The eLearning Coach Podcast #39
ELC 039: How to Plan, Design and Write Tests With Michael Rodriguez

theelearningcoach.com/podcasts/39

Hello learning people, you're listening to Episode 39 of the eLearning Coach Podcast. And with today's topic you can experience true nerd joy. I speak with professor and psychometrician Michael Rodriguez, PhD, about planning, designing and writing tests.

Michael teaches courses in educational measurement and assessment at the University of Minnesota Department of Educational Psychology. So why dedicate an episode to testing? Well, sometimes learning designers and educators think of a test as something that just needs to be thrown together at the end of a topic or course. But, really, testing is an enormous opportunity to better understand learners, to improve instruction, and to let learners know what is most important in a subject area. Michael provides lots of guidelines and advice on this. You can find the show notes for this episodes at theelearningcoach.com/podcasts/39. Here's the interview

Connie: Hello Michael, and welcome to the eLearning Coach Podcast.

Michael: Happy to be here.

Connie: Just to make sure that everyone understands the kind of work or research that you're interested in, can you explain what the field of psychometrics is? It's not exactly a common household term.

Michael: Yeah. And I had the same kind of reaction to the word psychometrics. It's not a term that I knew as an undergraduate student, and I was a psychology major, and psychometrics is a word that comes out psychology. I didn't learn it until I was in a research position following my undergraduate days. Psychometrics literally means 'mental measurement', psycho-metrics. And it's really the science of measurements, whether that measurement occurs in psychology, social sciences, education, it's the measurement part that is not physical measurement. We know a lot about physical measurement, we use rulers and scales, and essentially psychometrics is attempting to create something like a ruler or a scale in order to understand human characteristics that are more mental types of characteristics, like knowledge, skills and abilities. And so you'll hear me use that language a lot, the measurement of knowledge, skills and abilities. And that's what psychometrics is really designed to do, to create and use the scales and rulers that help us understand human knowledge, skills and abilities.

Connie: Great, thank you for that. Are you familiar with a typical instructional design cycle?

Michael: A little bit. I have some colleagues that do some work in instructional design. I haven't myself, but I know what you're after.

Connie: So where do you think that writing test items should be done in the instructional design cycle, because so often, even though many instructional design models have after-writing learning objectives, in practice people often write them at the end because that's when they have a better understanding of the topic.

Michael: Ideally, in my mind, test items can be developed throughout the entire instructional cycle. And I think it's best to begin when instructional design begins, at the very point of planning. Because, in part, thinking about how we're going to assess students' or participants' knowledge, skills and abilities really helps us ground our thinking in terms of what are the knowledge, skills and abilities that we need to be addressing through instruction and through instructional activities. And so having both in mind, what do we want students to know and be able to do, and how are we going to know that, helps us integrate those two concepts. And there are instructional designs perhaps that are more effective and help us achieve our outcomes, our intended outcomes. When we often get the best items, however, during instruction or from interactions with students and participants, we can keep a list of possible topics, issues, ideas that form the basis for test items during all of these phases. But, at the very least, we should take a moment after each instructional period or activity and write down ideas for possible test item topics. Otherwise, we do default into waiting too long in writing all the test items surely before the test date. And that really limits the power of items to reflect important topics and issues or challenges that students face during instruction or during training.

Connie: I see what you're saying. Now, a lot of us develop only for either a virtual classroom or, even more so, for self-paced eLearning, and in those cases we really often don't get much feedback from the learners, which really puts us at a disadvantage.

Michael: Yeah. In those cases where there isn't a lot of interaction, always trying to pull from students their current thinking in some way will help us write test items. Perhaps there are other kinds of activities that go on through instruction that are more formative, and that is just checking in to see how students are thinking about various topics or the kinds of questions that we might get from students.

Otherwise, the instructor really has to reflect on what are the kinds of errors that are possible, what are the kinds of misconceptions that, as an instructor, we have had to address in different studies, what are the ineffective problem solving routines that people might engage in, and use those in order to think about creating test items that might help us understand do my students or my participants hold these misconceptions or have the problem solving errors.

Connie: If there is a subject matter expert and we're just doing the design for that subject matter expert, then we can probably find those things out through the audience interview, and I'm always promoting the idea that we should never design anything without getting to know the audience quite well, so that goes along with my philosophy anyway.

