Jillian Derby, Dr Kelly Dore and Dr Eliot Rees

Webinar date: 02.12.2020

Putting together pieces of the puzzle: Defensibility in admissions and in-program assessment.

Transcript of webinar:

ELIOT Ok thank you all for joining this afternoon's ASME BITESIZE session. I'm Eliot Rees I'm a lecturer of medical education at Keele University. I'm the Chair TASME the training group within ASME. Delighted to be joined by Kelly and Jill for this session and I'll introduce you to them and the talk in a second. But I'll just go through a few housekeeping notes first if that's ok. So, the session is going to last for approximately 45 minutes and Kelly and Jill's presentation will be going for most of that time but there will be some time for questions at the end and they'll pause about half way through to answer some questions on the first part of the talk. If you do have a question to ask please pop it in the chat field and we're not going to use the Q&A feature so just in the chat and please make sure that you put it to the panellists and attendees so the other participants can see your question and then they'll know what we're answering. If we don't get time to answer all the questions today, we'll provide a document after the webinar with any questions that we weren't able to address and some written answers. The chat field's available for you to contribute to the conversation throughout the presentation so please do feel free to pop in any comments or questions in there. But again, please do put it to the panellists and all attendees. Just to let you know the webinar is being recorded and a video of the webinar will be made available on the ASME website along with any supporting materials after the presentation. If you have any technical problems during the talk please make ASME aware by emailing events@asme.org.uk and I'll pop that in the chat field in a second but email them rather than adding it into the comments and then someone from the team will be in touch to try and help you out. So, without further ado I'd like to introduce our speakers this afternoon Kelly Dore who's Co-Founder of Altus Assessments and an adjunct Professor of McMaster University. And Jill Derby a research scientist at Altus Assessments. And they're going to be talking about 'Putting together the pieces of the puzzle: Defensibility in admissions and in-program assessment'. So, over to you both.
KELLY Thank you so much Eliot. Thank you all for joining us today and we hope you're well and safe wherever you're joining us from. And thank you for taking the time, I know how busy things are especially this time of year and with everything going on in the world so Jill and I really appreciate your time today. To give you a sense of what we're going to go through we're going to spend a little time digging in to how to identify the constructs or what it is that you want to measure and then we'll talk about as well how to put that together in an assessment blueprint. As we work through, we're going to give you some examples that either we've lived through or seen as we've done a variety of different assessments talking about some pearls and common pitfalls we're

going to try and make it practical for you to take some of these tips home. And at the end we're going to pull all the pieces together and tie it into programmatic assessment. So, with that we're going to start off with a bit of a poll and help us get to know you. So, I'm going to launch a poll here and we're just asking to get a sense of the familiarity with your assessments in general, and what your role currently is with assessments, and how you sort of participate and are involved. So, if you can just do the poll within Zoom and complete it once I get a few responses I'll end the poll and I'll share it with everybody. So, that's lovely thank you so much for populating that as we go through. And again, we encourage you as we go through this talk to pop comments or questions in the chat. We'll both pause mid-way through to talk about some of the questions and we'll try and ensure that what we're talking about is relevant to all of you. Perfect so, I am going to share these with you so you can get a sense of where everybody's at, who's joining us today. So, the majority of you have been involved in the creation of some assessments. Somebody is lucky enough to write them in their sleep which is amazing. Everybody's involved at sort of different levels whether that's from admissions or in clinical evaluations as a lecturer or researcher. So, I want to thank you so much for doing that so we can ensure as we're talking through this, we're making it as relevant as possible for you. So, I'm going to pass it over to Jill to talk just a little bit of orientation as we start.

JILLIAN

Hello everyone, I'm so excited to be speaking with you all. So, just before we start to level set as not everybody has been involved in creating assessments, I'm just going to talk a little bit about what an assessment is. So, the first and most important thing to recognise about an assessment particularly within a selection or training process is that it helps you to understand your applicants and trainees and how they can contribute in various ways. Importantly also, it's part of a larger picture. Assessments cannot do everything. We'll be talking about that a lot and about how to narrow them to really speak about what they can do. So, it's part of making a case for competence you can think of it as. So, you're building evidence by having multiple assessments in play. Each assessment does one maybe two things. To have an assessment involves testing, measuring, collecting and combining information it's a very multi-faceted process. So, there's a lot of different components that go into it and its ongoing and iterative which is another key feature because you never have a perfect assessment first time, you never have a perfect assessment. But you have to be sure that you're looking at it and evaluating the kind of data that you've received and deciding how you can improve it for the future. So, when we talk about iterative it's really every time we administer an assessment we go back, we look at what we've learned and we change it for the better.

