Officer, for supporting this effort. Mr. Pacheco provided important guidance and direction throughout this project.

A special thanks to all the test administrators and raters, and to all the interpreters who participated in this study. Thanks also to Merced College in Merced, CA and to Boston College for allowing the use of their facilities free of charge; to Cynthia Bravo, the Director of the Boston College Language Laboratory for her assistance with test administration in Boston; and to the Latin American Health Institute in Boston for assistance with the statistical analysis.

This report was written in part by Cynthia E. Roat, MPH, co-chair of the NCIHC board, based in large part on a report submitted to NCIHC by Maria-Paz Beltran Avery, PhD. Recognition should also go to Jane Kontrimas and Wilma Alvarado-Little, MA, for their input into this document. This report was reviewed by the NCIHC Board of Directors in November 2003.

Introduction

As the population of limited English proficient (LEP) individuals in the United States grows both in number and in distribution, health care providers across the country have been struggling to find ways to facilitate communication with them. Research and experience have shown that the historical use of family and friends as interpreters leads to inaccurate and incomplete message transfer. But is the performance of bilingual staff or a contracted interpreter any better? How can the quality of an interpreter's services be guaranteed?

Quality assurance in interpreting is accomplished through a number of steps, including the recruitment of individuals with strong language skills followed by appropriate training, certification, and continuing education. Introducing one step without the others will not assure quality interpretation. Over the past decade, the availability of training for health care interpreters has increased significantly in many parts of the United States. Initial language assessment tests are emerging and continuing education programs are being developed.

There is also a great deal of interest in the development of a certification for interpreters in health care. Many working interpreters would like to be certified as a means of separating themselves from the larger candidate pool of less qualified interpreters. Many health care institutions and language agencies would like to see certification both as a way to guarantee quality and as a means of limiting their own legal liability.

Certification, however, is a large and difficult task. A certification process is a test of a specific demonstrable skill as measured against a publicly acknowledged standard. Practitioners in the field must agree on the skills to be measured and the standards to be applied. In addition a certification process must be proven scientifically valid and reliable in order to be accepted in the field.

The first certification process for social service interpreters in the U.S. was introduced in 1993 by the Washington State Department of Social and Health Services in response to a court order; two years later this office introduced a parallel exam for interpreters in health care. This test is administered only in Washington State, and, while a pioneering effort, is regarded by many as flawed.[1] That same year, 1993, the State of California Office of Personnel also instituted a "Medical Interpreter Certification" under the jurisdiction of the California Judicial Council, however this is a test for judicial interpreters serving in worker's compensation cases and not a test of clinical interpreting skills. In the past several years, two commercial telephonic interpreting companies (Language Line Services and Network OMNI) have also announced the introduction of proprietary certification tests; only the former is available to interpreters outside the company's network, and both are generally seen as being tests of telephonic interpreting skills only. Neither was developed with public input. There is clearly a need for a nationally available certification process, based on publicly vetted standards and protocols.

The National Council on Interpreting in Health Care (NCIHC) has been working since its incorporation in 2001 to develop the foundation for such a certification process. With support from the U.S. Department of Health and Human Services' Office of Minority Health, the NCIHC

[1] For a summary of concerns about the Washington State Medical Interpreter Certification exam, please see Roat, Cynthia. *Certifying Medical Interpreters: Some Lessons from Washington State.* The ATA Chronicle, 28(5), May 1999, pgs. 23 - 26

started in 2001 with the development of a nationally vetted Code of Ethics for Interpreters in Health Care. A process to develop a set of national Standards of Practice for Interpreters in Health Care will start in January 2004 with funding from The Commonwealth Fund and The California Endowment. In the interim, the NCIHC felt it would be important for the country to gain more regional experience around certification so that the development of a national certification could be started as soon as the Standards were in place. There was experience is Washington State; what other state might be ready to start certification of health care interpreters?

Two states did meet the criteria of having generally accepted standards of practice and making training widely available to interpreters in health care. In one of these states, Massachusetts, the Massachusetts Medical Interpreters Association (MMIA) had been working on a certification instrument for a number of years, however a lack of funding was hampering their efforts. In the other state, California, the California Healthcare Interpreter Association (CHIA) was eager to gain experience in certification in preparation for initiating its own process in the next several years. With a view toward building inter-organizational links and building a base of experience, the NCIHC sought and received in 2002 a contract from the U.S. Department of Health and Human Services Office of Minority Health to help MMIA and CHIA pilot this certification process. This is a report of that collaboration.

This report is organized in five sections. The first describes the intended purposes of the certification process and the principles that guided the design and development of the Medical Interpreting Assessment for Certification prototype (MIAC) developed by the MMIA both in its pre-pilot and pilot version. The next section describes the version of the assessment instrument that was used in the pilot. The third section details the work done on the pre-pilot. The fourth section describes the collaborative pilot process and reports on the analysis of the properties of the test as it was piloted in Massachusetts and California. The final section discusses what was learned through this collaborative pilot process and offers recommendations for future work.

Development of the Assessment Instrument[2]

Shortly after the Massachusetts Medical Interpreters Association developed and adopted the Medical Interpreting Standards of Practice in 1995 (published by the MMIA and the Education Development Center), its Standards Committee reconstituted itself as the Certification Committee and began discussing the creation of an assessment tool to certify medical interpreters in Massachusetts. Working on a purely volunteer basis, the members of the Certification Committee began developing a prototype of an assessment tool that could eventually become the mechanism for certifying medical interpreters in the state of Massachusetts.

Prior to the development of the prototype instrument, the MMIA Certification Committee seriously deliberated on the purposes of the certification process, the nature of the skills and tasks that would be assessed, the principles of assessment on which the tool would be based, and the methodologies and formats that would be most conducive to measuring what an interpreter should know and be able to do.

---

[2] This section draws from the following publications: The MIAC Candidate Manual, 2001 and Avery, Maria-Paz B. and Berinstein, Eduardo. The Massachusetts Medical Interpreters Association's Efforts to Achieve Medical Interpreter Certification. ATA Chronicle. February 2001.