Michael: Yeah, it's definitely the best way to start thinking about assessment, knowing where students are at, knowing what they know and can do helps us then avoid having to retrace all of that and get redundant information, rather than to help people move forward from where they're at.

Connie: Can you talk a little about validity and its role in test development, because I suspect that many of the test items that we write in the world of adult learning, they're often just thrown together so that someone can mark off that they have taken a compliance course. But if we really cared about assessment and used it as feedback to go back and improve the learning experience, I think that we need to know more about validity and reliability. So can you explain validity and its role in test development?

Michael: Yeah, that's a great question. And validity is a great place to start having conversations about test development and assessment, because it is really the most important concept in psychometrics, and the most important characteristic of all assessment and measurement activities. And actually there is a lot of misunderstanding around the concept of validity, so let me just start by giving a simple definition, which I think is current in the realm of educational and psychological measurement. Validity is the extent to which evidence supports our interpretations and uses of test scores. So, I'll say that again, validity is the extent, which means it's a continuum, to which evidence supports our test score interpretation and uses. So it's not an either/or, oftentimes people say things like 'the test is valid', or 'the test invalidated', and there is not a point at which validity has been achieved. It's sort of an ongoing and continuous characteristic.

But the other important thing is, validity is not really a characteristic of the test, we don't validate tests, we validate interpretations of test scores and the uses of test scores. So, because we say a test is valid doesn't mean we can use it for any purpose. It's validated for particular purposes, and that is we actually validate the uses and purposes and interpretations of tests rather than the tests themselves. And so we engage in things like validation. Validation is just the gathering of the evidence to support interpretation and score use, and whenever we introduce a new use for a test we have to gather evidence for that particular new use.

In educational and professional training setting, we mostly care about making claims about what students know and can do, and we want to be able to say something about the knowledge and skills of our students, given the topics of instruction or the learning goals. So, to support those claims about what students know and can do, to validate our interpretation of what students know and can do, the interpretation of the test scores, we need to defend the content balance of the test. So we need to collect content evidence, and for a long time this was referred to as content validity. But knowing that there aren't really different kinds of validity, that validity is really the extent to which we have evidence, we know there's many different sources of evidence, so we collect content evidence, do we have the right content on the test? If we want to make claims about what students know, that they have specific and skills, then that knowledge and skills must be represented on the test. And so, test content is key in most instructional settings and it must be. The test must clearly represent the important instructional and learning goals and objectives, and so we can support our claims about what students know and can do as a result of their test performance.

Connie: Okay, but aren't there ways to improve the validity of an individual test item by following certain guidelines?

Michael: Yeah, absolutely, we can write questions that we think tap certain kinds of knowledge and skills, but the way we write that question is really also critical. Every item is a building block for that whole test. And so we have to be really careful about the way we write items. And we have some guidelines that provide us with really important advice to avoid giving students clues to the correct answer, to avoid including information that is irrelevant or extraneous and confusing, that interferes with accuracy, accurate measurement, do we really know what students know and can do if we are asking questions that are irrelevant? So there are great guidelines that are available to help us.

Connie: Okay. You gave us three, do you have more just little quick tips for making sure that each item is valid?

Michael: Sure. There are some great item writing guidelines that are out there, and most of the people are talking about item writing guidelines use a common source and common set. Many of these come from research and are empirically based, but most of them come from practice, and we learn a lot about what kind of items work well and result in high quality test scores and what kind of item characteristics introduce measurement error and reduce the quality of test scores.

So, a lot of these just come from practice. But there are a set that we believe strongly in and they mostly are designed to avoid giving clues to students. But it also helps create results and test scores that we can defend so we really are sure about what students know and can do. So things like every item should reflect one specific idea or element of the relevant content and one cognitive task. Maybe later we can talk about what that means. It's important to keep the content of one item independent of the other items, and always our items should be testing important content, not trivial or irrelevant content.