KELLY

Perfect. And just as we sort of level set around assessments, we want to sort of share some themes around what we're going to be talking about. So, we are going to be working through some case examples as I mentioned and we're going to be talking a little bit about admissions just because that's the sand box that Jill and I happen to play in most days. But most of the principles that we're talking about even when we're working through a case example of admissions are applicable through the entire lifecycle of the trainee from their admissions and probably even before, to medical training all the way through to when they become a practicing doctor. And throughout this entire component as we talk about some of the basic principles, we are going to keep a lense on equity, and we're going to talk about opportunities to bring equity into your blueprint as you go through. So, that's another common theme that we're going to be talking about. And whether we're thinking about the outcomes for graduates or

the good medical practice competency frameworks, there's a variety of different constructs that we tend to measure. Oftentimes they're put into sort of two big buckets if we want to think about academic and non-academic. As Jill and I go through we are again going to be focussing a little bit more on the non-academic components just as we work through some examples sharing our experiences with you but part of the reason we've chosen to do that is those are sometimes a little bit of the harder constructs to measure we struggle a little bit more understanding how to measure those. So, that's why we're using those as our examples but it's meant to apply broadly across all of the constructs. And we're going to begin by talking about blueprinting within a single assessment but understanding again these principles that we're talking about also are applicable for blueprinting across multiple assessments and that program of assessments. And we will tie it back together at the end of this if you're thinking we're only talking about single individual assessments we're also talking about how to think, even if you're not the one creating or choosing assessments, how to think about putting them altogether to make sure that they're reflective of what you want to actually be assessing within a larger program scope.

JILLIAN

And with that note the reason that a larger program scope is so important is that there is no perfect predictor of the future. We know that for sure, that's the only thing we do know. So, what we have to be remembering throughout all of this is that there is going to be error in everything that we measure and that's ok to accept but it's all the more important then to consider defensibility when creating assessments. So, what we hope to do today is give you some of those tools to create defensible assessments whilst still understanding that there can't be any silver bullet or perfect assessment that captures everything.

KELLY

So, let's start off with our most political slide. So, if you're not an LFC fan I do apologise, that's where my background is. So, I want to start with an analogy that we can all relate to before we dig into a specific assessment example of how do you scout a player? How do you know when someone is good? And I'm using the example of Mo Salah here from LFC, and I think about what you want to do to know whether or not this player is good is to see them in as many different contexts by as many different scouts as possible. And you're thinking about them in a variety of different contexts as well. So, you're thinking about how they engage in different aspects of play, in the practice setting, as a team player and even within the community. And keeping in mind that the more aspects you see the broader you can understand but Mo's a great example of you're never going to be able to measure everything in fact there's great data which I think is just a really interesting aside that when Mo joined the LFC there was actually a decrease within the city of Islamophobia. So, there's certain things you're never going to be able to measure but it's important to keep track and think about what it is that you are measuring constructively. And so, this concept of sending multiple scouts to multiple different games or practice settings to see how this player would actually fit within your team and within your community is actually applicable to what we want to think if we're using an example of UME selection. So, this is an example we're going to work through and think about how this would apply, this concept of seeing multiple scouts in multiple games, to what we do in UME selection. And again, this is applicable if you're thinking about assessing competency in-program as well. So, what we know is that when we see our applicants for the first time, it's a little bit like a black box or in this case a blue box, is that we want to know as much as possible about these applicants. And in order to do that the answer is to sample as