Intended Purposes of the Medical Interpreting Assessment for Certification
The Medical Interpreting Assessment for Certification was designed to meet the following intended purposes:

1. To establish a mechanism by which to determine whether a candidate for certification in medical interpreting has the necessary basic, entry level skills and knowledge to function competently according to the standards and ethical practice of the profession.

2. To provide those who use interpreter services with a standard of quality they can expect from candidates who successfully complete the Medical Interpreting Assessment for Certification.

3. To provide interpreters with an assessment of their performance and the areas in which they may need to continue their professional development.

Principles of Assessment
The MMIA Certification Committee committed itself to creating an assessment system that would provide candidates for certification in medical interpreting with rigorous and yet fair opportunities to demonstrate what they know and are able to do. The committee was not interested in perpetuating the tradition of testing systems that serve as "hostile gatekeepers," excluding potential candidates even before they are given the opportunity to demonstrate the essential skills of the profession. Many certification tests prejudicially limit the entrance of many candidates on the basis of peripheral requirements, skills, and methodologies that preclude them from being able to demonstrate their proficiency in the actual skills of the profession in question. For this reason, the committee deliberately chose to create an assessment tool built on principles of authentic, criterion-referenced assessment.

Authentic assessments are based on meaningful tasks that are directly related to the desired outcomes identified as essential in the "subject domain," in this case the profession of medical interpreting. Assessments are "authentic" to the extent that the tasks are true to the concepts, knowledge, and skills of the discipline and based on real-world contexts. Authentic assessments not only provide evidence of the level of proficiency (or achievement) of the test taker but also reveal to the test taker what the actual standards and challenges are in the field. In other words, the outcomes of the test are evaluated against standards of proficiency established by the profession, in this case the MMIA Medical Interpreting Standards of Practice. Each test taker is assessed against the same standards (criterion-referenced tests) rather than against each other (norm-referenced tests). Such assessments provide the test taker with opportunities to show "evidence of knowing."

In developing the Medical Interpreting Assessment for Certification (MIAC), the MMIA Certification Committee was guided by the following principles of authentic assessment:

1. The skills and knowledge assessed reflect what competent members in the field identify as essential to the profession. With the exception of sign language interpreting, interpretation is a skill of spoken language. It is the ability to convert, orally, a spoken message in the source language into a target language, accurately and completely. It is this skill that is at the core of the profession.

   This principle was met by basing the MIAC on the Medical Interpreting Standards of

Practice that were officially adopted by the MMIA in 1996.[3] These standards were extensively reviewed by experts in the field both within Massachusetts and nationally.

2. The assessment uses a "standards model." In other words, there are clear and publicly defined standards of what the candidate should know and be able to do (content standards) and clearly described criteria for what counts as good performance (performance standards).

   These standards, both content and performance, are clearly presented in the Medical Interpreting Standards of Practice.

3. The assessment formats and methodologies do not interfere with candidates' ability to demonstrate what they know and can do in the areas of skill and knowledge required for competent performance in the field. Ways of demonstrating knowledge should not be so culture bound that candidates from other cultures are unsuccessful only because of the form in which the demonstration occurs.

   This principle is being tested through the inclusion of more than one methodology to test the same area of knowledge and skill. The pilot will help determine which methodologies might be most appropriate and create the least interference. However, this question will not be fully answered until the prototype is tested with additional groups from other cultural and linguistic backgrounds.

4. Attention is given to equity issues. In other words, the knowledge and skills to be tested are made clear to all. The criteria used in the assessment are visible and transparent and not hidden from the candidate. The resources to learn and to understand what is to be known must be available to potential candidates. This includes training and educational opportunities as well as access to review materials.

   Massachusetts is fortunate in that there is are multiple training and educational programs throughout the state. In addition, most of these training programs use the Medical Interpreting Standards of Practice as the foundation of the curriculum.

5. The design and structure of the test must be flexible enough that it can be used in a comparable way across all cultural-linguistic groups. While the standards and performance expectations will remain the same, the tool should be able to reflect linguistic and cultural variations (e.g. knowledge of specific diseases or syndromes prevalent in that group; length of time required to render a conversion).

   The design and structure of the current prototype was developed with this principle in mind. However, whether the current prototype meets this principle cannot be determined until it is tested with additional groups from other cultural and linguistic backgrounds.
6. The assessment tool should stand up to standards of validity and reliability in its construction, administration, and scoring.

---

[3] Massachusetts Medical Interpreters Association and Education Development Center, Inc. Medical Interpreting Standards of Practice. Newton, MA: Education Development Center, Inc. 1998 (third printing).

One of the major questions in any test development is the question of validity. Does the test measure what it is intended to measure and not some other skill or concept? And, do the scores represent true differences among the characteristics we are trying to measure?

With respect to the first question, the pre-pilot prototype was examined for face or content validity by presenting the design and content of the instrument to experienced professionals in the field. These presentations clearly specified the areas of knowledge and skill that were being measured in each module of the instrument, the format through which they were going to be tested and examples of items in each module. Presentations were made to the following groups: the MMIA Board of Directors, the MMIA membership, the coalition of Interpreter Service Coordinators of the Greater Boston area, and the Board and committee members of the National Council on Interpreting in Health Care. The same information was also shared at national and international conferences such as the Critical Link International Conference on Community Interpreting held in Montreal in 2001, and the Third National Conference on Quality Health Care and Culturally Diverse Populations held in Chicago in 2002.

With respect to reliability, this pilot is designed to test the reliability of the administration and especially the scoring system that has been developed.

7. Attention will be given to consequential validity. The assessment instrument will be analyzed in terms of the social consequences of the outcomes of the assessment or consequential validity. Assessments do not meet the criterion of consequential validity when competent candidates are excluded not on the basis of their proficiency but because of the inappropriateness of the tools of assessment. For example, an assessment tool for spoken language interpreting that excludes candidates solely on the basis of failure on a written test without giving that candidate the opportunity to demonstrate their ability to perform the essential skill of oral interpretation does not have consequential validity.

