

**Lessons Affirmed and Learned  
(The MMIA Medical Interpreting Assessment for Certification Pilot; Forum on  
Meaning and Accuracy in Meaning; Forums on Certification)**

Maria-Paz Beltran Avery, Massachusetts Medical Interpreters Association and  
Education Development Center, Inc.

Let me just say that it's really a pleasure to be here. Right now my mind is reeling so much that I don't even know where to start. Everything I have heard so far has made me want to start discussing the issues and not just listening. So, I'm going to work very hard to do the presentation I have prepared and not try to start a dialogue based on what others have already presented.

I also want to thank Elizabeth for actually doing my presentation. It's really interesting to hear another voice from a different position with a different perspective of the same project.

So, today, I will share a few things with you about the MMIA Medical Interpreting Assessment for Certification Pilot. I titled my presentation "Lessons Affirmed and Learned." As I thought about our experiences in the MMIA, and my own personal experiences with this pilot as well as with the pre-session on meaning that I did at the MMIA Conference two years ago ... [And to those of you who were there, I apologize. I have not written that session up, but I still have all the notes, and eventually I hope it will be written.] and recently with the forums on national certification which I first designed for a presentation at the Quality Health Care Conference in Seattle and that are now being conducted in other places across the country, I realized that we have begun to gather a lot of data on how the nation is thinking about certification. These experiences have affirmed lessons that we already knew, but they also generated new lessons for me.

I guess I was hoping, as I was thinking about lessons learned, that maybe we were already beyond those basic lessons. But unfortunately what these experiences did for me was to affirm a lot of the things that I think you've already heard from other people. But there were also other things that came up that really were lessons learned.

Let me start with a quick summary of the development of the MMIA prototype instrument - the MMIA Medical Interpreting Assessment for Certification (MIAC). The development of this prototype was an outgrowth of the work of the Standards Committee of the MMIA that had created and published the MMIA Standards of Practice in collaboration with EDC [Education Development Center, Inc.], which is where I work and where I directed a project to develop a college-level certificate program to prepare bilingual adults as medical interpreters. We started that project with Spanish speakers and Khmai speakers in 1992. At that time there was really very little in the field of health care interpreting, and so we were creating everything practically from scratch, even though we did work a lot with people who were ASL interpreters and educators. But we felt that the context of medical interpreting was very different from where ASL interpreting was, and, even at that time, we defined our approach to interpreting very differently. ASL interpreting, at that time, was still in the black box model and we were more into an interactive, culturally-based model. Out of this program, we developed the MMIA standards of practice using the DACUM process (a job analysis process) and based the development of the certification on these standards of practice.

We developed the prototype in Spanish because that's the biggest language group that we had, and most of us on the committee were Spanish speakers. We pre-piloted a version of the prototype in Massachusetts in January 2001, and then conducted a more formal pilot with a revised version and included CHIA in this pilot. Someone mentioned earlier a cost of a million dollars for development of a certification instrument. Our budget was zero. Eventually, we got, I think, twenty-five thousand from OMH through NCIHC to do the pilot that involved CHIA. That was like peanuts. Well, peanuts are good, but you need more.

We trained scorers for the pilot but I won't go through that because you've heard from Elizabeth [Nguyen] about the two-day training we did in both California and Massachusetts. The instrument itself was administered to thirty-seven people in Massachusetts and forty-six in California over two days. The first day we administered the written sections and the oral sentence conversion section. The second day we administered the role plays.

However, the final analysis included only forty-two tests. So, out of eighty-three people who took the test, we were able to use only forty-two of those tests for statistical analysis because there was a lot of incomplete data. The major reason for the amount of incomplete data had to do with the inconsistency of administration across the two sites. Also, because of inconsistencies in administration, especially on the written part of the test, we focused the analysis only on the sentence conversion part and the role plays. And even with the sentence conversion part, we had to eliminate some of the data. In Massachusetts we administered the sentence conversion section using a language lab, which was very controlled. For California, because they did not have access to a language lab, they used tape recorders to administer the sentence conversions. We had given instructions that once the audio tape was turned on, the test taker should have left it alone – that is, not stopped it at any time. The audio tapes were timed so that the test taker had a specific amount of time in which to provide the conversion. But we could tell, sometimes, from the recordings that there were people who stopped the audio tape. Where that was obvious, we threw those results out and used only those that seemed to have followed the instructions. That is how we ended up with only 42 analyzable tests out of 83.