much as possible so sampling, sampling, sampling across context, types of assessments, different constructs so those academic versus non-academic constructs we measure across different items, with different raters. The more you do that the more you can get information. The sort of multiple biopsies that you get of applicants the stronger your ability to actually measure their performance on the different competencies that you think are important. And why that becomes important is that the more samples and the more assessments you do the clearer and clearer the picture becomes of who these applicants are. I think of each assessment or each question within an assessment adding pixels to the picture of who your applicants are so, the clearer a picture that you get. And we're going to talk about how that can be valuable. When we're thinking about assessments just to further explain this context if you're not working particularly right now in the field of assessments, you have a pretty big job. You have to think about how do we establish at the time of selection how these people are going to perform in the future, how do we choose the right assessments to use and measure the right things so that we can predict how they're going to do in training, at licensure and further into practice. And as I mentioned before I'm broadly grouping these things into two buckets. So, if we think about the academic attributes that we might want to measure for somebody. Right now, in UME selection we're using tools that look something like the UCAT, BMAT or GAMSAT depending on the school that you're from, A-levels or GCSE's again depending on your school. And those are the things that we're counting on to help us predict for future performance. So, we want to make sure we're sampling broadly across those academic attributes. But we also have non-academic attributes we're trying to sample across and so, in times of interview we're thinking about whether that's a traditional panel interview or an MMI, some schools incorporate work experience at this stage. There's a variety of different tools that we use but our first step is to think from all of our applicants who we can actually bring in for interview and in that case, we're using things like personal statements, reference letters. We're going to talk a little bit about a situational judgement test, specifically about how as a case example of how we might blueprint that and also holistic review. So, all of these components we need to think about what information they're giving us across the different constructs that we're interested in. Quick definitions in case you're not involved in admissions, an MMI or a Multiple Mini Interview is likened to an admissions OSCE. So, originally when they were completed at the school level, they varied at the school so it was more aligned to the school's mission but of course it could be co-ordinated nationally at a higher level based on constructs that are important. Holistic review is individual consideration of experiences or attributes about that applicant to balance out some of those academic metrics and maybe experiences and again also aligned to mission. Now, as a component of what we're going to talk about today Jill and I are going to be talking about the blueprinting of a specific assessment called CASPer which is a situational judgement test that we work on at Altus and the reason we wanted to use this as a worked example is simply because we've spent a lot of time thinking about how we blueprint this. CASPer is an online primarily video-based situational judgement test. So, again measuring those non-academic attributes. But it is adapted for both geography and level so thinking about school leavers and how that differs from our expectations of people entering specialty training so all of those different aspects of blueprinting are things we work with on a day-to-day basis, and why we want to work through this example. So, as we talk about CASPer importantly across the 12 sections it evaluates ten different constructs that sort of give you one general score on somebody's personal professional qualities and we'll walk through what that looks like as an example.

JILLIAN
Ok so, assessment matters. It matters to a variety of people but it matters because particularly with high-stakes assessment real decisions are being made. So, when creating these assessments there's a four-part responsibility. The first responsibility is to the applicants or trainees, or students. You want to ensure that the assessment is accurately as much as possible reflecting the capacities that these test seekers have. So, you must be responsible to them in creating that assessment. There's a responsibility to the program, you want to create an assessment that accurately reflects your program's values, mission statements and classes as well. And there's a responsibility to society because particularly within professional programs such as medical education you will be serving the public so, the society expectations are going to play into the assessment itself. And finally, and perhaps most importantly there's a responsibility to be equitable and what this means is considering the viewpoints of people who have been othered, who have been unprivileged, disenfranchised and ensuring that when you create these assessments you are removing as much bias as possible. There will always be bias, it's impossible to remove it entirely but you want to include as many processes as possible to get as equitable an assessment as possible. And so, we're going to talk a little bit about that and about ways that you can ensure that going forward. So, where do we start? Three things to consider; the first is identify, the second is define and the third is consider. And that consider piece is the equity piece so we'll get to that at the end but it should really be infusing everything that you do as you start to define your constructs. So, identify; what do you need to measure? One big question to consider, what is important in your incoming applicants or the students that you have? What is important about these people? What is important to you about these people? So, four questions; what can you see currently with your existing selection measures? So, if you're creating a new assessment for selection or for in-program assessment, we're going to talk about selection specifically but what can you see currently with your existing selection measures? Do you have things that capture, let's say the academic competencies but you don't necessarily have something that captures communication as a construct, which I will be talking about as my example, as much as you would like. So, you then have to consider what might provide incremental value in your selection process. So, how can you add something else to help you get to a point where like Kelly was saying you have a more pixelated picture, I don't know if I used pixelated correctly, but a more clear picture of this applicant that you're selecting for. The other thing that's really important to consider is what you do in training. So, what skills are the applicants, your trainees going to need to demonstrate in order to really succeed. And then once you've had a chance to consider all of these, where are the gaps? What are you missing in your selection process and how can you create something that fills those gaps? So, you're going to ultimately identify any missing or under-represented constructs. It may be that in, for example, letters of reference you're able to capture some of communication but it doesn't capture all of the aspects of communication in sort of a broader context. So, you want to be sure that you're looking at communication in various forms particularly given that applicants will have to speak let's say which is why a lot of people include an interview or an MMI. And then the big question; what do these constructs mean to you as a program etc. And that's where we get into the defining of constructs. So, why does defining your constructs matter? That's an important question to consider because a lot of the time we think we understand what we mean when we say something like communication but that actually has a variety of definitions to pretty much every individual and pretty much every context. So, it's