   This principle is being met by including two modules that focus on the essential skill of oral interpretation. The pilot will help determine whether the module that uses a less expensive methodology can be used to predict success on the other module that, while more costly to administer, is also more “authentic.”

The Medical Interpreting Assessment for Certification Prototype Instrument

In designing the MIAC, the Certification Committee chose to focus only on the basic knowledge and skills that a competent, entry level, oral language medical interpreter should have. This focus on the skills of oral language interpreting was deliberate. The Certification Committee wanted the MIAC to test only the fundamental skills of oral language interpreting so that potential candidates who had not as yet mastered other related skills could be recognized for what they could do well. As a result, the MIAC does not test a number of skills such as sight translation or written translation of documents that medical interpreters are often asked to perform. It also does not test knowledge and skills in the area of mental health interpreting. Members of the Certification Committee felt strongly that mental health interpreting is a specialty skill that requires additional knowledge and training and therefore should be tested separately.

The Medical Interpreting Assessment for Certification prototype consists of a set of modules,

each of which is designed to assess specific areas of knowledge and skill that a competent entry level medical interpreter is expected to know and be able to do. Candidates for certification are expected to demonstrate basic competence in the following areas:

- Knowledge of basic human anatomy, medical terminology, and health care vocabulary in English and the non-English language (L2).
- The ability to convert oral messages accurately and completely from English into L2 and from L2 into English.
- Understanding of ethical and cultural issues in medical interpreting and the ability to make informed judgments based on such knowledge.
- Proficiency in integrating the skills of oral message conversion and ancillary interpreting skills that support the maintenance of accuracy and completeness and the goal of making possible the communication between a patient and a health care provider who do not speak the same language.

The prototype was developed in the pair languages of English and Spanish. Spanish was selected as the second language of testing principally because it is a language of high incidence in Massachusetts and many other parts of the United States. Spanish was also the language that was shared by the majority of the MMIA Certification Committee members. It was important for the developers of the assessment instrument to be able to assess the language of the prototype in order to ensure content validity and reliability in scoring.

Modules of the Medical Interpreting Assessment for Certification Prototype

The Medical Interpreting Assessment for Certification prototype consists of a set of modules, each designed to measure and assess the specific areas of knowledge and skills outlined above. The four modules are:

1. Basic Anatomy and Medical Vocabulary
2. Linguistic Conversion and Health Care Vocabulary
3. Ethical and Cultural Issues.
4. Integrated Interpreting Skills

The following is a detailed description of each module.

<u>Module 1: Basic Anatomy and Medical Vocabulary</u>

The purpose of Module 1 was to assess the candidate's knowledge of basic anatomy and medical terminology. It consisted of two parts. The first part focused on knowledge of the names of the principle parts in the major body systems in both English and Spanish and of the location of each of these body parts. The second part focused on knowledge of medical terminology relevant to these major body systems such as major medical conditions or diseases, specialties and specialists, commonly used diagnostic procedures and tests and major interventions and treatments.

The major body systems were identified as: the human body (front and back), the circulatory system (body and heart), the digestive system, parts of the eye and ear, the endocrine system, the reproductive system (male and female), the lymphatic system, the nervous system, the respiratory system, the skeletal system (front and back), the urinary system (male and female), and the skin.

The methodologies used in Module 1 were what are commonly known as paper-and pencil tests. Candidates had to perform three tasks: 1) provide the Spanish equivalents for a list of major

body parts in specific body systems; 2) label a diagram of the major body parts in a specific body system; and 3) match medical terms with their definitions.

Scoring was based on simple right and wrong answers.

Module 2: Ethical and Cultural Issues

The purpose of Module 2 was to assess the candidate's knowledge of the MMIA Standards of Practice related to ethical and cultural issues. This module was designed to determine whether or not the candidate could apply the relevant standards when faced with an ethical or cultural dilemma and more importantly, could offer a rationale based on the standards for the actions they proposed were most appropriate in that given situation.

Module 2, like Module 1, was designed as a paper-and pencil test**.** Two formats were tested in the pilot: 1) a modified multiple choice format in which an ethical or cultural dilemma was presented and response choices offered; however, in addition to the response choices, candidates were also asked to provide a rationale for that choice; and 2) open-ended questions to scenarios that described an ethical or cultural dilemma.

Scoring guides were created for both the multiple choice and open-ended question formats.

Module 3: Basic Conversion Skill and Health Care Vocabulary

The primary purpose of Module 3 was to assess the candidate's ability to convert messages in the language pair being tested (from English to Spanish and from Spanish to English) accurately and completely. Secondarily, this module also continued to assess the candidate's knowledge of health care vocabulary, including vocabulary commonly used by Spanish-speaking patients.

The methodology used was an oral test of the ability to convert a message in one language into its equivalent in another language (Sentence Conversion). This was a timed test in which the candidate heard an oral stimulus (i.e., a sentence or series of sentences) in one language and then had to provide the equivalent message converted into the other language within a given period of time. The candidate was not allowed to ask for repetitions or pauses.

For the purposes of scoring, the sentences were coded into units of meaning. By coding the sentences into units of meaning, scorers had to listen for equivalencies of meaning rather than word-for-word conversion.

Units of meaning were scored in three ways: mistakes, omissions, and additions. Mistakes were defined as conversions that had a different meaning than that of the original language. Omissions were defined as units of meaning that were in the original but not in the converted message. Additions were defined as units of meaning that were not in the original but appeared in the converted message.

Module 4: Integrated Interpreting Skills

The purpose of Module 4 was to allow the candidate to demonstrate performance of the many tasks and skills required of a competent entry-level medical interpreter in a simulation of a real provider-patient interaction. This module was designed to assess not only the skill of message conversion but also the ability to use the appropriate auxiliary skills during the course of an interpreter-mediated clinical encounter to maintain accuracy and completeness in the conversion

and to promote the goal of communication across language and cultural barriers between the provider and the patient.

