How to run a level test across 17 time zones

This is the text of a talk delivered at the New Directions conference, virtually hosted by the British Council Singapore in November 2020. The talk was given by Andrew Stokes (ClarityEnglish) and Ian Cortez (British Council).

Andrew:
Over the next 20 minutes, Ian and I are going to tell you a story. Let me start by setting the scene. You are a British Council Manager in Saudi Arabia and you have been working for months to secure a contract with a prestigious company. We'll call it the Gulf Transportation Company (GTC) for privacy reasons. This company wishes to employ 50 graduate engineers this year, which gives you an idea of its size, and it has received over 2,000 applications, which gives you an idea of its prestige and high profile. Some of these applicants are in Saudi, but others are studying in the UK, and right across the US, from East Coast to West Coast.

Like many large companies, prior to the face-to-face interviews, GTC will administer a battery of tests: maths, logic and reasoning, and also English. They will only accept applicants who are C2 in English. You as British Council manager have proposed Aptis, and GTC has agreed. The overseas applicants will fly back to KSA to take the test. Everything is settled by February, and the tests will take place in April.

Then disaster strikes. By March, Covid 19 is stretching its tentacles across the world and countries have started locking down. The original Aptis plan will no longer work, but equally, it is not an option to let down such an important client. So, what are you to do?

Ian:
What we needed was a 100% online test, with CEFR output. Obviously it has to be accurate and reliable and it needs straightforward and simple administration. We knew about the Dynamic Placement Test, published by Clarity, because we have used it in a number of projects around the world. We had confidence in the test construct which was devised by telc Language Tests, the European Union language testing experts. Also, the British Council has had a close relationship with Clarity since the 1990s, in particular with Road to IELTS which has run tens of millions of sessions.

The challenge is this. With Clarity based in Hong Kong in the East, the company and the BC admin in Riyadh in the middle, and test takers ranging as far as California in the west and the UK in the north, could we run a successful test in a limited time frame across 17 time zones? Add to this that the lines of communication are complex: HR at GTC interact with the test takers. GTC interacts with BC for all aspects of test administration. BC interacts with Clarity which will have a dedicated member of staff for this project. Well, it's difficult to understand even when I say it here, and it was much more challenging at the time. So, how did it work out? Let's look at three aspects. Andrew, what about test security?

Andrew:
Well, as you will be aware, clients often take a very black and white approach to test security. They will often not accept the validity of a test result unless it has been properly invigilated. So this was potentially an interesting challenge. Things have moved so fast this year that we are all now familiar with the term "online proctoring", but nobody had heard of it back in March.

So we were faced with two questions:

- What options did we have for security? (more practical)
- How do we categorise the test in terms of security requirements? (more theoretical)

Let's look at the practical side first. Basically, options were limited to non-existent. For reasons we won't go into here, it's quite difficult to cheat the Dynamic Placement Test, but of course there is nothing to stop you getting someone else to do it for you.

We had recently implemented a fairly simple system to solve this problem in a project with Brunel University which wanted a straightforward way of testing applicants in the absence of IELTS, which had totally shut down at that point. The system was to set test takers up a Zoom meeting. They would position their phones so they could be observed while they took the test on another device. When the Zoom meeting started, an invigilator in London checked the identities of the candidates and could randomly observe them as they took the test at home.

This system was successful and the client, Brunel University, was very satisfied. But it was significantly different from the GTC project. The Brunel project never had more than ten test takers at any one time, and they were all in the same country (China). The GTC project had potentially 2,000 test takers spread across the world. It's difficult to see how it could work. So we were effectively out of options. But was invigilation really necessary?

This takes us on to question 2: how do we categorise the test? How important actually was it to monitor these job applicants?

In simple terms, let's ask whether this is a high-stakes test. The interesting answer is Yes and No.

For the test takers, of course it was. In this case, if a test taker failed to achieve a C2, then that was the end of it, it was game over — they would have lost the job. But for the employer, this was not a high-stakes test. In this it differs from, say, IELTS. If a candidate cheats their way to an IELTS 7 when they are really 5.5 and ends up at a UK university, then there is a real problem for both sides: the student and the university.