essential to know what you want to measure before you try to decide how to best measure it. So, my background is education so, my research has been in pedagogy so I always play in the sandbox of teaching and a lot of you are lecturers we saw so you know exactly what I'm about to say; when you're teaching, when you're giving a lecture there's an objective that you're trying to get your students or these lecture listeners to know about. So, that objective is very similar to having a construct or blueprint before assessment. You create your lesson based on your objective, you create your assessment based on your constructs and your blueprint. The actual creation of the text let's say, the content of the assessment needs to be informed and directed by this guiding principle. And then the other thing that's really important to consider is that evaluators must know the definitions. So, if let's say you're in a Multiple Mini Interview where you have multiple people like community members interviewing your various applicants the evaluator should know what your specific definition of communication is because otherwise, they're going to be bringing their own prejudices, their own biases to the table and evaluating their applicants based on their personal subjectivity. It's not that subjectivity is necessarily bad and we could talk about that in the question and answer period because I think we had to cut those slides unfortunately however, it is really important that there's a clarity and a level setting for everyone going into the evaluation stage. So, defining your constructs, let's talk about how to do that. There's sort of three steps to consider. The first is reviewing common definitions so again we'll stick with communication as a construct. So, the first thing to do is a general literature review. Typically, particularly with non-academic constructs there's not one specific definition that has been given and accepted as true. But there are people who have defined it in different ways. So, even a brief literature review taking a look at different psychometric articles let's say that talk about communication as a construct can help you form a base definition that you can work from. Then you want to go to your competency frameworks such as outcomes for graduates because that will give you a lot of context within the specific profession or competencies that are needed there. So, it's not just one or the other it's looking at how everyone defines it so that it's a you know, a fair and clear definition there. But it's also looking at how it is within your specific area. Then you need to identify specific and relevant behaviours and traits because if we're just speaking about communication, communication is communicating effectively. That's great. There are a lot of competency frameworks including I think CanMEDS which is the equivalent of outcomes for graduates, which basically says that as someone who has a background in English you don't use the word to define it. So, how do we define this in a way that's clear? You find behaviours, specific behaviours to make these constructs a little bit more tangible. So, for example within the competency framework outcomes for graduates' communication has specific behaviours listed under it. Such as you must be considerate to those close to the patient and be sensitive and responsive in giving them information and support. If that's a specific behaviour that you're looking for you can create an assessment that actually assesses for that behaviour because you know exactly what you are looking for. So, that's really important too. The other thing if you want to step away from the theory and more into the practical and your own experiences is to think about trainees who have succeeded in your program or struggled, and identify those stories and look at the specific qualities or behaviours that they demonstrated. And finally, it is very important to consider equity and cultural context in your blueprint, in your training, in your content, in your evaluating at every step of the stage and we'll talk about that a little bit more. So, when you are defining your constructs you include your competency framework as well as your school's mission which is also key. You want to ensure that

you're bringing all of the pieces of what your school actually does to the table when defining. As well as community voices. And this is where we start to talk about not only the patient population but the population of people who are othered or who do not have the same privileges that most people at the table have. And so here we get to equity. So, in terms of equity it's very important to consider a few questions, these are just four, there are many but they're ones that I think are important to consider going forward with any blueprint or any constructs that you define. And this comes from, as Kelly has been saying, we have a lot of hard-won experience around this, about trying to incorporate perspectives into our blueprints for CASPer. Because initially our Canadian version for example was blueprinted to the CanMEDS roles and those CanMEDS roles directly conflict with indigenous ways of knowing and indigenous means of giving healthcare. And so, we have been collaborating with indigenous folks in our community as well as indigenous folks in admissions in Canada to learn more about what that healthcare actually is and starting to incorporate those perspectives into our blueprint. It's a huge process, it's a necessary process we absolutely have to do it because if we don't then we're creating a test that is not equitable. And so, it's an ongoing process as we learn more and we keep going back and reflecting. So, these are some of the questions that we have been reflecting on that I hope will be helpful for any of you. So, the first is to consider who has a seat at the table when defining the constructs or the blueprint for that matter, but when we saw that we had created a blueprint that didn't have these other voices in play it was very important to go back to the drawing board and really restructure how we were thinking about things, in a different sense with people who are able to be at the table who are indigenous, who are black, who are BIPOC people, queer people etc. The other thing to consider is what population your graduates will be serving. So, you want to ensure that you have representatives of the communities such as rural communities that are present at the table so those people are able to give perspective that you might not otherwise have. This is an important one because there's a lot of assumptions we make on a day-to-day basis. What assumptions are we making that come from a place of privilege? So, most of us have some level of privilege what assumptions are we making about how things should be based on that place? And that requires a lot of self-reflection and learning as well as reflection on your program in general. And then also how can we account for implicit bias? Implicit bias is not going to go away but there are ways that you can train your evaluators, as well as yourself, to recognise it and adjust for it, and account for it. It is again a very reflective process, a lot of this has to do with self-learning it's a lifelong journey that we have to do to un-learn our privilege and un-learn our ways of thinking. And I am happy to provide, I forgot to include a citation for implicit bias training, but I'm happy to provide that after if anyone is curious. So, I think at this time we'll pause and if anyone has any questions on the constructs or anything, we've spoken about so far, we're happy to answer. And if not, that's fine too we can wait until the end.