It's important to keep the language complexity at an appropriate level, given your test audience, so that we make the test accessible to all whom we might be testing. It's important to minimize the amount of reading, unless it's a reading test, because oftentimes our reading will interfere with our assessment of what students know and can do. We also know from empirical evidence that it's important to write the questions or the phrase in our test items, to write them positively and avoid negative phrases and negative words. People often overlook the negative term and will get the item wrong for the wrong reasons.

There is a lot of evidence about the number of options, how many options should a multiple-choice item have, for example? Three is usually sufficient, and in fact there is over eighty years of research on the number of items, and all of it unanimously endorses three options as being optimal. But what we usually recommend is to write the number of options that is really appropriate for the item. Because sometimes for some items there might be four or five options that are really plausible, and we want to include all of them just to be complete. So, there is sometimes multiple misconceptions or multiple ways that students can make errors, and we want to include them so we know what kind of misconceptions students have, and so they can be included. There is psychometric reason why every test item should have the same number of options. So the number of options should naturally reflect the nature of the question. Three is usually sufficient.

Connie: Now, that's kind of interesting, because I was always under the impression from my reading that one test should have the same number of options on all questions, and I guess the reason was that it lowers the extraneous cognitive load if it's consistent, but that hasn't showed up as significant in research?

Michael: That's right. In fact this is a very common misconception in item writing. And what happens is the item writer will then try to create every item so that it has five options, and then many of the items have options that are just not functioning, they're not really plausible, they use filler options like 'none of the above' or 'all of the above', which are also highly recommended to avoid because they introduce all kinds of measurement error generally. And so people are forced-- and it's really hard to come up with four or five options that are really good.

Connie: It really is.

Michael: And we rarely see them function. And so when students respond to those items, they say, well, for example, C and D are clearly not correct, so now I'm just going to try to figure out between the others. And that is an important test taking strategy, but it's one that we should be avoiding from the item writing design side, that there's always options that students will discredit or discount. As an example, the GRE, the Graduate Record Exam, that students take for admissions to many graduate programs, it now employs items with anywhere-- I think some items might actually have two option, but definitely three option, all the way up to seven options. And so even on that very highly developed and standardized instrument, items have many different numbers of options. And the number of options is really just relevant to the nature of the item.

Connie: Yeah, that is really interesting. Speaking of poor answer choices, I am always working with subject matter experts trying to get them to perhaps not put a negative in, and they only listen to me maybe 50% of the time. But one big one that they always want to use is 'both B and C' or 'both C and D', and I'm guessing that that is not a good answer. What can you say about that?

Michael: Yeah, that's right. There is some empirical research on that, and it suggests that, one, it makes the item harder, more difficult. And, two, it might actually interfere with our ability to measure knowledge and skills clearly, because it's often confusing to students. It's not recommended, and particularly for instructional purposes. Where it does appear to be used the most often and perhaps more effectively is in the health sciences, but oftentimes in health sciences there are multiple presenting symptoms, there are multiple treatments, multiple routes to take, and the important piece is for students to distinguish between which combinations are relevant and which are not.

Now, there is another way to have that kind of item in a way that is perhaps more functional, and one of my educational measurement mentors is a real advocate of this format, it's called the multiple true/false format. It's essentially a multiple choice item that has a question for a stem and has multiple options, and now we can actually have many more options, you could have five to even ten options, where for each option we have the student state whether it's true or false, or whether it's appropriate or inappropriate, and whether it should be considered or not considered.

So that way we get much, much more information, because now students are providing a response to every option, instead of just pick the correct options where we sort of default to-- well, if you haven't picked an option, that means you think it's not true. And so to be more explicit about it, you can just have students declare for each option, is this true or not given the question of the phrase in the stem.

Connie: You're saying that's clearer essentially to the student than a multiple-selection multiple-response?

Michael: Right. Because it's never clear to the student how many options should I select, if I select one or two is that sufficient, should I select the best one or the best two, how do I know? And instead of trying to make all of those other decisions, we tell the students, you know what, respond to each option and tell us whether or not you think it's true or false. It's just a little more direct, I think.

Connie: Often in the eLearning authoring tools that we use, we typically have templates for different types of question formats, and I can't say I've ever seen the one that you're talking about. What about true/false questions, I was always under the impression and from my readings that single true/false questions do not necessarily provide good measurement because it's so easy to guess, what would you say about that?