The methodology used was the role-play. Role-plays provided a format through which the candidate could demonstrate the ability to perform major interpreter tasks as identified in the MMIA Standards of Practice. Thus, in addition to the scripted dialogues, the role-plays also included situations that are often found in real encounters such as inappropriate expectations on the part of the provider and/or patient, highly technical terminology that most entry level interpreters would not be likely to know, and lengthy complex messages. The candidate was asked to function in the role of the interpreter as if the role-plays were "real life" encounters using whatever tools/techniques s/he would normally use to ensure accuracy and completeness, including but not necessarily limited to asking for clarification, taking notes, or asking the speaker to pause.

The role-plays were scored in two ways: 1) accuracy and completeness of the interpretation, and 2) performance on key standards in the MMIA standards of practice focusing on the principle auxiliary skills that support accuracy and completeness and the role of the interpreter. As in the sentence conversions, the scoring for accuracy and completeness was based on "units of meaning." Performance on key standards of the MMIA standards of practice was based on a set of rubrics that described the quality of the performance on each of the standards that was being measured. Rubrics are scoring tools that assist in clarifying and making transparent how performance on a task will be judged. Three levels of performance were used: unsatisfactory, basic and proficient.

Overall, then, this was the design of the testing instrument used in the pre-pilot and in the later pilot of the certification process. Although adjustments were made based on the experience of the pre-pilot, the basic design remained the same.

A comment should be made here about why the committee chose to use a written testing methodology for certain modules (Modules 1 and 2) even though it was not the committee's purpose to test reading and writing skills. Written testing is simply much less expensive than oral testing. As this first test was to be a pilot, the committee was interested in seeing whether a candidate's performance on the written tests (the less authentic but less expensive tests) could be shown to be predictive of the candidate's performance on the oral test (the more authentic and more expensive test). As such, a written form was used for these sections, subject to evaluation at the end of the testing pilot. At the same time, the MMIA Certification Committee designed Modules 1 and 2 in such a way that accommodations could be made in the administration of the test for those candidates who had problems reading and writing in English. For example, Module 2 that tests for knowledge of ethical and cultural issues could be administered orally or in writing in the candidate's non-English language.

The Pre-Pilot

In 2001, the first version of the MIAC prototype was pre-piloted in Massachusetts. The purposes of the pilot were threefold: 1) to compare different formats and methodologies to see which provided the most valid information; 2) to determine the reliability of the instrument, in particular with respect to its administration and coding; and 3) to determine whether any of the modules could be used to screen candidates prior to administering the integrated skills module

which, in this instrument, is viewed as the most "authentic" measure of a candidate's performance as a medical interpreter.

First, the MMIA Certification Committee wanted to make sure that the methodologies and formats used in the different modules authentically and validly measured what they were intended to measure. Beyond this, the committee wanted to make sure that the methodologies and formats were fair and equitable for potential candidates from a diversity of cultures and educational experiences. The committee wanted to ensure that the ways of demonstrating knowledge and skills used in the assessment did not serve as barriers to potential candidates. It is for this reason that some of the modules used more than one methodology. The MMIA Certification Committee was especially concerned with the formats that relied on paper-and-pencil tests, formats that also tended to rely on reading and writing skills in English.

Second, the reliability of any instrument rests in large part on the reliability of the administration and the scoring system. With respect to the administration of the test, the biggest concern had to do with the consistency of administration of the role-plays. There was a concern that inconsistency among the role-players would result in some candidates having an unequal advantage over others. In terms of scoring, there was interest in whether a greater definition of what an "equivalent conversion for a unit of meaning" meant in order to obtain adequate inter-coder reliability that would ensure consistency of scores across raters and respondents.

Finally, a major issue in the development of any certification process is the cost of testing. The cost of administering a test increases when authentic forms of assessment are used, since these are usually more costly both to administer and to score. It was hoped that the pilot would provide information on the value of the modules that were less costly to administer in predicting success on the more authentic but more costly module: the integrated skills module that employed role-plays. If some predictive association were found, future test administrators could use the less costly modules to screen out those candidates who were unlikely to be successful in the role-plays. At the same time, the less costly modules would provide candidates with feedback on the skills they needed to work on to successfully complete the MIAC.

Focusing on these purposes, the volunteer Certification Committee did its best to attract volunteer candidates to take the test. Unfortunately, due to the very limited resources of the MMIA Certification Committee, the number of volunteers who completed all sections of the test was too small to conduct any statistical analysis of the properties of the tool. However, several important lessons were learned through the pre-pilot.

1. General validity of the prototype instrument.
   The feedback received from experienced professionals was that, on the whole, the instrument had face validity. The general consensus seemed to be that the MIAC prototype was measuring the relevant areas of knowledge and skill and that, taken as a whole, the instrument was a fair yet challenging test of what an entry level medical interpreter should know and be able to do.

   The only module that was repeatedly questioned was the section that purported to test language proficiency in the interpreter's language pair through the method of "shadowing." Shadowing is defined in this context as the process of repeating, almost simultaneously, everything that is said in the same language as it is being heard. This

methodology is premised on the belief that unless a person understands the syntactical and lexical structure of the language in question, it would be impossible to repeat what is said with a high degree of accuracy. The module was questioned on two points: 1) "shadowing" is not an adequate measure of language proficiency; and 2) it taps a skill that is more appropriate to simultaneous interpreting than to consecutive interpreting which is the preferred mode for medical interpreters in Massachusetts.

Objections to this module were noted but the pre-pilot, in the spirit of testing a variety of methodologies, included the module anyway. Later deliberations of the MMIA Certification Committee based on the results of the pre-pilot and recent experiences of some of its members in using shadowing as a predictor of success in training programs resulted in the conclusion that there was not adequate evidence to warrant using "shadowing" as a measure of language proficiency. It also became evident that candidates who did not have adequate command of both languages did very poorly in the oral sections of the instrument, making the shadowing test of language proficiency superfluous.

The MMIA Certification Committee came to the conclusion that the certification process should not be in the business of testing for language proficiency. Rather, it assumed that a candidate who did not have adequate command of the structure, syntax, and lexicon of both languages would not do well on the MIAC. As a result, the language proficiency module was eliminated from the pilot version of the MIAC.