KELLY

So, we're going to give you a few minutes to populate that in the chat and maybe while we wait to see if there's any questions that come up, I'll just sort of present sort of the time through piece because Jill has spent a lot of time talking about constructs and what it is that we want to measure and considerations. And to take through the example of you know measuring admissions or thinking about what we want to measure in our medical schools. We do have that common framework that we have in terms of measuring the competencies that we want for a physician but because Jill has talked about the layering the analogy that I use is that while we all might be trying to

bake a cake and this will resonate for Nick because I think he and I talked about something similar at an INRESH workshop on baking bread, that we all want to bake a cake but there's different ingredients that we're going to use and we're going to maybe use those ingredients in different proportions. And as we think about what those proportions are so we're going to end up baking a cake which is all sort of you know towards the same end goal but they're going to end up looking different and that's influence of our own specific community, our own curriculum, our own mission within our medical schools that end up doing that. I'm just seeing that Helen's got a question here about how we can start benchmarking assessments with qualitative aspects. I think that's potentially, Helen can you tell me what you mean a little bit in terms of, do you mean that they're constructed responses, or that they're verbal responses rather than multiple choice is that what you're asking? I'll just give her a second to make sure I'm going to answer the right question. But what we want to say here is that we're all working towards baking a cake. You can have your list of ingredients but the goal with identifying your constructs appropriately and actually blueprinting them is that it's ok to be different but you need to be an informed different. And I clarified with Eliot at the beginning that this is not my Pinterest fail, this is just a Pinterest fail I found on the internet. I do have my own versions of failed baking but I thought I wouldn't share them with you. And I should say if it's easier you are welcome to unmute yourself to ask your questions if that's easier for you. So, Helen's just clarified when asking observers to form subjective judgements and providing qualitative feedback exactly when observing practice. Perfect. So, Helen when we're asking observers to make subjective judgements, I think there's a few things that are important and I'll let Jill jump into this as well and we'll talk about them. The first is making sure that you've level set. That they have a clear understanding and also, this is for qualitative feedback, of the things that they should be observing and the expectations of that level of performance. So, if you have them looking at a first year versus a final year medical trainee engaged in a similar activity your expectations are going to be different from them and making sure that those are clearly delineated and outlined for the faculty supervisor or reviewer. So, making sure that they understand the level of expectation of performance but also the content that they should be evaluating. We all have biases as evaluators, there's certain aspects of performance that we all tend to gravitate to whether that's communication or empathy or maybe medical knowledge. And so, making sure that people are informed as to what actually needs to get evaluated on different rotations or after a specific lecture is critically important and that's part of the blueprinting process. You're never going to get it right because obviously people's performance will vary significantly and oftentimes that's where the Likert scale on the assessment, if you're thinking about a clinical evaluation is insufficient, you know rating them on a 1-5 or a 1-7 scale doesn't give you all the information you need and that's really where you need to provide those qualitative comments which are the richness of the information. Actually, Shiphra Ginsberg out of the University of Toronto has done some great work on that as well. Jill did you have anything?

JILLIAN

I can just add to that two things. The first is how important it is to define your constructs precisely for this reason. If you were going to have evaluators who are evaluating on communication if you have clear behaviours for what you're looking for it's a lot easier to get consistent judgements and benchmark ratings is what we call them. So, benchmark ratings. If everybody knows exactly what they should be looking for within the definition of communication. And then the other thing that can be helpful, again this comes from my pedagogical background, is rubrics. So, making sure

that you have clear and well-defined definitions of what an excellent applicant does at this stage versus a mediocre applicant, versus a poor applicant. So, having that clarity is one of the most important things that you can help to create a less-subjective assessment.