Michael: Yeah, that's the one danger with the true/false item, you have a 50% chance of just guessing correctly. But, it turns out, if you have enough of them, the probability of getting a very high score because of randomly guessing is pretty small. The more you have, the smaller that probability. I use true/false items in my classes for quizzes, because you can write very clear declarative statements, and it's a great way to check a lot of content. You can cover a lot of content in a single declarative statement.

Rather than a multiple-choice item which has multiple options and requires more reading and more consideration, you have to ask less of them in the same time. So if you just want to take ten minutes, for example, to do a quiz, you can cover a lot more content with a series of true/false statements.

And when I write true/false statements, they're often based on really hard concepts, so it's not simple knowledge or remembering kinds of skills. And I try to select the more challenging kinds of considerations. And I write it so that it's positive, and I write it so that it's negative-- or not necessarily negatively worded, but I write one so that it's true and then I write it again so that it's false. And having to write the true and false phrasing allows us to be more clear and direct in the statements themselves. I think one of the errors when people true/false items is they just write it either true or they write it false, and they haven't really considered what the alternative is to make sure that it's really true or false and really clear. And then I go through it and I pick some of the false and I pick some of those true statements and I do a combination.

Connie: Okay, so you're saying you write true and false to ensure that you're writing an accurate statement, but then you don't use both?

Michael: That's right.

Connie: That's a really good tip. Thank you. Let me ask you something if we can go back a little bit, how do you recommend that someone plan for designing a test?

Michael: That's a great question, and it's a really good question in the context of instructional design. So we have been thinking about test design, because it really is a big process, and, ideally, the test is being designed throughout the instructional process. The basic framework for a test should really be defined prior to instruction and should reflect the instructional and learning goals and objectives. But to create a strong test, no matter how short it may be, always start with a test blueprint.

Just like we use blueprints when constructing structures, like buildings, bridges, monuments, or even technology devices, we have blueprints for new technology, we need a guide to develop a test that results in the intended product, which is the tool. Psychometrics is about designing tools and rulers that allow us to make claims about what students know and can do.

So, a test blueprints simply lists the content topics that need to be covered on the test and the kinds of cognitive abilities are tasks that should be represented in the items. Really, those two things make up most test blueprints. The content covered on the test should be based on the learning objectives, so learning objectives should be clearly articulated prior to instruction or training. And then these learning objectives can be translated into specific topics of knowledge and skills, and that forms the basis for the test blueprint.

And then the test blueprint also describes those cognitive skills, and we usually include things like remembering, understanding, applying, analyzing, evaluating and creating. These are the typical instructional, relevant, cognitive skills. So we can develop test items and tasks that require those skills so that we can support our claims about students' abilities.

Then the final thing that a test blueprint should do for us is to provide a guide regarding how each content topic and cognitive skill should be represented on the test. And we've all taken those tests where there are too many items on a small topic that was not emphasized in class, and it's like where did all these test questions come on this, what we thought was an irrelevant issue or topic. And maybe it is, maybe the topics are too specific and there's too many items on it. So it doesn't really capture the core content and learning objectives. When topics are not appropriately represented on the test, students get the wrong impression about what is important and valued in the course or in the field of study. And then we as the instructors, we have the wrong information from where we can support claims about what students know and can do. And really the ideal blueprint states what percent of items should cover each topic area and cognitive skill, so that we secure balance of items across the topics that is appropriate and intended by the learning objectives. This test blueprint is the strongest source of validity evidence to support our claims about what students know and can do. If we want to make those claims, we have to make sure that their knowledge and skills are appropriately represented on a test, and a test blueprint will secure that for us. What percentage of items are on this topic, or what percent of items are on this topic, and what percent are knowledge items, what percent are analysis items, what percent are application items.

Connie: Now, those cognitive skills that you mentioned sound like an updated Bloom's taxonomy, is that what you're talking about?

Michael: That's exactly what it is. Bloom's taxonomy has been updated a bit to be more instructionally relevant, and so those terms I think are just put more in the context of what we might think is more appropriate in instructional settings.

Connie: I keep coming across a debate about Bloom's taxonomy in general that it was never based on any sound research, and that there are a lot of alternatives to it, what is your opinion of that?

