Regulation Matters:
a CLEAR conversation

Episode 27: What Makes a Difference for Candidates Taking Computer-Based Tests?

March 10, 2020

Line Dempsey: Welcome to our podcast, Regulation Matters: a CLEAR conversation. Once again, I'm your host, Line Dempsey. I'm currently the Chief Compliance Officer with Riccobene and Associates Family Dentistry here in North Carolina. I'm on the CLEAR board of directors as well as the current chair of the National Certified Investigator Training committee with CLEAR. If you're not familiar, CLEAR is the Council on Licensure, Enforcement and Regulation, and it's an association of individuals, agencies, and organizations that comprise the international community of professional and occupational regulation. Our podcast is an opportunity for you to hear about the latest and greatest in our community. And to truly make this international, today I'm joined by Paula Lehane, a PhD candidate at Dublin City University in Ireland with the Center for Assessment Research, Policy and Practice in Education. We're very glad to have you with us today.

Paula Lehane: Thanks very much, Line. I'm delighted to get the opportunity to contribute to the podcast.

Line: Well, perfect! And also I wanna thank our listeners for joining us. What we'd like to talk about today was a topic of an article that you wrote for the latest issue of the *CLEAR Exam Review*, and the title of that article was, "What Makes a Difference for Candidates Taking Computer-based Tests: Issues Surrounding Device Comparability and User Interface Modifications." You know, as technology is constantly advancing, I guess regulatory organizations really do need to think about these issues, both from the perspective of the candidate and their expectations and from the perspective of how technology will impact licensing and examination programs as we move forward. So I guess let's start with this question, what are some of the big things that we need to consider for candidates that are taking computer-based tests?

Paula: Well, I think in the perspective of regulatory organizations, they need to think about what the candidates expect themselves. And I think nowadays when we move towards computer-based tests, candidates are expecting a certain level of flexibility that might not have been there before with brick and mortar exam centers that they had to attend that had invigilators and things like that. Also, what people are really expecting is the flexibility in terms of how they will take their test. So there's the

flexibility that they could possibly bring their own device to the testing center, maybe even take the test in their own room rather than having to travel maybe hundreds of miles to a place. And that level of flexibility, while it kind of cuts down on the logistical nightmare for regulatory organizations, it does mean that they need to take into consideration what impact that flexibility might have on their standardization practices, which is a really important cornerstone of the assessment process. So I think the biggest thing that we need to consider is, how are we going to manage this expectation of flexibility in testing with the requirements of having a very clear and consistent standardization process?

Line: Well, it's very interesting. I guess the term is device diversity or bringing your own device. I've recently taken some courses, even a regulatory course for certification, and I had to go to a testing facility actually to take the test. I couldn't do that. But I guess, is this something that the industry is... Obviously, it looks like it's something that they're considering. I guess, what are the pitfalls or what should they be carefully looking at when considering using this device diversity?

Paula: So certainly within the education sector, which is the one that I'm based in, Bring Your Own Device is becoming more and more common in testing just because there's a huge range of schools out there with different levels of income, different access to different types of devices like tablets, desktop computers, laptop computers. Some of them, their only device they might have access to would be a smartphone. So device diversity is great for people like yourself who might not be able to get to an exam center; they can just open up their device and work with us. Though, that does have some issues there, because you need to consider the idea of device comparability. And this is something that Dadey and colleagues have kind of been working on for the past six, seven years. And what it says is, it basically means that device comparability refers to the comparability of scores produced by test takers taking the same test on different devices. So if I take a test one day on a desktop computer, my score should be relatively comparable to the next day when I take it on a laptop, or maybe a tablet.

This issue of comparability has been around for decades; it's written into the 2014 *Standards*; Bennet discusses it himself. And we usually define it on a continuum of content and score level. And that is, if a test and its variations--be it a desktop, a laptop, or a tablet--use the same test items to assess content, then we should have seen the test scores are interchangeable.

This assumption is kind of flawed in the area of device diversity and testing because it doesn't really take into consideration how there could be major variations in content presentation for identical test items, because of how different the devices themselves are. We call that the form factor of a device, and that form factor could interfere with the presentation of the test items but also how the test taker interacts with the item they need to answer.

Line: So, this issue with form factor, I think I kind of follow what it is, but I guess ultimately how can it interfere with test-taker performance?