But the GTC case is different. In the GTC case, not getting a C2 would filter the applicant out, but getting a C2 would not qualify the applicant for the job. It would simply keep them in the pipeline. And if they had, in fact, cheated, they would be caught out at a later stage, perhaps in the face-to-face interview. So there was no risk that if a candidate cheated there would be serious consequences, and it was not therefore a high-stakes test from the client's point of view.

Would they appreciate this distinction though? As it turned out, we were very lucky to be dealing with a highly sophisticated HR team at GTC who understood that test security should, in fact, be calibrated to the function of the test. So they accepted that this is not a black and white picture, but perhaps one with many shades of grey. And as a result they were quite happy from both a practical and theoretical standpoint to let the test takers self-regulate.

So, that was challenge number one taken care of. What about running the test itself Ian?

Ian:
Normally with DPT the admin is pretty simple. You set up all the test details on a screen. Then you import the test takers' email addresses from an Excel file, or similar and put them into a group.

Finally, you assign the test takers you have imported to the test you have created. The system will then email them all the details: the url, the test time and a confirmation of their username and password. Simple!

However, in this case it couldn't work like that because the HR department was not able, or at least not willing, to release the email addresses of the applicants. So the process had to be more complex — and as we know, the more complex the system, the more likely it is to go wrong. As we were going to find out...

Luckily we had two very competent people working on this: Eithar Diab at the British Council and Dora Tsang at Clarity. They worked with GTC HR to devise a system of unique anonymous IDs, each matched with a password and cross-referenced to a test taker's real identity in terms of their name and email. This would enable HR to email the test details to each applicant before the test. After the test, using the same system in reverse, it would be straightforward to marry up the test results with each candidate in an Excel file.

What could go wrong?

Well, in fact, this all went very smoothly and nothing did go wrong with this part of the process. The test was duly set up to run between the 9th and 11th April 2020 and the candidates were informed. But late on the evening of April 11, there was a panicked phone call from GTC to the British Council. They'd had calls and emails saying the test had closed earlier than they thought, and some of the applicants were unable to get in. They'd gone back to their invitation emails and realised that they had put the wrong times in the emails. They acknowledged it was their fault — but was there anything that the British Council could do to save the day?

It was actually no bad thing that this happened because we were able to show them excellent customer service. Changing the end date of the test is a simple matter and by 8 o'clock the following morning, we had contacted Clarity, and they had extended it to a new end date matching that in the GTC email. Actually, now we have a little more experience with the test, we could have done it ourselves within 5 minutes.

The result of all this is that the client was extremely happy. Not only had we set up an efficient admin system, but we had also saved them from the consequences of an embarrassing mistake.

I'd also like to highlight one additional benefit of this system. Remember at the beginning Andrew said that the overseas applicants were going to fly to Saudi Arabia to take the test. Just think of how wasteful that is: New York to Riyadh is 6,526 miles, or 13,052 miles return. It is certain that by using this 100% online test we saved hundreds of thousands of unnecessary air miles, probably millions. That is perhaps twenty to thirty thousand tons of CO2.

This may not seem relevant to testing, but surely it relates exactly with what we were talking about earlier: the importance of calibrating the test security and making responsible decisions around what we find. So in this case one of the decisions was that we didn't need invigilation -- and an additional outcome was avoiding a significant environmental cost.

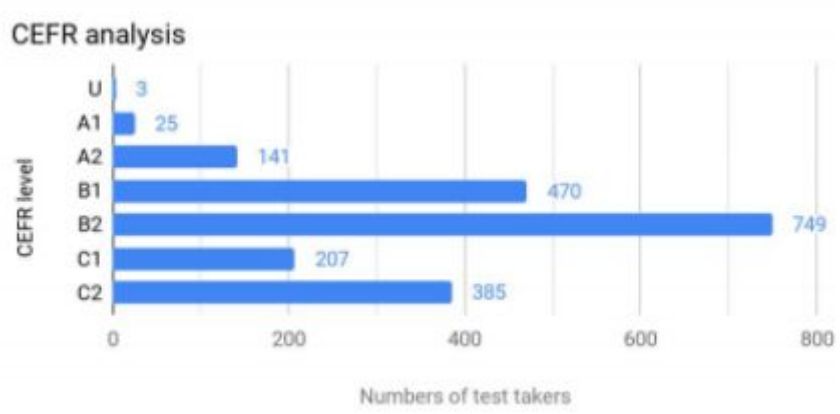
Okay, Andrew, let's complete this talk by looking at the test output.