KELLY

Nick has asked another question so, thanks Jill. He likes the idea of sample, sample, sample and how do you know when it's enough? That is an amazing question because I think that's one of the hardest questions to answer. And we're going to talk about that a little bit as we frame our blueprint to help us identify what it is that we're actually measuring and is that sufficient but part of that feeds into the iterative nature of the assessment. So, do you collect enough information to predict problems or excellence in performance and that you may need to iterate the assessments that you use or chose based on the data that you're getting about your applicants because if you're not able to sort of identify people who are maybe bumping along the bottom in terms of their level of performance, they're never dipping below, they're not failing but they're bumping along the bottom. Then maybe you need to adjust your assessments in a way or identify you're flagging in such a way that you do pick up on those people or maybe you need to add a different assessment that focuses on something else. So, Nick we'll try and address that as we go through the blueprinting talk but I welcome you to add back in questions if you want to dig a little bit deeper into that because I think that is the key question here. We only have so many resources so how do we make sure that we're using them effectively? So just to transition to the next but again please keep populating the chat so we can answer your questions. Institutions' tests may perform very differently so your applicant pool, your training pool influences the assessment that you use so you could, and you need to understand how these tools build on each other. So, to Nick's point when do you know when enough sampling is enough and that's thinking about what is each tool for assessment adding to your understanding of these applicants or trainees. And to think about be different, but be an informed different, I want to give you an example that's a little bit different that's an in-program example. So, I've had the opportunity to work with a variety of different programs and specialty programs as they have been doing their assessments working with them. And I was asked to observe an OSCE for an orthopaedic specialty training program. And when I saw it, I was really excited. They were trying to make it new and different, and add different question types and things like that. And it is very resource intensive for them to put together an OSCE for their specialty trainees and get them all together, and devote the faculty time and the trainee time. And what ended up happening is that as they requested from broadly across their faculty everybody had to create questions and to send them in and everybody was creating their own question. In some cases, they were creating their own marking or scoring rubric on it which is problematic in itself but what ended up happening is they didn't get enough questions. That faculty became busy and they didn't send in questions so it fell actually upon one specific faculty member to create a lot of the OSCE questions. And what ended up happening is that this very resource intensive OSCE that they'd blocked off residents and faculty time for ended up focussing quite a bit on orthopaedic cancer situations because that was the specialty of that particular faculty member who was developing the question and because there was no blueprint to tell them exactly how they should be framing these different questions and how many different questions they needed at different times and on different topics they ended up falling back to their own familiar area. And so, a lot of the assessment ended up being on orthopaedic oncology which skewed actually what was being measured and in fact became a wasted resource of

time. And that's why Jill and I get so excited and passionate about the concept of blueprinting because it can be incredibly helpful. So, let's jump into this because we've talked about multiple biopsies or sampling across your trainees but we want to talk about now how to put those pieces together and connect them. Blueprinting is one of probably the most commonly overlooked pieces in assessments in programs and while we talk about blueprinting our curriculum, when we don't blueprint out assessments as well, we actually end up undermining our curriculum. So, if we're not assessing it oftentimes they won't pay attention. How many of you have given a lecture and people wanted to know if this was something that was going to be on one of their examinations or their licensing exam etc. and that's sort of where they end up focussing. So, if we measure it, they will assign weight to it too. So, we need to make sure that we do create a blueprint for our assessments and that it's aligned with what we actually want to measure and what we intend to measure because otherwise as the example I just gave we're wasting our time and resources for all of us. So, why do we want to blueprint? Well, I think one of the best examples is that it creates a lot of transparency whether that's for applicants or our current students, they understand what they need to do. For applicants they know is the right program I should be applying to? Are they aligned with my scores or how I'm doing or the things that I'm interested in? For students they know whether or not they're up to speed. They know the areas that the programs are focussing on. For faculty we talked about you know if they're assessing somebody in a clinical situation, are they looking at the right things? Are they setting the right levels of expectation? If they're giving a lecture what are they putting their focus on? Is that the right topic? We all know that oftentimes faculty may use a - we use the term can-talk, I'm not sure if that's a common term - but a talk that they've given before and oftentimes that may not be exactly aligned with what we want the students to know. So, the more transparent we can be the better. It also is helpful for programs and accreditors whether that's the MSC or others, it has them actually know that what you're intending to teach and the mission and the goals of your program if you're focussed on you know particularly on rural education how are you measuring that your trainees are going to be set up for success when they go into practice in that area. It also creates buy-in. So, we have linkages with our priorities and what we're assessing so we know we're putting our resources in the right way. It also - to Nick's point on when are we doing enough sampling - can identify gaps, redundancies or when we're just completely inconsistent in what we're intending to do. But I think importantly it also builds a defensible process so both for the individual assessment as a whole that we know this is to some degree a defensible assessment obviously there's other psychometrics that go into that but that's a piece of it. But also, when you think about programmatically do we have a defensible assessment process? And so, factors that go into that, that we can eliminate with blueprinting, is construct over or under-representation. That's when we're measuring too much or too little of a particular construct. So, once you've gone through all the work that Jill's described of identifying what those constructs are that you want to measure, well again like my OSCE example if you just ask one faculty member to create the assessments and they have a particular interest or area of specialty you're going to end up with too much of some constructs and too little of something else. So, while orthopaedic knowledge, or disciplined knowledge is important what we've actually seen when we break down what that disciplined knowledge is that was expected is that we had too much orthopaedic oncology and too little of foot and ankle, or spine if I'm getting those correct. But it also can help us make sure that we're actually thinking within and across assessment what is important to our program? And then in the case of selection at