Michael: Yeah, I think that's right. In fact I have read some research that shows that those cognitive tests are not necessarily hierarchical or ordered in the way that Bloom intended. Bloom made an argument that in order to understand something you have to have some basic knowledge, and in order to apply something you have to understand it. And actually that hierarchy doesn't really hold true across all six or seven levels of cognitive tests. So I think this revision, there have been some revisions and updates on this for instructional purposes, they seem more realistic. The other thing that people are not able to do is if an item writer writes an item that is supposed to tap, for instance, analyzing or evaluating, then another person comes and reads that item, they don't necessarily see the same cognitive task that was intended. And so that kind to calls into question, too, the accuracy of those classifications and whether or not they're really meaningful. But they're absolutely an important to guide for us to think about the kinds of skills we want students to have. And most of the time when I'm writing test blueprints I break it down into three basic areas: the knowledge-- well, actually two, I combine knowledge and comprehension, remembering and understanding, and then application. And in application I try to do a variety of things like analyzing and evaluating.

Connie: I was thinking about that when you were saying that a lot of reading in a test, perhaps you were saying, reduced the validity of it, or it's just a general guideline to ease the cognitive demands of a test. However, I was thinking that one of the ways to use multiple-choice items that will test or assess critical or higher level thinking, you almost have to put in a lot of reading because you're usually presenting problems that someone needs to solve or analyze or evaluate. What are some ways that we can test higher order thinking without using essays?

Michael: A typically approach is what you described there where you create a scenario, and of course a scenario requires lots of text and reading. One of the best ways to assess higher level thinking is have students consider the things they've learned in a new or novel situation or a novel context, or introduce a challenging condition, and sometimes those can be done pretty briefly. Another one is through the use of graphical displays, and it doesn't necessarily have to be numerical, it could be pictures or even a video, and so the context could be presented verbally. There are ways to introduce reasoning in the item, and have students provide reasons. This is an effective tool for extended multiple-choice items, you could present a specific piece of content knowledge in a multiple-choice item and then ask students to defend their answer or give a reason.

Connie: A rationale, yeah, that is a good idea.

Michael: We can ask students to interpret something, interpret a problem, or hypothesize, or predict outcomes, given some condition or a change in conditions, sometimes those can be introduced really in a short way. We can ask them to determine which principles or components are relevant to an issue. It's also interesting to get students to think ahead, so what's the next step in a process, what's the course of action that might be appropriate or useful, given some problem or condition.

Connie: Those are a lot of good alternatives. Thank you for that. Let's go back to if someone's planning a test, they're looking at the cognitive skills, whichever criteria they're using, and then they're thinking, well, what kind of selected response options do I have. What are some criteria you might use to select which type of item format you're going to choose?

Michael: That is a good question. Let me open that question up, first to the bigger issue about all the possible different kinds of formats. And then in most online instructional settings there are some practical constraints of course. But, ideally, the item format should really meet the purpose of the test and support the intended claim about what students know and can do. And, of course, I'm talking about validity, so we want to pick an item format that is going to support the validity argument. And if we're making claims about what students know and can do, it's best to get evidence that is direct rather than indirect. For instance, if we hope to make claims about student skills regarding the creation of an innovative product, then we have to create a task that allows students to do so, which might mean a performance assessment or some other kind of constructed response format where people actually create something. We can ask selected response items about ways to create products or the kinds of basic knowledge that's required to be innovative and to create something, so we understand that people have the basic knowledge and skills in order to do that. But the actual product creation itself is not easily observed through selected response items.

However, no other format allows us to cover a broad range of topics and skills than selected response items. And particularly in the multiple-choice format, we can cover a lot more knowledge and skill areas in a short period of time with well-developed multiple choice items. So sometimes it's really practical limitations that makes the choice for us. When we have a limited amount of time, when we have a limited way of scoring, multiple choice items are often the best choice, not just by default but also because experience has shown us that we can do a good job of measuring knowledge skills and abilities with carefully developed multiple-choice items, and that those measures have been shown to be important indicators and predictors of future achievement and job performance. So it's not just, oh, we have to do multiple choice items, there is a lot of benefits and a lot of strength in high quality multiple choice items.

Connie: Yeah, that's really good to hear.

