Developing Common Assessment Measures: A State and District Collaboration

By Jorge Allen, Tim Eagan, and Catherine Ritz

LEFT TO RIGHT: Tim Eagan, Catherine Ritz, and Jorge Allen

The Path Begins

If you ask a teacher in Massachusetts what they think of "DDMs," be prepared for an eye roll possibly followed by a rant. These innocuous little letters stand for the new student growth measures that are tied to teacher evaluation—"District Determined Measures"—and have become something of a controversy. As the name indicates, each district is now left to develop their own assessments, and many teachers have felt overwhelmed in trying to come up with something meaningful.

"Teachers in Massachusetts were aiming blind," reports Catherine Ritz, President of the Massachusetts Foreign Language Association (MaFLA). "We were hearing about DDMs that focused entirely on grammar points or vocabulary, since they're easier to grade. We needed to do something." Enter Craig Waterman, Assessments Coordinator at the Department of Elementary and Secondary Education (DESE), who reached out to MaFLA to offer support in developing model DDMs for world language teachers, leading to a statewide collaboration.

Three district leaders who ended up working on this new initiative were Tim Eagan, Department Head for Classical and Modern Languages of Wellesley Public Schools; Jorge Allen, Foreign Language Coordinator of Andover Public Schools; and Ritz, who is the Director of World Languages in Arlington Public Schools. The three came together to share their assessments, revising them, giving one another feedback, reviewing national materials, and of course meeting with Waterman for his feedback and critique. This collaboration culminated in a presentation at the DESE "Fall Convening," a conference for district superintendents, assistant superintendents, curriculum directors, and other school leaders. As representatives of MaFLA and their own districts, Allen, Eagan, and Ritz provided an in-depth look at their assessments, discussing the development process, planning for consistent administration, setting parameters for student growth, and managing and using data.

Presented here are accounts from each of the three districts involved in the development of their common measures, which are now posted as model assessments on the MaFLA website (***mafla.org/links-2/ddms***). While the process of developing assessments to use as state models has been ongoing, this collaboration was born in districts who first had to build assessments that were meaningful to the teachers there who would pilot them.

Arlington Public Schools, MA

Our first iteration of common assessment measures 5 years ago was pretty sparse. As the new World Language Director in Arlington, I worked with two teachers in my team of 19 to develop assessments focused initially only on Spanish. We had no idea what we wanted to do, but we knew at least that we wanted to develop the assessments around the three Modes of Communication—Interpersonal, Interpretive, and Presentational. We came up with some prompts that we thought were satisfactory and tested them out. In Year 2, we expanded. I brought the initiative to the entire department and used meeting time to discuss the assessments and what could be done at each level. We wrote language-specific prompts for each course, and most teachers piloted them. We also started working on department rubrics—a huge task! It became too much to tackle with the full department, so two teachers volunteered to spend time in the summer working together on developing rubrics for both our interpersonal speaking and presentational writing assessments.

By Year 3, we had more or less solid assessments in place, and everyone was administering them with our department rubrics. We now spent meeting time looking at student samples and doing common scoring. This led to intense discussions about the rubrics, which we quickly realized needed a major overhaul. Back to the drawing board! In Year 4, I started learning more about performance assessments, and started getting frustrated that we had different prompts for each of our four spoken languages—French, Italian, Mandarin, and Spanish. There were discrepancies between what teachers in one language were asking their students to do versus another, and this was causing tension and frustration on the part of the teachers as well. The more I thought about it, the more I felt like we needed to all have the same prompts and focus instead on proficiency development rather than content in one particular language. This led to yet another overhaul—both of our prompts and our rubrics—which brings us to where we are today.

Arlington now has a series of performance assessments designed for each proficiency level that span across languages. Our Level 1 course proficiency target is Novice High, and we came up with the following prompt for the Interpersonal Speaking assessment:

> You've been accepted into a study abroad program in France (Italy/China/Spain) for the summer and have just arrived in Paris (Rome/Beijing/Madrid). You've been assigned to a dorm room with a student from another country. You try saying hello in English, but get a blank look. It's time to test out your French (Italian/Mandarin/Spanish). So that you don't spend an awkward summer staring at the walls of your dorm room, say hi to your roommate and introduce yourself in French (Italian/Mandarin/Spanish). It would be nice to have a friend while you're in France (Italy/China/Spain), so find out as much as you can, and share some information about yourself to see what you have in common.