Paula: So the form factor is the fancy term used to describe how the size, style, shape, layout, and position of the big parts of a device are laid out. So if you think of it as a standard desktop computer and a laptop, they are pretty similar form factors. They have a screen, they have a key board, they have a mouse. If you think about a tablet, the tablet is really different to a computer or a laptop. In a tablet, you're normally using a touch screen interface or you might be using the keyboard that's on the screen. Or you might have an external keyboard or you might have a tablet pen. So these are quite big differences in how we interact with whatever we're looking at on the screen. But, it also might have an impact on how we actually view something.

So if we just take, for example, a standard, the screen size of a computer, say. So the screen size has been found to be one of the greatest factors of construct irrelevant variance in comparability studies. So, when researchers compare how people do on a test that they take on a laptop and test that they take on a tablet, and what they found is that the screen size and the differences in screen sizes available on laptops and desktops, when compared to a tablet, can have a big influence on test-taker performance.

So, Bridgeman and Sanchez & Goolsbee--those are two research studies I cited in the paper in this issue of *CLEAR [Exam Review]*--they found if there's an increase in scrolling, there's a reduction in factual recall and item performance. So, on tablets, you have smaller screens, so you might do more scrolling, and that means that you might be more likely to forget something if you're doing something like a reading comprehension item. But if you're on a desktop or a laptop, you scroll less--you might only scroll once--you're more likely to remember some key bits of information that you need to answer the question. And King et al. did a really interesting study about that, that in test items that require a lot of reading, if you do it on a device that makes you scroll a lot, you don't do as well.

So going back to your original question of how a form factor can interfere with test-taker performance – if I take my test on a smaller screen, I might not do as well as someone who takes their test on a bigger screen.

Another study that might be of interest is that Kong found out test takers actually take more time to answer questions on touch screens, except for drag and drop items. Whereas another study found that people who took tests on touch screens, particularly if they were older test takers, people in their 50s and 60s, the performance gap was smaller because they didn't have to work with a mouse that they weren't comfortable with. So, again, there's lots of little issues directly related to the form factor of a device that could have an impact on test-taker performance then.

Line: That's interesting 'cause anecdotally from my experience, having done both on an iPad tablet as well as taking a test on the computer, I actually found the test on the computer was more difficult because the screen was larger, so the sentence was longer when presented on the screen versus on a shorter screen, word wrap allowed me to see things a lot easier. I tend to be a fast tester anyway, so I tend to scan the question and look for things- and that may or may not be to my benefit, but very interesting.

So I guess to be devil's advocate, for these populations where maybe a testing center is not something that's available to them or to make it easier, should we go back to some type of proctored paper-based test and forgo our movement towards computer and technology?

Paula: You can't put anything back into a box when it's been opened [laughter], and I think it would be a bit remiss to go back to the very traditional days where you have the paper-based test going back to the exam centers. So you can't just ignore this huge technological development because it poses a risk. The best thing you can do when you see a risk is to absolutely minimize it so that you have the confidence that you need in your testing standards. And one of the best ways to do that is to design your test from the very outset with these form factors in your head so that you know that test-taker performance won't be influenced by a device's particular features.

So if we go back to that example about scrolling, how that if I take a test on a desktop computer with a big screen versus taking a test on a tablet with a smaller screen. What I could do is make sure that between the two screens, I scroll the same number of times, so that if I take the test on a tablet, I'm going to scroll twice to read the entire passage and then I can go on to my test item. And again, if I take the same test on the desktop, I was scroll twice so that there's consistency in test-taker experience. And that might mean you might have to get a little bit creative with how you present your items. You might have to work with using consistent font sizes or you might have to have a paging interface, which is what some people did in the education sector in Australia. You might make sure that people in the tablet condition can only read the text in landscape mode rather than portrait mode, so that you're making everything very, very consistent.

I was lucky enough to attend a conference in Madrid recently where some test developers were starting on that, using this idea of principle assessment with design. I think it was from the company ACT that were working on it, that at the very outset they were trying to make sure that there was consistent usability for all their test takers, that at the start, they would examine the factors that could introduce construct irrelevant variance that are not essential to the task and then work to make sure that there is that consistency. So when that they're aware of a difference, they're doing something to modify it and rectify it.

