Q&A: The Development of Automated Online English and Spanish Writing Assessments

Introduction

In 2021, the eLab's team at Emmersion pursued the development of a new product line of writing assessments the TrueNorth Writing tests. The first languages included in this product line were English and Spanish. While a technical paper highlighting each tests development and validation will be forthcoming, we feel that some may benefit from a high-level introduction to these assessments that would be more appropriately formatted as a QA with the test's creators.

Why the prioritization of a writing test solution?

There was a pretty natural draw for us to look at writing. Our TrueNorth Speaking tests (offered in 8 languages) use elicited imitation (listen & repeat) as key task type. So they target the critical skills of listening and speaking. We also have reading/grammar/vocabulary assessments in seven languages. So with offerings that cover three (Speaking, Listening, Reading) of the four major skills, there has been a natural gap in direct writing measurement.

However, our discovery was more than just an inventory check. We've heard directly from clients and observed industry trends that more and more ability to communicate in writing is a valuable skill to screen bilingual applicants. More and more contact center companies are being tasked with omni-skill demands. Agents conduct customer and technical support not just over the phone but also through chat and email. Also with the move to more and more remote flexibility for contact center work, more and more agents, supervisors and administration are using chat, email, and message boards for internal communication.

We also have seen that while a person's ability to speak in a language gives some indication of general ability (including writing skill), a person can be a competent speaker of a language and yet have gaps in the language aptitudes required to express clear and accurate information when writing. In short, in contrast to some of the deep thinking and questioning our team does, this was no brainer.

When customers asked for a writing assessment did they know what they were looking for?

As always, we were so grateful for the opportunity to have extensive conversations with industry leaders from around the world. Fairly consistently we heard frustration and frankly surprise that there wasn't already a solution that would meet this need. Some had been using internally developed assessments that were cumbersome and admittedly poorly implemented. Others relied on indirect measures or even just self-report screening and the expensive intervention of quality control when inability was exposed.

Consistently, we heard the need for something similar to our speaking test: quick to administer, fully remote capable, immediate results, alignment with international standards. Reasonably, most expected that a writing test would require writing (crazy right;). The tasks used to prompt writing performance didn't need to be restricted to chat or email but there was agreement that they should fit general language functions of description, narration, problem resolution, and outlining, selecting, and justifying a course of action.

Finally, there was a desire that as with our speaking test where administrators not only get scores and certificates certifying performance, they wanted to be able see the sample of writing (or script) associated with the test.

Could you tell us a little more about the discovery process around the types of writing tasks that were included in the test?

This was an area where despite all of our careful thinking and research, we knew that it was important to make data defensible decisions. When you look at writing proficiency assessments, there's a pretty conventional form. A task is presented and the test taker creates a script in response to the task within a time constraint that has a relationship with the expected length of the text and the expectation for revision before submission.

Oftentimes a writing proficiency test will present burdens of meeting the characteristics of a specific genre of writing (i.e. academic supported opinion essay) with specific rhetorical demands (introduction, thesis, body paragraphs, conclusion). In creating the most universally appropriate tasks for language screening, we knew we wanted steer clear of any task that would be so constrained. We didn't want writers to think there was a single right way to respond to the tasks.

Also, as we mentioned above, our conversations with industry leaders gave us a pretty good frame of reference in terms of the specific domain and language functions that we should be targeting. We shared a desire that the test be fair in targeting these language functions. We drafted tasks that targeted these functions. We put them in front of hundreds of prospective test takers who had the choice of which tasks they would write towards. Tracking items that were selected and not selected and the success of understanding and meeting the task demands revealed a core of items that seemed the most universal. This helped us to have confidence that absent of having a choice between different tasks, we'd present a test taker with the best chance to show what they could do.

We knew that we wanted the administration of the test to be efficient. That was not just about making the test short but it also made it more authentic. Unlike tests that might be assessing ability to write for academic purposes where the writer will have longer ability to formulate thought. The writing skill that our clients need to be confident the test is measuring for is the more responsive, automatic, even dynamic writing inherent to chat and email. So a shorter time limit would help reveal what this more natural less edited type of writing performance would look like.

So what does the current test form look like?

After an initial language background survey and test start-up screens including instructions are completed, the test taker has two parts to complete. Part 1 consists of fully adaptive multiple choice grammar and vocabulary questions. Part 2 takes what is learned about ability from Part 1 and presents a level appropriate writing task that should provide a fair opportunity for the test taker to evidence their writing skill. This is followed by a score report where ability is described within the terms of our TrueNorth Scale with estimated CEFR level, descriptions of ability at this level and the script submitted during Part 2.

The Part 1 - Grammar Multiple Choice section is kind of surprising. What motivated you to use this structure and content?

This was in part directly informed by our conversations with organizations in terms of what they were already using. Many were using a grammar skills test as part of their screening process. So this was a comfortable place to start for the clients and frankly for test takers who have had similar tests be prominent in their language learning experiences.

Grammar knowledge does have well supported predictive power for other literacy skills in a language including writing. Within our WebCAPE test offerings, we had item banks of hundreds of items that have been accurately measuring language ability and helping organizations make decisions based on language skill for a long time. Just because we were creating something new didn't mean we couldn't harness the wisdom of years of success.

Before repurposing this test content, we did take the mountain of data from these assessments (English Grammar and Spanish) and updated our understanding of each items difficulty and usefulness. We also modernized the adaptive algorithms that would select items and calculate a final score. Machine learning would also help convert the estimate of ability from Part 1 to our TrueNorth scale with its accompanying CEFR estimates.