2. Consistency in the administration of the test was crucial.
   Training of the test administrators, in particular the integrated skills module that uses the methodology of the role-play, was crucial. Again because of limited resources, training of test administrators for the pre-pilot was minimal. As a result, the role-plays were not administered consistently. Some role-players followed the instructions not to pause until their section of the dialogue ended, unless requested by the candidate. These instructions were deliberately set to see whether the candidate was able to demonstrate several skills having to do with managing the flow of communication and maintaining accuracy and completeness. However, other role-players paused a lot, giving their candidates an advantage of interpreting relatively short segments and preventing the candidate from demonstrating mastery of other skills needed to maintain accuracy and completeness.

3. Reliability in scoring can be achieved only with rigorous training of the coders. Prior to the pre-pilot, the members of the MMIA Certification Committee developed the outlines of a scoring system for the oral sections of the assessment instrument. They then spent a day using videotaped role-plays to calibrate their scoring. Many hours were spent discussing differences in how individual units of meaning had been scored and coming to consensus on what the score should be.

   Still, when the actual scoring was done, a comparison of the results showed significant discrepancies between coders. While the overall scores given by two coders were often similar, the inter-coder reliability on specific items was not very high. In other words, the two coders gave the same candidate a similar overall score but the rating of individual units of meaning often did not correspond. The same unit of meaning was scored as "incorrect" by one coder and "correct" by the other.

Based on the results of the pre-pilot, the MMIA Certification Committee revised the prototype and developed a second version. This second version was used in the current pilot of the MIAC.

The Pilot of the MIAC

It was 2002 when the National Council on Interpreting in Health Care approached the MMIA about funding a more complete pilot of MMIA's prototype certification, in collaboration with the California Healthcare Interpreters Association (CHIA). By this time, the small pre-pilot had been completed although with insufficient numbers to establish reliability, and the MMIA Certification Committee had made decisions about how to revise the original prototype. When NCIHC approached MMIA and CHIA about a collaborative pilot, both organizations were receptive.

The purpose of the collaborative pilot

Each of the organizations involved in this pilot had its own purposes for doing so. As mentioned previously, the NCIHC was interested in creating a wider base of national experience in interpreter certification in order to inform a future process to develop national certification. NCIHC was also interested in supporting the work of regional interpreter associations and in facilitating working relationships among such organizations.

At the time of this pilot, CHIA had recently published its own standards of practice and was focused on disseminating them. While CHIA had no immediate plans to begin certifying interpreters in California, the organization was interested in gaining experience with testing and in building on the extensive work that MMIA had already invested in its certification tool.

As part of its collaboration, CHIA partnered with Healthy House, a non-profit organization in the Central Valley of California with a long history of training and testing interpreters for health care. The key players from Healthy House were acutely interested in comparing their experiences with those of the MMIA, as well as gaining more experience with this sort of testing.

For MMIA, fully intent on implementing certification in the state of Massachusetts, this collaboration represented an opportunity to re-pilot the test instrument with greater financial support and a larger, more diverse pool of candidates, assuring the numbers necessary to probe the validity and reliability of the test.

All three organizations, then, had compatible if not identical reasons for working together to pilot this certification process.

Building a coalition

The first step in building this partnership was a meeting convened in Boston in February 2002 and attended by the key leadership of the three organizations.[4] At this meeting, each organization clarified its interest in the project and outlined what it considered its role might be. Specific tasks were enumerated and assigned and a general timeline developed. The meeting ended with a historic commitment among these three organizations to implement this project together. This commitment was subsequently formalized in a contract between NCIHC and MMIA and another between NCIHC and CHIA, outlining the scope of work that each organization would undertake and the funding that would be disbursed to underwrite the work.

Revising the instrument

The first task facing MMIA was to create a new version of the test itself, based on the same principles as the test that had been used in the pre-pilot. This was necessary because many of the same candidates might take the test, and those who did so would have an advantage over others if the same version of the test were used. The Certification Committee of MMIA also revised the preparation materials for test candidates and developed a Test Administrator and Coder Manual to use in training those who would be giving and correcting the test.

Training of Administrators and Coders

The training of administrators and coders was central to the success of the MIAC as a valid and reliable instrument. To this end, highly experienced interpreters in Massachusetts and California were recruited through their respective associations to participate as administrators and coders. The following criteria were used to identify this cadre of interpreters:

1) three or more years of experience as a paid medical interpreter;
2) respected by others in the field as a competent medical interpreter;
3) experience in training and/or supervising medical interpreters; and
4) commitment to, and passion for, the development of the profession.

In California, an additional criterion was sensitivity to language issues among heritage speakers, although all administrators/coders were expected to be aware of regional differences.

Ten interpreters were trained as administrators/coders in Massachusetts and thirteen in California. All except one of the California administrators/coders were native Spanish-speakers, representing a variety of Spanish-speaking countries from Europe, the Caribbean, and Central and South America.

All 23 administrators/coders participated in a two-day training conducted by one of the MIAC developers from Massachusetts. Two training sessions were held, one in Massachusetts and one in California. The training consisted of an introduction to the instrument including the rationale for the content and formats for each module, the administration of the Module 4: Integrated Skills (or rather, the role-plays) and the scoring for those sections and modules that required judgment on the part of the coder (Module 2: Ethical and Cultural issues, Module 3: Message

[4] Attendees included:
NCIHC: Cynthia E. Roat (Co-chair of the Board), Karin Ruschke and Linda Haffner (Co-chairs of the Standards, Training and Certification Committee)
MMIA: John Nickrosz (President of MMIA), Jane Kontrimas (Chair of the Certification Committee), Maria-Paz Avery and Eduardo Berinstein (Members of the Certification Committee)
CHIA: Beverly Treumann (President of CHIA), Niels Agger-Gupta (Executive Director of CHIA), Elizabeth Nguyen and Ann Chun (Co-chairs of the Standards and Certification Committee)
Facilitator: Joy Connell

Conversion, and Module 4: Integrated Skills). Instructions for the administration of the sections other than the role play (Module 1: Anatomy and Medical Terminology, Module 2: Message Conversion, and Module 3: Ethical and Cultural Issues) were conveyed through an administrator's manual. The administrator's manual described in detail the process for administering each section, the personnel or staffing requirements, and the tasks for each person involved in administering each section of the test.