Michael: The kind of selected response items we write then should also serve the purpose of the test, and really should reflect the nature of the content being tested. And we have a variety of multiple choice item formats, we've got just the typical multiple choice, there is actually a multiple-choice format called the alternate choice, it's this or that. I'm not a big fan of that because it's more like a true/false item, and I think true/false items are probably more effective than alternate choice multiple-choice items where there are just two options. Then there are the complex multiple choice items you described earlier where it's like 'A and B', 'B and C', and we try to avoid those in instructional settings.

And then also matching items, matching items are really quick, they're easy to respond to, you can cover a lot of material, but it's almost always restricted to remembering, the skill of remembering. And if we want to check people's ability to match things, persons, places, dates, but it can also match characteristics, definitions, things like this, it's a very quick and easy way to do it. And my recommendation is often if you're just testing recall and remembering, don't take a lot of time by writing lots of multiple choice items, just put together a great matching exercise. And remember in the matching exercise that one of the lists should be longer than the other so that students don't just do process of elimination to figure out which one is which. And also provide for opportunity that in the one list from which we select the characteristics or the matching options, that it's possible that some of those could be used multiple times, or that maybe some of them aren't even used at all. And it creates more of a challenge, it requires students to think more deeply, and that's another way to get a little more sort of analysis skills in the task, so we can structure it in a way that does that.

Connie: So you're saying have an uneven number of definitions versus the terms?

Michael: And maybe some of the definitions don't even apply to any other terms. Or maybe a single definition applies to multiple terms.

Connie: Before we wrap up, I just wanted to ask two more quick questions. What are some common errors that you see novice test developers make?

Michael: We actually have some research on that. It has been an interest measurement folks and item writers and educational management trainers in terms of what are the things we need to look for and address in test developers. And a lot of it comes from teachers, because of course teachers are doing more testing than anybody, they're writing tests all the time, but they're also assessment, different kinds of assessments all the time. Some of the most common errors I think are having the wrong content balance on the test. That the content is not really balanced in terms of how much time did we spend on the different topics, how important are the different topics, and what is important in the field, and what do we value in terms of knowledge and skills.

It's a really hard to thing to kind of balance. So we might actually spend more time on a topic, but, relative to the other topics, it might to be as important, it's just that it took more time to cover. And so it's not just a simple reflection of how much time did we spend on each topic, but it's also what's important and relevant in the subject area and in the field. This is a great thing about instructional design. So if we're talking to people who care about instructional design, then we know that people care about what is important and relevant in the subject area and in the field. Many teachers, and this includes K12 teachers and college and university teachers, we don't have the training in instructional design, oftentimes explicitly, so it's hard for us to think about how is everything we're doing conveying to the students what is relevant and important in a subject area or in a field of practice. And we aren't really good at writing either instructional or learning objectives. And without having clearly written learning objectives, we don't have a ruler or a guide to go back to, and students never really get a clear message about what's important or what's relevant or what's valued in a subject area or the field.

Oftentimes they get the clearest message of what's important by what shows up on the test. And I even see this with my graduate students. My graduate students still ask the question, Is that going to be on the test? Because if it's on the test, that means it's important. And that's kind of what we've trained our students to believe and understand, because we haven't done such a great job on the instructional design side. So if we're not really careful about how we balance content on the test, then the clearest message a student gets about what's important in a subject area is poorly constructed, and they're getting the wrong message about what's important and relevant and valued in a subject area and in a field. So it puts a lot of pressure on the test developer and on the instructor to really think hard about how do I want to represent what's really important on the test. So, wrong content balance, clearly the most significant and most common error novice test developers make.

And then there's things like poorly constructed options in the multiple-choice item, that are options that are just not plausible, sometimes they're even absurd. Sometimes instructors try to be humorous by putting in a humorous option. And there is a little bit of advice about using humor in tests. If humor is an important part of instruction and humor is in the training activities, then it's probably okay to use a little bit of humor on a test, sometimes it sort of breaks the anxiety or breaks the ice a little bit on the test, but if it's a high stakes test, if it's a really important test, it's probably best not use humorous items or humorous options, because it will confuse some students. Students will wonder "Is the instructor being serious, is this a real question, am I not understanding something?" And it might actually increase the anxiety level of students, because test anxiety is a real thing.