Line: And it's interesting you mentioned the landscape and portrait mode as well. I always get these great comments when people come into my office because I have two monitors for work and one of them is put in a portrait mode because I'm looking at PDFs or patient records on a regular basis, and I can look at the entire thing without having to scroll, which is something that's important. So it's interesting that you mentioned that they're even thinking about that as far as forcing you to take the test in landscape mode, so that you don't have this unfair advantage necessarily between different devices. So I guess you mentioned some concerns surrounding user interfaces. Could you be a little bit more... Could you expand on that?

Paula: Yeah, so when we go back to that idea that we have flexibility, that means I can bring a device

that I'm comfortable with into a testing scenario and things like that. So just in relation to device diversity, another related issue that also needs to be considered is that when people have flexibility in the device they bring in, they might have some flexibility to personalize the device to their own preferences, in terms of the user interface. So user interface basically means that users interact with the hardware and software of a particular device; the UI determines how commands are given to the device and how information is shown up on the screen. So things like topography, the font size, the font color, menus - all of that contributes to a particular device or software.

If you pay even a little attention to the computer industry, you know that the user interface for a MacBook and the user interface for a Windows device or an Android device, they're all slightly different, and within each of those, you can personalize this further. So say for example, my friend can't use my particular computer because I have personalized suggestions so that when I do the three-finger swipe, I can see all my documents all at once, and he can never quite understand why this new screen suddenly comes up.

So in relation to this flexibility about how personalized devices, being able to personalize the font color you see your test in and being able to personalize the font size you see your test in, and that has the exact same issue- it might have an impact on standardization.

This isn't new to the testing industry. It certainly is older to the testing industry than the use of computer-based tests themselves because testing accommodations have always been there. So we expect certain testing adaptations to be given to the needs of certain individuals. If the person has a visual impairment, they might have very specific field of vision; they might need really big font sizes to work on the screen.

Those adaptations are perfectly necessary, and they're absolutely fine in terms of test-taker performance. But it's when modifications like that become a choice or a preference, there is some research to argue that these modifications might not be suitable unless there is a particular reason. If we leave it down to choice or personal preference, you might actually be doing yourself a significant disservice in the test.

Line: So, obviously, some of the things that you've touched on already, and I immediately thought to my mom's iPhone, she's got a larger text on it, all those things, right? So I understand that it's way different for someone that may need that. But what are some of the other examples of, I guess, tweaks that people could do that may or may not be helpful in that process?

Paula: Yeah, so again, if we go back to topography, recently, again it's based in the education sector, but also within third level in particular that they were able to modify the topography. And that's basically anything related to the onscreen style and appearance of the written word. They could change the color combinations- they could read black on white, they could read yellow on black, they could change from serif to sans serif font. And research is saying that, well, there are certain things that maximize your performance in a test in relation to topography. So sans serif fonts have been

shown using eye tracking studies to be the best font in an onscreen environment, that the serif fonts (I always remember they're the Times New Roman and things like that) they tend not to do quite as well as sans serif fonts, which would be like Arial and stuff.

So we can read those more quickly, and if we read them more quickly, we're more likely to remember them. We're not using a lot of our cognitive power to understand them. In relation to font size, it goes back to the issue of scrolling. If my font size is so big that I'm scrolling 15-20 times to read a text that might have an impact on my performance.

Another study that people did and they repeated the same study in 2008 and 2014, and what they did, they looked at color combinations- reading certain color text and certain background color combinations. And depending on the type of screen, if it was a modern LCD screen, certain color combinations were a lot more legible. So again, that was another thing that sortof needed to be considered.

Another thing that has come up more and more is this idea of a clock, that timing is very, very important in testing as we all know. And how should that time be represented in an on-screen environment? Should it be a standard clock like the one we would traditionally see in an exam center? Should it be a kind of a countdown timer where we see how much time has elapsed or we see how much time is left? And those are three very different ways to convey time, but we don't really know if they're different. There's some very, very, very preliminary research about this that they're saying, well, some of those clocks might be more anxiety-inducing, so don't give people the option to pick that clock.

So again, it's knowing what's best for yourself, but balancing that with what research is saying about the testing environment. In a work environment, certain things will work for you, but when you move into a testing environment, it needs to be optimized for the test itself but also for you as the test taker.