what stage of selection you use some things. So, we're going to talk through an example of if you have a measurement that you use in selection so, perhaps their GAMSAT score or something, if you have that and you're using it to identify who you want to interview and then you're also incorporating that score later on to evaluate who you're going to give an offer to, that's absolutely fine but you just need to be aware that you're double weighting that particular assessment and therefore potentially over-representing those constructs again unless that's what you want and that can be done intentionally. And then you can also have something called construct irrelevant variance which means you're measuring something outside of what it is that you actually want to measure and it impacts your ability to actually use the score to understand competency etc. So, I give an analogy here of a simulation setting where a team is actually going to intubate a patient and you want to measure team communication during a difficult intubation. Well, you could be measuring it in this team setting but I think one of the things that we would think about is construct irrelevant variance, is the actual simulator that you're using. Potentially it has sort of a rubber tube that requires a certain amount of lubrication in order to properly do the intubation. If the person being assessed doesn't know that they're going to struggle and it's going to cause behaviours that are completely outside of what you would want to measure and are not representative of how somebody would actually perform in their actual true behaviours because they're going to be struggling with the intubation and perhaps getting frustrated with the simulator rather than actually being able to focus on maybe the team communication that you wanted to measure that would be reflective of a real life scenario. So, questions that you might want to ask yourself include Jill's work on what are the important constructs? What are the key aspects of the constructs that you really want to measure? So, we talked about constructs being very complex like we'll talk through communication for example. What aspects of that construct are really, really important for you in your program? Where are you trainees struggling? What are the aspects of it that you've measured in other assessments and so therefore what is still left to measure? And how much weight you should be giving to something, how many times you need to sample something and then thinking about are you doing it with defensible or reliable measures? And that actually allows you to create this beautiful matrix and so I'm sharing this as an example, this is obviously a bit of a made-up example of our cast for blueprint. So, we have 12 different scenarios, we have ten different constructs that we want to measure in these scenarios, you could imagine this likened to the different constructs that you want to measure across a multiple-choice examination as well. So, the way that we've particularly framed this is that we have primary, identified in the darker blue, as A and secondary constructs, in sort of a very light blue/grey as B. And what we try and do is lay out each test to identify what is our primary construct that we're measuring in each scenario? What is the secondary construct that we might want to measure because in these sorts of things it's complex, if we're measuring communication, we're not just going to measure communication we also might be measuring things like motivation or resiliency in their response because to Helen's point when you have these sorts of qualitative responses, they tend to get complex. So, having in our case our raters or in other cases faculty members understand what is the point if you're sending them into an OSCE that you actually want to measure in here, it can be really helpful and directive to know if they're aligned on measuring the same thing across different trainees coming through. And so, now we just want to quickly pull all this together in turning these individual assessments into a program of assessments. And why this is so important is that we've talked about all these little pieces together but we need to think about how we pull

them together in the bigger picture. So, let's talk about the construct of communication and we work through this a lot but we wanted to sort of carry this through. Communication doesn't happen in isolation as we've just sort of talked about. There's verbal, there's written communication, there's non-verbal communication, there's empathy that overlaps with it in breaking it down. So, the more we take these larger constructs and we break them down into their sub-components the more accurate or aligned our assessments end up being to the things we actually want to measure. And that could be quite overwhelming but when we think about communication it actually overlaps with things like collaboration and empathy. It's not just happening in isolation as we've talked about before. And so, I've used the example of an interview as well as CASPer which are both focused on measuring non-academic attributes but what I'm trying to show with sort of this Venn diagram is that they're identifying different pieces, that they're focussing on different aspects and that even when you're using both of them, you're actually measuring different pieces of communication. And what that means is that the more times you're measuring communication the clearer picture you get of somebody's competency in that larger construct. So again, highlighting the value of needing to have these multiple assessments across a construct as well. And so, this was the example I gave earlier of thinking about how you're strategically, in our admissions example, combining assessments so that they might be measured intentionally multiple times and then you take that matrix that you had before where you do it across a single assessment and I've just simplified a version here. I'm happy to share a much larger version of this, I just couldn't fit it on the slide in a way that was readable, where you actually think about what are the constructs that are measured by different components across the different constructs that you want to measure. And, in this case, I gave the example where a construct, and that might be communication, you actually break it down into sub-components and you think about how that might be labelled or identified through different assessments. And you can sort of get through that matrix understanding of whether or not you're measuring the different components enough or appropriately across.