In Massachusetts, the two-day training totaled approximately ten hours but an additional 3 hours of training were provided to a subgroup on the scoring of Module 2: Ethical and Cultural Issues. In California, the two-day training totaled approximately 12 hours.

In both settings, the primary focus of training concentrated on the scoring of the oral portions of the instrument.

Administration of the Pilot

The pilot version of the MIAC prototype was administered in mid 2003 to 37 participants in Massachusetts and 46 in California. Participants in the pilot were recruited to represent different levels of experience, employment status, and training. To this end, participants were asked to complete a background questionnaire designed to elicit information on their employment history as a medical interpreter including volunteer work, the type and amount of specific training or education in spoken language medical interpreting they had received, and the estimated number of interpreted encounters they had performed. The objective was to be able to identify participants who represented three levels of experience and proficiency: beginner, intermediate, and advanced.

The administration of the MIAC was divided into two major sections: the Screening section and the Integrated Skills simulation.

The Screening section consisted of Modules 1, 2, and 3. In the design of the MIAC, these three modules were intended to serve as tools to screen out those candidates who could not demonstrate basic knowledge of the terminology, the standards of practice relevant to ethical and cultural issues, and the basic skill of linguistic conversion. All three of these modules were administered on one day.

The Integrated Skills simulation (Module 4) consisted of two role-plays. Role-plays provide a relatively authentic assessment of how the candidate might perform in a real situation. However, role-plays are more expensive and labor-intensive to administer. Therefore, one of the purposes of the pilot was to determine whether the screening modules, which were less authentic but also less expensive to administer, could be used to predict the likelihood of success in the role-plays, which were both more authentic and more costly. To be able to make this determination, participants in the pilot took all the modules. Module 4 was administered on a separate day.

Although the intent was to administer the different modules of the test in a consistent manner, there were some variations in administration between Massachusetts and California. For example, in Massachusetts, Modules 1 and 2, the written or paper-and-pencil parts of the test, were administered without any time limit. Participants took as much time as they needed to complete these modules. The maximum time required was approximately two hours. In California, however, Modules 1 and 2 were timed. Participants were given an hour to complete

these two modules. As a result, many of the participants were unable to complete Module 2. This resulted in the loss of comparable data across the sites.
The administration of Module 3: Linguistic Conversion also differed. In Massachusetts, Module 3 was administered in a language laboratory. This meant that the amount of time that the candidate was given to render the linguistic conversion into the other language was controlled by the administrator and not by the candidate. Each candidate, therefore, had exactly the same amount of time in which to produce a linguistic conversion. In California, participants used two tape recorders, one to play the stimulus and the other to record their response. This meant that the participant could control the amount of time available to make the conversion. In a number of cases, the coders noted that the candidates were able to correct their first conversion. Thus, California participants had an advantage in generating a correct response.

Statistical Analysis
Prior to entering the data and conducting the statistical analysis, all scoring sheets were reviewed for consistency and completeness. Various discrepancies were noted:

1. Missing data: Missing data was the result of a number of factors: unrecorded responses, unread sentences or phrases in the role plays, and unscored or ambiguously scored units.
2. Incomplete test results: A few of the participants did not complete all sections of the instrument.
3. Scoring by only one coder: In order to determine the inter-coder reliability of the scoring system, all test results had to be scored by two coders. In a few cases this did not happen.

Priority was given to consistency and completeness in the scoring for Modules 3 and 4 (Linguistic Conversion and Integrated Skills). By the time the data was cleaned, close to half of the cases in Massachusetts were eliminated from the statistical analysis and slightly over half in California. As a result only a total of 42 cases were included in the statistical analysis, 20 from Massachusetts and 22 from California.

Statistical analysis was performed only on Modules 3 and 4. These were the only modules that provided consistent and reliable scoring data. Statistical analysis was conducted to determine inter-coder reliability on the linguistic conversion of utterances in both modules. In addition, a comparison of inter-coder reliability between Massachusetts and California coders was performed. Statistical analysis was also conducted to determine whether Module 3: Linguistic Conversion served to predict performance on Module 4: Integrated Skills. Finally, regression analysis was conducted to determine whether or not there was any relationship between participant characteristics (i.e., employment status, number of estimated interpreting encounters, and type of training) and performance on Module 4.

Results of the Statistical Analysis
1. Inter-coder Reliability
   The interpreted responses of each participant on Module 3: Linguistic Conversions and Module 4: Integrated Skills were scored by two coders. To determine inter-coder reliability the scores given by each coder were analyzed in two ways. First, the percentage of overall correct answers given by each coder was computed. Second, the percent agreement on each unit of meaning was computed.

The results showed that there was considerable variability in inter-coder agreement. Based on the percentage of overall correct answers given by each coder on the English to Spanish sentence conversions, only 64 percent of the coder pairs had a difference in scores that was ten points or less apart. The range of differences in scores between pairs went from a low of 0 points to a high of 27 points. On the Spanish to English sentence conversions, the correspondence was higher. Seventy-six percent of the pairs had a difference in scores that was ten points or less apart. The range of difference in scores went from a low of zero points to a high of 29 points.

However, when this data was disaggregated by site, it was found that Massachusetts pairs had a higher percentage of agreement than California pairs – 80 percent for Massachusetts versus 50 percent for California in the English to Spanish conversions and 80 percent for Massachusetts versus 76 percent for California in the Spanish to English conversions.