Line: What about annotating things? I know when I took my test, I went old school. We were allowed to bring a sheet of paper in and a writing utensil. I know there were ways that you could flag questions to come back and re-look at them or whatever. I typically just wrote them down, or if I had some little bit... usually my first response or thought on a multiple choice question is probably right, and I noted that so that I could come back and take a look when I looked at the question again. But is that something that would also play in as far as making it challenging to - I guess, I don't know what the right word is but - to standardize in these systems?

Paula: Well, not even within the issue of standardization, but I think you're making a really good point that research has realized recently at the moment this idea that we need to be able to annotate our test papers. And so marks made as we read a test so we can make them very explicit, like text going "oh, refer back to Marshall's Law on X" note to yourself, or even like what you said there, highlight or underline your first thought or answer. And a lot of the tests that I would have seen that have moved towards a computer-based environment haven't made the effort to include annotation tools. And

marking your test paper with those really clear notes to yourself are really, really important in a testing scenario and should be facilitated in an online environment.

Now in your situation, you brought in a sheet of paper, but there was a test in 2010, and they found that... it was a really interesting study in 2010 with Taiwanese students trying to complete an English language test. And what they found some of the students were allowed to mark questions with stars, underline, highlight, however they wanted, like they would do in a paper-based test. And in this study by Chen in 2010, they noted that the really high-ability students didn't really seem to benefit from marking their test papers because they had the knowledge already; they didn't need those extra clues or extra notes that marking gave them. It also didn't improve the performance of low-ability students, because they didn't even have the knowledge to answer correctly, even when they did use marks to help them. But for those average-ability students, marking the test items seemed to give them a much better opportunity to show the knowledge because having those marks allowed them to recall the information.

And that study is a really good example of the value of annotation tools. And I think that's one area that the testing industry needs to try and get caught up on, that in terms of user interface tools, that's the tool that they really need to allow for and develop and give people flexibility in that area, 'cause research says that when you give people flexibility in that area, it can help them.

Line: That's fascinating. So I guess, what's next to further this?

Paula: Well, what's next? So many things, so many things. But I suppose the biggest thing would be conduct more research and make it public. So you mentioned that I'm based in a university; I'm not specifically within the testing industry, and I've been in the very fortunate position of being able to go to industry conferences, and the things I've learned there are absolutely fascinating. They're brilliant, like I refer to them here in the state of this 20-minute podcast. But I wouldn't have had access to them otherwise because it tends to stay within the industry. And I think that the research needs to continue, but it needs to be disseminated among the public and needs to be shared between testing industries, different... So if you found out something in a dentistry test, the people in the accounting test could also use it because again, this is a new format of testing; we want to make sure it's the best across the board.

I think the other thing in relation to research is that it needs to be based within the assessment industry. A lot of the research I refer to in the article in *CLEAR [Exam Review]* comes from educational backgrounds or a very small usability study. I think they need be based within the assessment industry, they need to be exploratory in nature, but they also need to be experimental in nature. And yeah, so just doing more research, diversifying it, and making sure that it's public so most everyone can have access to it.

Line: Well, this has been a fascinating discussion. I think we could talk for a great deal of time longer, but I know you probably have to get back to teaching another class or finishing up your day. But I do

really wanna thank you for being a part of this podcast. It's always wonderful to have an opportunity to talk about these things and learn from each other and especially from your article from the *Exam Review*- it's been great. So thank you for speaking with me today.

Paula: Thank you very much for having me. I appreciate it.

Line: I also wanna thank our listeners. We'll be back with another episode of Regulation Matters: a CLEAR conversation very soon. So again, thank you to our frequent visitors. If you're new to the CLEAR podcast, you can subscribe to this on a variety of different mediums. It's available on Podbean, iTunes, Apple Podcast, Google Podcast and Google Play, Stitcher, Spotify, and TuneIn. If you've enjoyed this podcast episode, please leave a rating or comment in the app. That helps us improve our ranking and makes it easier for our new listeners to find us. Feel free to also visit our website at www.clearhq.org for additional resources as well as a calendar upcoming training programs and events. Finally, I'd like to thank our CLEAR staff, specifically Stephanie Thompson; she is our content coordinator and editor for this program. Once again, I'm Line Dempsey and I hope to be speaking to you again very soon.

The audio version of this podcast episode is available at https://podcast.clearhq.org/e/Difference_CBT/.