One of the purposes of our pilot was to assess the validity of the prototype test. Does it measure what it's intended to measure? Was it reliable – here, we focused only on inter-rater reliability. Another purpose was to try out different methodologies. So we included different and more formats for testing the same thing than we would need in an actual exam. We wanted to see what formats worked best and what didn't work for different people. We also wanted to determine whether any of the modules could be used as screeners prior to administering the role plays.

As a committee, we spent a lot of time talking about why we would want to have certification. The purposes we came up with were: to determine basic entry-level proficiency; to provide a standard of quality to the consumers of the service; and to provide interpreters with an assessment of their proficiency.

We also spent a lot of time discussing principles of assessment, and those principles that we wanted to develop the test around. Whether we were successful or not, I'm still not sure. But these were the things that we were aiming for:

- We wanted clear and public content standards. We felt that we had those because we were basing the content of the instrument on the MMIA Standards of Practice.
- We wanted to have clear and public performance standards. Again, those of you who know the MMIA Standards know that it's not just a statement of a task or a function. It's also a description of what it looks like if you are able to do the task in a masterly way and what it looks if you lacked mastery.
- We wanted to use authentic assessment methodologies. We felt that the role play was at least one way of getting to that authenticity of the assessment.
- We wanted to address issues of equity, and for us, it meant that the resources to acquire the knowledge and skills should be available to the candidate prior to taking the test. So whether that's training or education, it's got to be in place. Or if it's about the methodology and formats used in the instrument, we wanted to make sure that the methodologies and formats were accessible to all and comparable across languages while accommodating cultural and linguistic differences. We didn't want the format and methodology of the test to get in the way of the candidate being able to demonstrate what they really were able to do and knew.
- And we wanted to be able to meet the criterion of consequential validity. We wanted the process and the instrument to be designed with a concern for the social consequences of the measurement. That's the hardest piece to do without a long-term process. Consequential validity refers to the accuracy of the decisions made on the basis of the instrument.

The instrument had four sections:

- Section one: knowledge of basic human anatomy and medical terminology vocabulary for which we used paper and pencil, diagrams, equivalencies, and matching terms with definitions.
- Section two: understanding of ethical and cultural issues. We felt strongly that cultural issues and ethical issues had to be part of the assessment process. We measured the cultural issues in two ways. We addressed them in the written section by using scenarios. But we also addressed them in one of the role plays, in which we posed a cultural dilemma. We wanted to see how the candidate would react to the dilemma. This part was scored separately and not as part of the accuracy score. "How did you do on that aspect of addressing the cultural issue?"
- The third section had to do with the ability to convert oral messages accurately and completely. This was done through the sentence conversion. "Can you go from L1 to L2? Can you go from L2 to L1?"
- Finally, the fourth section integrated all the knowledge and skills into the simulation, which is the role play. And that's the part that is costly. That's what's time consuming and labor intensive, and difficult to ensure consistency in the administration. That's what's the hardest to score.

So, in terms of scoring for accuracy in the role plays and also for the oral sentence conversions, we looked at mistakes, omissions and additions, but then we simply used a "mistake" score as the overall score.

What did we find? We found that the sentence conversion section did predict how well the candidate would do on the role plays. But what was interesting about this

finding was that it was only the Spanish to English conversion score that was the good predictor of role play performance; not the English to Spanish. We can talk about why that happened later.

In terms of overall predictive validity of the test score to on-the-job performance, we obviously had no available data to test for that. [With regard to] consequential validity, we also had no data to test for that.