These trends were corroborated in the computation of agreement on each unit of meaning for the Sentence Conversions. Sixty-two percent of the coder pairs had inter-coder reliability of .80 or higher in the English to Spanish sentence conversions and 86 percent had inter-coder reliability of .80 or higher in the Spanish to English sentence conversions. When disaggregated by site, a t-test showed that Massachusetts coders had significantly better inter-coder agreement than California coders

On the role-plays, 76 percent of the coder pairs had inter-coder reliability of .80 or higher on Role Play 1 and 80 percent had inter-coder reliability of .80 or higher on Role Play 2. Again, disaggregation of the data by site indicated that Massachusetts had higher inter-coder reliability than California. The t-test difference was statistically significant for Role Play 2 and although the difference did not reach statistical significance for Role Play 1, the trend was maintained.

2. Sentence conversion as a predictor of success in the role-plays
   Scores on the sentence conversion were analyzed using a Pearson r to determine whether they could be used to predict success in the role plays. A Person r of .744 existed between the Spanish to English sentence conversion scores and the scores on Role Play 1. This indicates that the Spanish to English sentence conversion was a very good predictor of performance on Role Play 1. This was the only correlation that was performed due to technical difficulties with the size of the variables in Role-Play 2.

3. Relation of participant characteristics and performance on the role-plays
   A multiple regression was done to determine the relationship between the participant characteristics of employment status, number of estimated interpretations, the type of training received, and the results of the certification test. No relationships were found. However, it should be noted that the measures used for these characteristics were not very reliable. For example, using number of hours of training is really not an appropriate measure of the quality of training received. Without knowing the content and methods used, two training programs could provide the same number of training hours but the quality could be very different. This finding suggests a correlate need in the field: the need to establish standards for training programs.

Discussion of the Results

It appears that the scoring system for the oral sections of the assessment instrument has the potential for greater inter-coder reliability. The fact that Massachusetts coding pairs were able to obtain a relatively high degree of inter-coder reliability indicates that with more intensive training, a high rate of inter-coder reliability is possible. It should be noted that this assessment instrument is based on the role of the medical interpreter and the standards of practice as contained in the Massachusetts Medical Interpreting Standards of Practice. The cadre of Massachusetts coders had had more exposure and more practical use of these standards than the cadre of California coders. Massachusetts is also a smaller community. The cadre of coders in Massachusetts for the most part had also worked with each other professionally for some time. California, on the other hand, is a large state and the cadre of coders came from different parts of California. Many of the coders had had little opportunity to work together previously.

It should be noted that these statistical results in inter-coder reliability masked an important observation: that certain coding pairs were more discrepant than others. Are some coders consistently much more lenient or much more rigorous than others? Do some coders fail to understand the coding system, making them less likely to approximate the true score of the test taker?

With respect to the predictive value of the Spanish to English conversions in relation to success in the role-plays, this finding could be explained by the fact that all the participants who took the test were native Spanish speakers. It is usually more difficult to go from the dominant (in this case, native) language into the second (or non-native) language. Therefore, the fact that the participants did well converting from their native language to their second language was a measure of their bilingual fluency.

Lessons and Recommendations

This pilot of the MMIA medical interpreter certification process has taught us valuable lessons about the test itself, the testing process and also the process of collaboration among organizations located in different parts of the United States.

Lessons learned about the pilot test and the testing process

This pilot still has not answered all the questions about the properties of the MIAC. However, significant findings allow us to make recommendations that could lead to the implementation of a formal certification process in English-Spanish for Massachusetts in the near future.

1. Scoring reliability
   As was indicated earlier, much of the reliability of this tool rests on the reliability of scoring. However, scoring of the message conversions is a formidable task that requires intensive training to achieve a high level of inter-coder reliability. It is often difficult to provider coders with the exact instructions as to what constitutes a “correct” interpretation, since there are many ways of saying the same thing even in the same language. In medical interpreting, when the provider and patient do not share the same cultural frame of reference, there is also the dilemma that some words are “untranslatable.” In these cases, not only is there no equivalent word in the target language, but the concept itself may be one that is outside the frame of reference of the

listener. What the interpreter does with this situation is often a matter of judgment. The challenge in scoring is to judge the variation in responses consistently and reliably across the test takers and across the raters.

Other steps need to be taken to increase inter-coder reliability. First of all, there is a need for clear and appropriate criteria for choosing administrators and coders. During the pilot, none of the individuals who were recruited as coders was denied the opportunity to participate in the actual coding, regardless of their performance at the training. In the future, the training of individuals as coders should also serve as a screening opportunity. Only those who achieve an adequate level of inter-coder reliability should be accepted as coders.

Secondly, this pilot clearly showed that a two-day training is not sufficient to cover all the aspects of scoring involved in the MIAC. Administrators and coders need more extensive training, including a more detailed instruction manual and more time devoted to practice coding under supervision. Separating the administrators and coders also could provide improvement in the accuracy of test scores.

2. Consistency of administration
   Each module of the instrument needs to be administered in a consistent manner across all sites and times. Inconsistency of administration puts into question the validity and reliability of the instrument. Individuals responsible for the administration of each module need to be very cognizant of the instructions and follow them closely. Otherwise, some participants may be given an advantage while others may be obstructed.

   Achieving consistency in the administration of Modules 1, 2, and 3 is fairly easy to remedy. In each of these modules there are technical ways of ensuring consistency. The biggest difficulty arises with Module 4, which makes use of role-plays. Although the role-plays were scripted, they also attempted to reflect normal interactions and dynamics found in many clinical encounters, thus providing the candidate with the opportunity to demonstrate use of key auxiliary skills.

   Given all this, delivering the role-plays in a consistent manner across different role players and at different times is not an easy task. In the pilot, the following variations in delivery were noted:

   - The pace of delivery: Some of the role players did not speak at a natural rate. Instead, they paused at regular intervals which resulted in giving the interpreter cues as to when to begin interpreting. Frequent pauses on the part of the role players also did not allow candidates to demonstrate how they would apply such skills as managing the flow of communication, asking for clarification, or note taking to assist in maintaining accuracy and completeness.
   - Consistency in responding to candidate initiated questions: The role-plays were written to include medical or cultural information that was likely to be "unknown" to the interpreter. The role-play scripts, however, did not provide adequate scripted responses to such queries, resulting in variation in the stimulus across test situations.