The most common item writing error that teachers make is writing the correct option so that it's much longer than the wrong options. It is the biggest clue to students, this must be the correct answer because it's the only option that has a conditional phrase, it's the only option that explains something. And teachers often do that, they put more information in the correct option because they know that there will be some students that challenge the correctness of the option. So teachers build in a defense and make the correct option longer, inadvertently, without realizing it. So this is an important item writing guideline, that all of the options should be about the same length.

Connie: Right, that's so important. That was good to hear about those errors. I'm just curious in terms of academic research, what don't we know about evaluation and assessment, and what kinds of things are people researching now?

Michael: There are maybe two areas. One has been underway for quite a while, and one we're just getting started at. The one that's been underway for a while and we still haven't figured out is how to use information from the wrong answers. Particularly with multiple choice items we worked really hard to write distractors or the wrong answers so that they actually contain information, they contain information about what kind of errors students are making or what are the misconceptions students might still hold. And some of those options are more important than others. And so there's actually information in the wrong answers, and we don't use it in scoring. Really good instructors will use it as instructional feedback, what the errors that we need to fix, and what are the misconceptions we need to clear up that students still hold. But in scoring, especially with the large scale tests, there are ways to extract information from which of those distractors did students select, and some of them contain more information than others.

So we're working on that on the technical side, but most of that occurs with the large scale testing.

The other area where there's a lot of research that's gearing up regards technologically enhanced items. Because we now have so much testing that occurs online or through computers, we can do all kinds of really interesting and innovative things with test items and have students engage with test items in many different ways. So now instead of writing a multiple choice item, like which sentence in the opening paragraph exhibits the main idea, normally we would write four or five options. But now we can actually have students go back to the paragraph and highlight the sentence of all the sentences. So if in the first paragraph there's ten sentences, there's actually like ten options, and students can highlight the item or highlight the entire sentence, or highlight a word, students can move things around. This is very interesting in the area of architectural design. Students can go in and move things around in a design, they can correct errors, they can look for things, they can create new things. So there is some ways, but it is more interactive, and of course it requires a lot more test development. Even things like computer programming, we can have students go in and identify inappropriate code or irrelevant code or errors in the code, so we can ask questions, ask them to go do something, and then that can be evaluated online.

So the question is, are all these technological enhancements improving our measurement, are we getting better information about what students know and can do, is the construct relevant? All these options may or may not be improving our measurement. I'll give you one example in the world of survey design. Oftentimes we have these rating scales that go from 'strongly disagree' to 'strongly agree'. On a computer you can have a little bar that people move, they slide from one side to the next, and they can place themselves on this really infinite continuum from 'strongly disagree' to 'strongly agree', and we can actually measure how far they are from one point to the next in number of pixels, which some people might think we're going to get a really precise estimate of where people are on the agreement continuum. Turns out that people cannot replicate where they are, which means that their score, their placement on that continuum is not reliable, because people are inconsistent over time. People can't make such fine distinctions in a reliable way.

Connie: Right, because you're trying to quantify feelings.

Michael: Yeah. And in a way that is beyond what people are able to report. They can't report their feeling on an agree/disagree scale at the level of the pixels. So it actually results in lower reliability.

Connie: That's interesting. Is the field of learning analytics becoming bigger in your field, in your research, where people are using the data to make big predictions?

Michael: We are. We are seeing a very interesting combination of cognition and learning sciences and large-- we call it big data, big data and data mining, and trying to bring together multiple sources of information in order to better inform our practice and better inform educational programming and policy.

Connie: It is an exciting time to be in this field.

Michael: There is a lot going on.

Connie: Yeah. Thanks so much, it was really nice to meet you and to speak with you.

Michael: Yeah, thanks for the invitation, and good luck with all your efforts.

Okay, I hoped you enjoyed this episode on testing. I've been wanting to get a psychometrician on this podcast for a few years, so I was just thrilled to come across Michael. I think he is a true teacher in that he explains things very clearly. Anyway, if you're interested in the show notes, you can find them at theelearningcoach.com/podcasts/39. Have a great few weeks, and I will talk to you next time. Take care.