With regard to reliability, we focused on inter-rater reliability. Each test was scored by two independent scorers. In general, our inter-coder reliability was pretty reasonable, but when we looked at the degree of agreement between specific pairs, there was great variability, which really says that some people were better raters than others. [With regard to] inter-coder reliability on the sentence conversion, sixty-two per cent of the coder pairs had inter-coder reliability of .80 or higher in the English to Spanish sentence conversion, and eighty-six per cent had inter-coder reliability of .80 or higher in the Spanish to English. Again it's interesting to see that there's a difference in the direction of the findings. A t-test showed that Massachusetts coders had significantly better inter-coder agreement than California coders. I've already talked about some of the reasons why that might have been.

How did the test takers perform when you look at who passed the sentence conversions from English to Spanish? Who passed the sentence conversions from Spanish to English, and who passed the role plays? What I want to draw your attention to here is that five people who passed the role plays failed the sentence conversion. What that said to us was that actually our sentence conversion module was much more difficult than the role play itself. And why was that? In the role play, we were looking not just at how accurate and complete the conversions were, but we were also looking at the auxiliary skills the candidates were using to compensate for maybe not having the highest level of language proficiency, particularly in English. But they were able to use auxiliary skills like asking for a pause, a repetition, an explanation in order to maintain accuracy and completeness. On the other hand, we also looked at the manner in which they used these skills – in other words, were they *skillful in using these skills*. Let me give you an example: In the pre-pilot, I was administering the role plays, and in one instance there was this candidate who kept interrupting after every five words or so, not even at logical places to stop. I was playing the role of the provider and I found myself getting so impatient I wanted to scream, really. I know that if I were a real provider, I would have been pissed, you know? This is not who I want helping me. So, we looked at that aspect - how skillfully was a candidate able to use the auxiliary skills in order to maintain accuracy and completeness?

What are the next steps for MMIA? What we really need at this point is to develop the blueprint for the instrument based on what we learned from the pilots. What we have is a prototype. So, when people call and say, "Can we use your test?" I say "No, you can't use our test, because the test is just a prototype." What we need is a blueprint that will allow us to create comparable forms, different forms of the test to be given at different points in time, but that maintain the same level of difficulty across the forms, that have the same number of scoreable items, that address lexical issues, that address idiomatic expressions, all of those kinds of things. So what we really need is a blueprint and as part of this blueprint, we need an item bank for each of the modules that we can draw on. We also need to test the blueprint, based on this prototype, with other languages and document the modifications that might be needed, but still again keeping

in mind consistency and comparability across forms. We also have to do a better job with training and screening the administrators as well as screening the process.

So what have I affirmed? Measuring the skill of spoken language interpreting is a complex endeavor. That's my "duh" affirmation. We're still at that stage. We're still trying to get that message across to our constituents across the country. Here's another one: excellent interpreters do not necessarily make excellent administrators or scorers. These functions require different skills. We, I know, should have done a better job of screening out some of the people that we had as administrators and scorers.

Many stakeholders in the field of health care interpreting do not understand the complexities of developing a certification process, especially one that has to address multiple languages, cultures, and ways of knowing and expressing that knowledge, let alone understanding how to develop a process that is valid and reliable across many administrations and versions. We in the field of health care interpreting have to come to a common understanding of what we mean by certification. We need to get away from just describing what people are currently doing that they call certification and say, "No. These are the requirements for good certification, and that's what we need to hold people to. If they don't meet these criteria, then it's not a valid certification." In addition, as we develop rigorous, valid and reliable assessments in a limited number of languages, we also need to develop other rigorous ways of measuring and acknowledging the competencies of interpreters of other languages.

There are a lot of questions about which we have little empirical data, I think. For example, do we know what the minimum level of language proficiency is in order to achieve accuracy and intelligibility? It's good to say that we need to have the highest level of language proficiency. But that is an unrealistic demand. We need to recognize that we're not going to get that all the time. When we developed the MMIA Standards of Practice, we did it very consciously knowing that we had to pay attention to the auxiliary skills that would allow an interpreter who did not have the *highest* level of proficiency, but who had an *adequate* level of language proficiency in both languages to be able to maintain accuracy and completeness.