In Level 2, where our proficiency target is Intermediate Low, students are asked to do the following:

> Your family is on vacation in Québec (Tuscany/Guangzhou/Peru). You're super excited to finally see [local historical site]! As you're waiting in line, you notice someone who looks very familiar—an exchange student who spent a month at your school last year. You don't have too much time to chat, but you want to find out how he/she has been and what he/she has been doing since you last saw each other. Find out and share some highlights from the last year.

Since we are trying to elicit spontaneous speech, students are given the prompt the day of the assessment and are paired at random; we really want to see what they are able to do. We use our iPads, phones, or laptops to film students during the assessment so we can go back and score together. Since we have recordings of the assessments, students are able to put them in their digital portfolios, which we have set up throughout the department using Google Drive. As we look back at the same student's speaking assessment from previous years, it's amazing to see their progress. What's also exciting to me, as the program director, is to see how much these assessments have had an impact on instruction. For better or worse, we all teach to the test. But, don't you want to teach to this one?

If there is one thing I have learned, it is this: *It's the process, not the product*. While I certainly hope we are done with the massive overhauls we have gone through in our path toward developing meaningful assessments, I realize that we will never have the perfect assessments, the perfect curriculum, the perfect lesson plans. We will keep revising and refining, but we have learned, grown, and challenged ourselves (and our students) along the way . . . and that's what matters!

— Catherine Ritz —

Andover Public Schools, MA

I wanted to start the process of developing DDMs in Andover by drawing on what was already familiar and comfortable for teachers, while also aiming for assessments that would focus on developing students' proficiency in the language. The three Modes of Communication gave me a framework to articulate what we wanted to assess. Of these, the Interpretive Mode seemed to be the most familiar and comfortable one for teachers, since it focuses on reading and listening comprehension, and teachers in my district already assess these skills consistently. Furthermore, we could structure the assessments in a multiple-choice format which would allow for control in calibrating scores. To avoid our natural tendency of just assessing what has been covered in the curriculum, I decided to look at how interpretive skills were assessed in other assessments, including those from integrated performance assessments (IPAs), exams from national language organizations, Advanced Placement (AP) exams, and other examples.

Our assessments are comprised of discrete listening and reading comprehension items. However, we are having conversations aimed at collectively identifying authentic online videos and writing series of questions that focus on different aspects of interpretation: main idea detection, supporting detail detection, guessing meaning from context, and inference. Here's a sample of our current development for a Mandarin class:

Video link: ***www.youtube.com/watch?v=7lC2CRLQ5Fw***

Question #1:

This video mainly introduces . . .

a. Lu, Chuan, Yue, and Min cuisine
b. Hui, Chuan, and Yue cuisine
c. Chuan and Zhe cuisine
d. Yue and Zhe cuisine

Question #2:

Your friend likes spicy food. Which cuisine will you recommend?

a. Lu
b. Chuan
c. Zhe
d. Min

Andover is currently in the second year of piloting and test development. Our immediate goal is to figure out a testing timeline that is good for our assessments while avoiding times when the students might be overburdened with other district or state testing. In addition, being a district with multiple middle schools and a high school which is moving from semesters to a yearlong schedule, we have some local challenges that the second year of piloting is helping us address and resolve. However, an immediate positive outcome of starting the process of using DDMs is that instead of having anecdotal student data buried in course grades, we are now isolating and focusing on one mode of communication and discussing it objectively. We feel like we are moving from a trial-and-error approach, where individual teachers try different strategies on their own to improve student performance, to a more scientific and structured approach which will help us systematically improve performance across the district.

We are still struggling with some questions in terms of how to best administer these assessments, and are constantly finding aspects of the DDMs we want to tweak. We are also working to revise prompts so that we include more authentic material and include a balance of types of questions. The process is ongoing, but we are on the path!