Andrew:
Thanks, Ian. First of all, and most importantly, GTC got exactly what they wanted. In terms of operations, the test was overwhelmingly successful as we can see here.

- 2,007 candidates took the test
- 1,980 completed it successfully
- 23 signed in and then gave up
- 3 took the test, but did not complete enough answers to get a grade
- There was 1 error.

So that's a technical success rate of 99.95%.

We can see from this next chart that they were able to whittle down about 2,000 applicants to the 385 who achieved a C2. So not only do we save the CO2, but removing 1,600-odd applicants from the process represents a massive saving in cost. That's even more dramatic when you consider that each Dynamic Placement Test only costs $5.



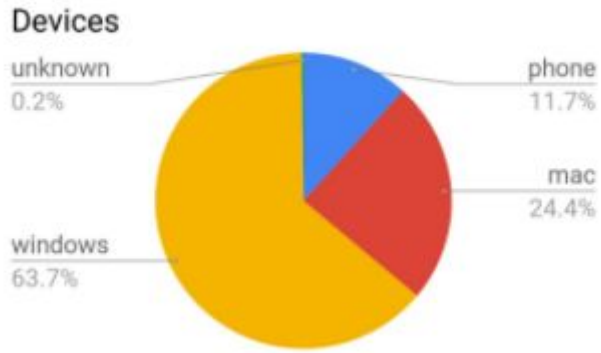
So for GTC that is the end of the story, and they walked away as a very satisfied British Council customer. And the managers responsible were able to rightly congratulate themselves on a job well done.

But the story doesn't end there for the rest of us. There is data here to make the most jaded testing specialist salivate. Let's look at three data sets and consider what we could do with them. I'm sure you have noticed something slightly odd about the graph above. If we compare this with a normal distribution curve, the number of C2s seems too high. Why might this be? I would postulate that there is a correlation between the CEFR level and the geographical location of the test taker. I

would suggest that the seemingly abnormally high level of C2s is explained by the cohort that has studied in the US or the UK, perhaps since secondary school. How can we test this? Well, we know the IP address of every test taker, so we can run queries to see if there are correlations between CEFR level and the country where the test was taken. Then if we wanted to, we could take it further: perhaps we could see whether correlations exist between urban and rural areas within Suadi Arabia itself. And perhaps GTC would find this information useful in their next round of recruitment.

And if we are interested we could do exactly the same thing for the time at which the test was taken. Who performs better: the night owl or the lark, adjusted for time zone, of course. Or does it not make any difference?

Here is another interesting chart.



For 99.8% of the test takers we know which device they used. Here it is broken down quite crudely into Mac, Windows and Phone — but we can get very much more granular. We have, in fact, researched other data sets to see whether there is a difference in performance depending on device and found that overall there isn't. But within that overall conclusion there are some interesting titbits. We find, counter-intuitively, that test takers complete reading items marginally quicker on mobile. Given that they are working on a much smaller screen, and having to swipe between text and questions, why would that be?

There's lots more work that could be done on this. And we have other comparable data sets too, from BC projects and others. So if there's anybody out there who would like to do some research, please get in touch.

So what are the main conclusions?

1. The British Council and Clarity were able to draw on an effective long term partnership to arrive at an outcome neither could have achieved alone. The success of the managers working in a fast-moving, complex environment must not be understated.
2. With GTC's cooperation, we were able to effectively calibrate security to achieve multiple benefits: cost and time savings and even environmental benefits.
3. We have harvested a wealth of exciting data that can further contribute to the relationship with the client.
4. And finally, cast your mind back to the scene I set out at the beginning. As manager, you have overcome the challenges, you have solved the problems and you have ended up with a very satisfied customer.

andrew.stokes@clarityenglish.com