How do we measure accuracy of meaning, and what does it mean to have accuracy of meaning? Do we have agreement on what we are measuring when we say that? What do we mean by equivalency in meaning? Can there be accuracy without completeness? How should we establish our cut scores? Is there a *good enough* score, and if so, how do we know it is good enough?

I want to share a couple of things that came up when we had that pre-conference on meaning at the 2006 MMIA Conference. Here are some of the themes that came out from what people said or wrote. There was agreement that meaning has many levels and aspects. The three big areas that everyone mentioned relate to the content. First, what is the semantic content of that message? Next, what are the context and the purpose of that message? Context and purpose contribute to meaning. And third, what is the intent of the speaker in conveying that message? That aspect of *intent* presents a lot of difficulties in terms of measurement. Is that something that we can really measure, the intent of the person? I know for myself, and I do a lot of work as a mediator, to get two people to understand the intent of what they said to each other takes a long, long time, if we ever get to it. So should we expect interpreters to be able to capture the intent of the meaning? I don't know.

What is equivalence of meaning in an interpretation? Here is what some people said: equivalence between the message in the source language and the target language. Did the message get through? Is it likely to be understood? Is there clarity of expression? Here's one quote: "*The message presented in one language elicits in the listener the same image, message, connotations in the mind of the speaker,*" and I put that in italics because I don't really know how we do that. Can we do that, really? Or rather is equivalence of meaning fidelity to the source message? But what does that mean anyway: does it mean equivalencies in the conversion, not in the understanding? Is it fidelity in the *conversion*, which means to me that the receiver of the conversion is able to respond to the message as if they had heard it in the source language? And if misunderstanding is part of that response, that's what happens.

What did I learn? If we are to move to national certification, there's a lot of work we need to do in arriving at a shared understanding of what we are measuring, how we are measuring what we say we want to measure and what the result means. And then we need to convey all of that clearly and explicitly to those who will be seeking certification and to those who will be availing themselves of the services of certified interpreters. In addition to developing an assessment for certification that is valid and reliable, we need to have agreement on the purpose of the certification among all the stakeholders. We can't assume that we share that same purpose. The certification process is not just about the assessment tool itself, and sometimes we forget that. I know I did. And we've now had many discussions to determine the requirements that will qualify a candidate to take the assessment. This, I think, is often the hidden but major point of disagreement among stakeholders based on different beliefs about things like formal education, status, literacy, etc. Certification is a political process and we shouldn't forget that.

What else have I learned? We need better training on ethics in general. Training on the profession's code of ethics appears to be inadequate based on responses to the test questions on ethics. In our test, we asked people not only to answer multiple-choice questions on ethical issues, but also to explain why they chose a particular response and how that response related to the code of ethics. If we had scored only on that basis, we would have failed almost everyone. Although they could check the right answer, they couldn't say why that response was the ethical thing to do based on the code of ethics.

We need ongoing, focused dialogue - this is what I recommend - ongoing, focused dialogue that builds a body of knowledge about what we know, what we don't know, and the mistakes we have made. We learned more from mistakes than we learned from the successes. We need ongoing, focused dialogue to develop a common understanding of what competent interpreting looks like at different levels.

We need ongoing research. I would strongly suggest that one of the things we should agree on is that all current certification efforts should include a data gathering and research component. Most of them do, but there should be an agreement that it should be publicly shared. Research questions should be developed collaboratively and contribute to a coherent body of knowledge, not just about measuring the skill of interpreting, but also about the theory of interpreting.

We need ongoing public awareness at every single level.

We need more rigorous training. Somebody noted the difference between training and education. I agree that there's a difference between training and education. We need training of trainers and educators. It's great that RID is doing that. But we need training not just on the "how-to's," the mechanics of interpreting. Maybe that's all that we can measure. However, I think for a really fully professional interpreter, someone who can do their job really well, they not only need to know the "how-to's," but they also need to know the "why-to's." I think that the "why-to's" are part of education. And when I say education, I don't mean just formal education. I think there are other ways of educating people. We need a common core of content, of knowledge and skills, and we need standards of excellence for training and educational programs.