— Jorge Allen —

Wellesley Public Schools, MA

What is our shared vision as a department? Implementing DDMs required my department to formalize some of our existing performance assessments to collect data on student learning and growth. Several thoughts emerged for me with this challenge. How could this work be meaningful? How could it highlight and improve upon our several years of success with 90%+ target language and performance assessments? How could I leverage our resources to examine student work, our beliefs, attitudes, and practices? So began my multi-year journey trying to answer these questions.

Having used rubrics for years, designing scoring guides for our presentational writing assessments would be easy—or so I thought. Several hurdles surfaced when using our existing rubric. Criteria like task completion and mechanics had too much weight. Domains like vocabulary were ill-defined or ill-matched to proficiency targets. So I assembled a small committee to design a more focused rubric measuring language control in the larger context of proficiency. We piloted scoring student writing with the rubric in 2014. Unwieldy and impractical describe the experience with that rubric! More consequential, however, was the fact that the conversations we were having when using the rubric focused on everything but proficiency. I needed help.

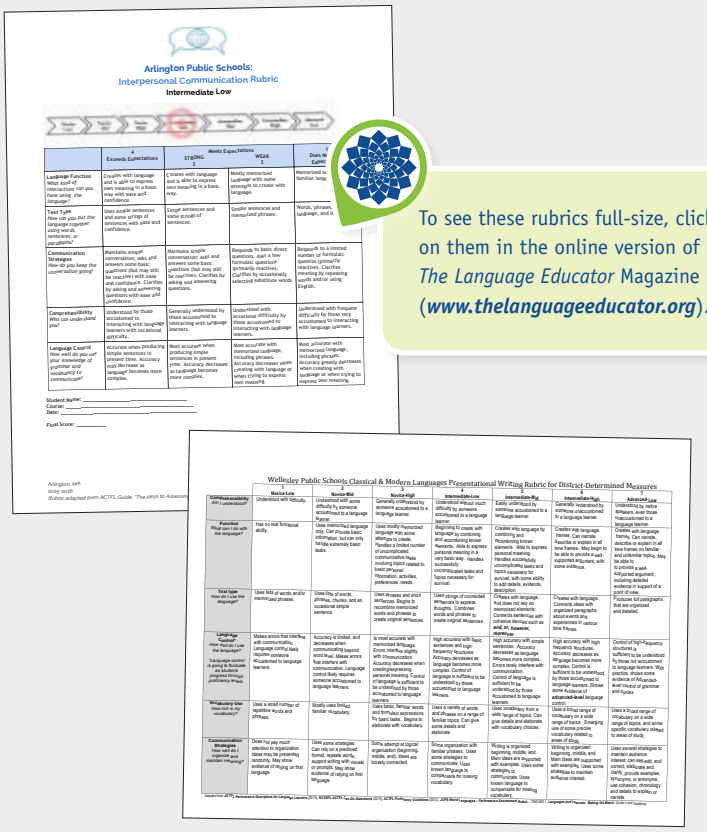
I turned to my shelf of books and began to read. Relying heavily on experienced educators and authors like Donna Clementi and Laura Terrill (2013), Helena Curtain and Carol Ann Dahlberg (2010), and Paul Sandrock (2010), as well as Greg Duncan's amazing website (***https://resourcesfromgreg.wikispaces.com***) has been fundamental in successfully navigating this journey. After many late nights and rubric iterations, I developed a viable draft, a rubric describing performance ranging from Novice Low to Advanced Low. With the rubric ready, the interesting work began.

Teachers here in Wellesley value collaboration; they share, work in teams voluntarily, and always support each other. But do we have a culture, as Boudett and Moody (2005) describe, which looks carefully at data, and uses data to inspire teachers and improve practice? Do we know how to look at student work and describe the work objectively? Are we skilled at staying low on what has been called the "Ladder of Inference" (Senge, Cambron-McCabe, Lucas, & Smith, 2000)? Before we act on any data, we must address these questions. We must develop procedures aimed at improving our scoring reliability, such as scoring small samples of work and calibrating our practice. This task means learning to see student work differently. According to Senge et al. (2000), humans are good at interpreting, but "seeing" is not natural; it takes discipline and lots of practice. It requires the suspension of judgment—an inquiry stance. He warns that our tendency to move quickly up the ladder, interpreting and drawing conclusions, leads to misguided beliefs about what we see.