JILLIAN

So, David had a question in the chat which I might be able address a little bit here but I'm sure Kelly will have other things to say. So, some take-home thoughts before we open it out to more questions. Three take home thoughts; there's no perfect solution so we've talked a lot about sample, sample, sample. Sample across multiple constructs, across one construct to get as clear a picture as possible. Point number two; map a defensible assessment and process which is aligned to the school or the program's mission and goals. You want to identify your gaps, define the constructs and blueprint and that will give you that defensibility when looking at the evaluations. And finally measure what matters. So, ensure that you're not spending too much time measuring construct irrelevant variance and you must consider bias so, this goes back to the equity conversation that we've been touching on throughout, construct relevant variance, and consistency between instances. So, this David, is where I will speak a little bit to and Kelly will continue and build on mine, but I think the key with consistency between instances is to have very clear guidelines particularly for how you develop questions. So, ensuring that you can return to specific aspects of a question that you're including so that the questions might differ but they're targeting the same kinds of behaviours. And this comes back also to a blueprint. So, the reliability necessitates a blueprint that you return to again and again that highlights the same, well-defined constructs and that gives you a reliability that you would otherwise not be

able to have particularly with qualitative assessments. And in terms of assessors the training is key and making sure that they understand each construct that you're going to measure. So, I'm speaking to CASPer specifically because that is my experience, my sand box, so you want to ensure that every assessor or rater has a very clear understanding of all of the constructs that you're measuring so that they can pick out these different pieces as they look at these complex responses.

KELLY
And David maybe one thing I'll add to that is that the importance of piloting your questions and having them reviewed because some of those things like construct irrelevant variance especially if you're using different forms or different versions of an assessment across different times. So, very simplistically using an example of an OSCE where you want to run two different sessions of an OSCE or even an MMI where you want to run it on different days, making sure that you have the scenario reviewed not just by other faculty members but ideally perhaps by your current trainees maybe who aren't doing the assessment or others who can make sure that it's going to be interpreted as measuring the same constructs underlying. And making sure there's no construct irrelevant variance in one which would make it perform, of course there's a variety of different item analyses that you can do in terms of differential item functioning and things like that which can look both across item performance but can also be used to determine whether items are performing differently for different sub-groups in your population as well. So, there's lots of things you can do about that and I'm going to just pause here and ask if there's time for maybe some questions, of course Jill and I are willing to stay on. And of course, happy to dig into any of these other questions as well. So, Eliot maybe I don't know if you can see anything?

ELIOT
Yeah, so I think that was the only questions so far. There's a question from Ramya so, how can you explain the difference between an assessment and an evaluation?

KELLY
Jill did you want to? Oh perfect. So, that's something I would say that sometimes the term evaluation is used interchangeably with assessment for some people. I think we try and think of them a little bit different. So, assessment is actually the tool or the items that you're using to measure a particular construct and sometimes we think of evaluation as sort of the program of evaluation of the process. I think it differs a little bit between countries as well. So, I'm not sure if that's the proper UK definition Eliot if you want to chime in there, I'm just talking about how we have defined them as separate like thinking about a program of evaluation as a bit of a QA process but an assessment is actually the tool itself.

ELIOT
Yeah, I think broadly similar in the UK context. I think we would usually consider evaluation in terms of the quality of educational programmes and assessment being of individuals.

KELLY
Yeah, and I think there can be assessment of teams and things like that as well. So, I think that's great. Thank you Ramya.

ELIOT
Are there any more questions for Kelly and Jill?

KELLY
And of course, if you want to dig into any of these topics feel free to message us. I was saying to Eliot earlier when Jill and I get together to talk about blueprinting even as

general concepts we get quite excited, so if you are one of those people who want to dig into this a little bit more, we are more than happy to have a virtual coffee and chat about these things in a lot of detail because we try to make this a bit of a high-level overview but happy to dig into any of these concepts.
ELIOT Well, I think unless there's any immediate questions I just want to say thank you very much Kelly and Jill for a really fascinating talk. It looks like from the comments that people have really taken a lot away from it and certainly I've found it interesting to see. So, the only other thing I wanted to say was just thank you to everyone that's come along and participated in the session. Just to remind you that the video will be available on the ASME website in a few days and to keep an eye out for any future BITESIZE sessions. So, the next one is on Wednesday 16th December at 16:30GMT and that's a coffee break session for anyone's that's missed those coffee break chats that you have at conferences this year. So, have a look on the ASME website and sign up for that if that's something you're interested in. And yes, thank you so much Kelly, thank you Jill for a really great talk
KELLY Thank you all for coming and your time and thank you Eliot for moderating so well.
JILLIAN Thank you everyone.
ENDS