---

Q: You mentioned the item bank, developing an item bank. You're not thinking of that in the context of the role play part, are you? And if so, what would you have in mind?

A: Well, I mean it might be...I haven't thought it through, but you know it might be that we would have a sort of deposit of many different role plays that have been calibrated according to the blueprint. This brings up issues of security. I come from the world of K-12 education, and in statewide testing there is no way that any state would give the same test over and over again.

Q: You were just talking about developing parallel versions that have been equated according to your test specifications.

A: Yes, and you need a blueprint in order to do that.

Q: I'm concerned that the questions that you ask, which I don't hear being asked very often, get addressed during these days that we're here, specifically the ones related to how we decide on what's an equivalence and those kinds of things. Do we have any way to make sure those questions don't get lost?

A: Well, let's not forget that we're going to have a series of dialogues. And I think that we can begin to plan for those kinds of things. I don't think it's something we can answer right away quite honestly. I think the understanding comes from people talking about it rather than writing it down.

Q: I just wanted to get a clearer understanding of the sentence conversion part of the test. I see it's an oral, so do they hear a sentence and then...?

A: Yes. It actually is designed so that it starts with very simple, short, and not very dense utterances. And it progressively develops into more complex, denser, and longer utterances.

Q: So then it wasn't like a written statement that they would then convert, or like a sight translation?

A: No. This is all oral.

Q: It wasn't a sight translation?

A: No. That was a decision we made. We were talking about entry level. We were talking about the basic skill of spoken language interpreting. Further down the line we said, yes, we would -- we *should* have a module on sight translation. We should have other kinds of modules, but this was starting with the basics.

Q: As you went down your list of things that we need to do, I was reacting with, "Yeah, those are the kinds of questions that academics would demand." To what extent do you think those are important questions with respect to where we are on the practical utility of testing and legal defensibility? And I maybe know the answer in the way you sort of deflected a previous question.

A: I think in order for us to do a good job of actually measuring what we want to measure, we have to know what we mean by it. We, the people who are developing the tests, who are doing the scoring, etc, have to have a better understanding of what it means.

Q: Do you think that the published test specifications (let's say, of a court interpreter test like on the consortium website) are inadequate in terms of presenting the objectives of testing? You know, this is what we're trying to do?

A: I've looked at what you have on the website, and I think it's something that somebody could follow and do. But I'm not sure I really grasp the training that you do for people who will be scoring it. I agree with you. We get to a certain level of objectivity, but there is also a certain level still of subjectivity. But I think the more we're able to articulate what it means to each other, the better we'll be at coming to a consistent, objective measure.

Q: I mean for the federal test, these people are exposed to one full week of training that involves a lot of practice, automated feedback, and inter-rater reliability stuff. "Why did this get twenty percent..? Why did you diverge? I mean really intense, expensive stuff. And yet, in a real testing environment, and we've got statistics on this, there's roughly fifteen per cent of the scoring units that are not unanimous. And then among those, there's debate. We were able to figure out that fifteen per cent of the time, in a three-person situation, a minority view prevailed for the official score. I mean, isn't there inherently always going to be some renderings that your experts just aren't going to agree with? I mean no matter how much time you spend trying to articulate what conservation of meaning means.

A: I think articulation through the training is important, but then the people who are training the scorers need to understand what they're getting at. That's all I'm saying. I think we still need a lot of discussion, and you might have had it more than I think we have had it in the world of health care interpreting, because we're still very new at the measurement and testing piece.

Q: Other than the federal government, who can afford forty hours?

Q: I was just going back to one of your slides where you talked about equivalence. You know, you want to make the person who is hearing the utterance feel the way he would had that utterance been in his own language. And that is really the old Eugene Nida translation, kind of the fundamental goal of translation, and that's called dynamic

equivalence. And really, that is what you're trying to do. And so I think it really is a good model. But you have to kind of go back and deconstruct it to realize that really, that's the kind of equivalence that you're looking for. Because you're not only looking at the meaning of the utterance, but you're looking at how it was said so that you get all that intent and register and feeling.

A: Yeah. I totally agree with you.

Thank you.