In a recent issue of *The Language Educator*, Donna Spangler (2015) shared three quality protocols for analyzing data, student work, and classroom practice. She also shared resources for protocols, referencing one of my favorite sources, the National School Reform Faculty (***www.nsrfharmony.org***).

The Harvard Data Wise process (Boudett & Moody, 2005) defines three phases of looking at data: prepare, inquire, and act. We remain here in Wellesley, after some time on this journey, at the first phase: preparing for inquiry. The work is slow, but I have learned that only by going slowly to understand what students know and can do, and where our strengths and areas for growth lie, can we define our shared vision as language educators.

— Tim Eagan —



To see these rubrics full-size, click on them in the online version of *The Language Educator* Magazine (***www.thelanguageeducator.org***).

ABOVE: *Global assessment rubrics from Arlington Public Schools (left) and Wellesley Public Schools (right)*

TOP RIGHT: *Craig Waterman, Assessments Coordinator at the Department of Elementary and Secondary Education in Massachusetts presenting on assessments at the DESE Fall Convening.*

Moving Forward

While the new state DDM requirements were the impetus for the work done by these three districts, it was imperative that the assessments themselves be developed by teachers and content experts, rather than be dictated by the state. According to Craig Waterman, "the development of high quality assessments should be led by a deep understanding of instructional goals. It is for this reason that professional organizations, led by teachers, are best positioned to lead the work on developing and sharing good assessments. The model assessments, developed by MaFLA, reflect an excellent example of the type of leadership that professional organizations can offer the field." Through collaboration within districts, among districts, and with the state, the assessments developed as model DDMs, "provide a clear example to educators across Massachusetts and beyond about what precisely we want students to be able to do, and provide a common framework for the important conversations around improvement of curriculum and teaching," explains Waterman.

While the world language departments in Andover, Arlington, and Wellesley will undoubtedly continue to revise their assessments over the years, thanks to ongoing collaboration their assessments have become stronger, better articulated, and are of higher quality than if they were developed by each teacher individually. Furthermore, by supporting this work and making these model assessments available to everyone on their website, MaFLA is providing guidance to other districts so that they too can develop meaningful assessments that measure what we most want our students to be able to do in the language:

COMMUNICATE.

Jorge Allen is the World Language Program Coordinator at Andover Public Schools, Andover, Massachusetts

Tim Eagan is the World Language Department Head at Wellesley Public Schools, Wellesley, Massachusetts

Catherine Ritz is the Director of World Languages at Arlington Public Schools, Arlington, Massachusetts

REFERENCES

On the Web

- ***https://resourcesfromgreg.wikispaces.com***
- ***mafla.org/links-2/ddms***
- ***www.frenchteachers.org/concours***
- ***www.nationalspanishexam.org***
- ***www.nsrfharmony.org***

In Print

Boudett, K. P., & Moody, L. (2005). Prepare. In K. P. Boudett, E.A. City, & R.J. Murnane (Eds.), *Data wise: A step-by-step guide to using assessment results to improve teaching and learning.* Cambridge, MA: Harvard Education Press.

Clementi, D., & Terrill, L. (2013). *The keys to planning for learning: Effective curriculum, unit, and lesson design.* Alexandria, VA: ACTFL.

Curtain, H., & Dahlberg, C.A. (2010). *Languages and children: Making the match.* Boston: Pearson.

Sandrock, P. (2010). *The keys to assessing language performance: A teacher's manual for measuring student progress.* Alexandria, VA: ACTFL.

Senge, P.M., Cambron-McCabe, N., Lucas, T., & Smith, B. (2000). *Schools that learn: A fifth discipline fieldbook for educators, parents, and everyone who cares about education.* New York: Crown Business.

Spangler, D. (2015, October/November). Effectively analyzing student data, student work, and classroom practice using protocols. *The Language Educator*, *10*(4), 34–39.