The problem of consistency in administering the role-plays can be remedied in two ways. One way is to train a small cadre of role-players by having them practice under observation until their delivery is consistent over time. Another way is to use actors in the same manner they have been used to serve as "standard patients." Actors already have the training to keep their performances of the same role consistent. This option, while the ideal, is too costly at this point in time.

3. The use of screening modules
   The written modules of the instrument (Modules 1 and 2) did not show any value in predicting success in Module 4, the role plays. They do, however, measure basic knowledge that all entry-level competent interpreters should have. Therefore, they should be included as part of the screening section of the tool. Candidates who do not have this basic knowledge should not be certified as interpreters.

   As discussed earlier, however, the concern about consequential validity remains: whether the written parts of the instrument test only knowledge of medical terminology and ethics, or also the candidate's skills in reading and writing English – a skill we do not wish to test. The MMIA Certification Committee has concluded that issues of cost dictate the need to use written tests despite this concern, however, the option of taking such tests orally should be available to interpreters who find the written nature of the test a barrier. Another option could be to allow candidates to write their responses in their native language where appropriate, such as in Module 2.

Lessons learned about collaboration

Follow-up interviews with the key players in the MMIA, Healthy House, CHIA and the NCIHC revealed that everyone felt the collaboration had been extremely valuable. From MMIA's point of view, the process allowed them to include enough candidates to more effectively evaluate the reliability and validity of the test. For CHIA and Healthy House, the collaboration allowed them to learn important lessons about certification that they will integrate into their own certification process if and when they decide to implement one; in fact, they may choose to build on the MMIA model when designing testing for the languages common in California. The NCIHC learned more about how to facilitate inter-organizational collaboration, as well as learning the same lessons about certification.

For CHIA, there were additional unexpected benefits. Participation in this certification pilot helped provide a focus to unify CHIA members and improved CHIA's reputation as an organization that is taking serious steps toward building health care interpreting as a profession. CHIA membership has increased during the period since the pilot, possibly as a result of CHIA's participation in this project.

The road to collaboration, however, was not an easy one. There were many challenges, and many recommendations of how such a collaborative process could be done more effectively in a future.

1. Allocate more time
   In the end, this project took about 16 months from funding to finish. Even though a tremendous amount of work had already been done on the testing instrument, this timeline was too short for a project implemented almost entirely with volunteers.

2. Allocate greater funding
   All those interviewed about this collaborative process agreed that the lack of paid staff was a serious limitation on the quality of the work and the speed at which the project could advance. The funding for the project went principally for photocopying materials and for paying test administrators and coders, although even they put in much more time on the project than they could be paid for. The collaborators felt strongly that the addition of a paid project manager to handle setting up testing logistics and coordinate between the two organizations would have helped the process immensely. Additional funding would also have allowed for a second face-to-face meeting of the key collaborators in each organization to establish good working relationships and discuss logistics.

3. Better communication
   One of the greatest frustrations mentioned by all those interviewed was the breakdowns in communication that hampered the collaborative process. While the MMIA had a stable group of people working on this certification pilot, CHIA's key participants seemed to shift throughout the process, making it difficult to keep people up-to-date with discussions that had gone before. Neither group was exactly sure who to contact with any particular question. Because of this, both CHIA and MMIA participants felt that the assignment of one key contact person in each organization would have gone far toward improving communication.

   A face-to-face meeting among all those involved in the collaboration also would have improved communication. While NCIHC, MMIA and CHIA leadership did have an initial meeting in Boston to establish initial agreement about the collaboration, a second meeting should have been scheduled just for the key individuals in the organizations who were to actually implement the work. A lack of funds led the project designers to prioritize spending for other aspects of the work.

Conclusion

The most important lesson learned by all the collaborators in this certification pilot is that assessing the skills involved in medical interpreting is a complex process. The very concept of what linguistic equivalence means is one that language interpreters in all fields and in all settings still continue to debate. What does it mean to convey the same meaning of a message when a conversion is made from one linguistic system to another? What does it mean to render this conversion accurately and completely?

On the whole, however, it appears that the Medical Interpreting Assessment for Certification prototype is a valid instrument that assesses the major areas of knowledge and skills that a competent entry-level medical interpreter should have. While the use of the English-Spanish prototype as a formal certification tool is a strong possibility for Massachusetts, there are still many steps in the process that have to be strengthened and refined. Foremost is the development of an intensive training program to screen and prepare highly qualified coders and administrators for the role plays. In addition, the test specifications (i.e., clear articulation of how test items for each module and section within each module are to be created) need to be more clearly defined.

There are, also, still some unanswered questions. For example, the question of the transferability of this prototype to other language pairs is still unanswered. We still do not know whether or not the formats and methodologies used in the English-Spanish version will be appropriate for

additional groups from other cultural and linguistic backgrounds. Answering this question will require the development of the instrument and its piloting in other languages, preferably several non-Western languages.

The development of a certification process in medical interpreting is one that requires a tremendous amount of resources, both human and financial. Both the pre-pilot and the pilot of the MIAC were constrained by limited resources and a relatively short timeline. This affected the consistent administration of the modules in both sites and the reliability of the scoring. Such a complex and high-stakes process cannot continue to be the responsibility of a dedicated group of people who essentially are volunteers. When and if the National Council on Interpreting in Health Care embarks on the design of a national certification process for health care interpreters, support to develop an infrastructure that will provide consistent oversight of the conceptual, technical, and logistical aspects of the process will be essential.

This pilot of a certification process for Spanish-English health care interpreters represents a major step forward in the professionalization of interpreting in health care. While certification is only one aspect of quality assurance, it is an important aspect whose time has come in some parts of the country, and for whose advent we must prepare at a national level. It is only through the dedicated and selfless work of people such as those who participated in this pilot that the lessons of this difficult and delicate task will be brought to light and disseminated. The result will be a higher standard of quality in health care interpreting, which will lead to clearer communication between patient and provider and better health care outcomes for all.