ANDREA CHEVILLE: What we call item response theory has become very hot recently, and there are some compelling reasons for that. Historically, we would give a fixed length test, the way we all used to take when we were in high school and college, meaning that everybody who sits down gets exactly the same questions. And classical test theory suggests that if you don't administer those exact questions in those order-- in that order, that it's no longer valid.

And item response theory, in contrast, says, well, gee, each item actually tells me something. If I ask you, can you walk a mile without getting short of breath, and you say, sure, I've learned quite a bit about you. And it eliminates the need to ask a whole bunch of questions like, could you step up on a curve, that now aren't as informative.

So item response theory takes it-- it requires what we call a unidimensional trait, something that is-- and that could be pain, could be a symptom, it could be an aptitude like mathematics, verbal aptitude. So it's a unidimensional trait, and it takes items and orders them along that. We call it calibration, with people who have more of the trait being able to respond accurately or affirmative to certain items, and people having less of the trait responding in the negative.

So if it's functionality, can you jog for five minutes on uneven ground, and if someone says, sure, I can do that, they have quite a bit of that trait. They're high functioning, whereas if we say, gee, can you turn yourself over in bed, and they say, no, they're much lower. So now we have these items all along this continuum.

And where CAT, or Computer Adaptive Testing, comes in is we can create what we call computer algorithms by which the computer selects the next most informative item. So for example, if I say, well, can you walk a mile without stopping to rest, and you answer in the affirmative, sure, I can without any problem. Then the computer is going to select something harder, maybe like jogging for five minutes, or going up a steep staircase rapidly, running as if to catch a bus, harder, harder. And if, again, you respond in the affirmative, no, I can do that no problem, it's going to pick something still harder. If you say no to that-- after your initial question, gives you a harder one, you say, oh no, I can't do that, it's going to pick an item in the middle so that it is serially honing in on your level.

It is clustering the items that are administered right around your performance level, whereas anything that's fixed length we cover the whole range, and we do it pretty poorly. Most fixed length instruments are most-- can discriminate most precisely around the average. So we're pretty good at assessing the average patient, we stink when it comes to higher performance levels and lower performance levels. And so with administering more items we have less precise estimates and the hope with IRT and CAT is that we can get around that. And that's absolutely-- efficiency is critical if we are going to embed these measures in clinical workflows. Nobody has extra time.

For quite a while now, there has been an increasing mandate-- and this is coming from many stakeholders, including the federal government-- to amplify the patient's voice, and the caregivers voice, in their experience, in their decisions about care. And one way in which we are doing that is through the more broader integration-- the more broad integration of what we call patient reported outcomes, and those typically query patients about a subjective internal state, quality of life, their anxiety, their level of depression, symptoms, shortness of breath, fatigue, pain. But they can also ask them for summative judgments about things like their functioning.

And some of the nice attributes about patient reported outcomes, they're very inexpensive, for the most part, to administer, and they don't require point of care. So if I'm going to use a performance measure like a six minute walk time-- we're timed up and go, the patient has to be in a facility. That requires travel time. It's often not patient centric. They'd prefer not to pay the parking fee. Whereas a PRO can be administered via smartphone, over the internet, over a telephone, via mail survey, and all of these are robustly validated modes of administration.

So it's very patient centric, and it's because of that it's relatively facile to assess patients at multiple times after a procedure or after an important clinical event, which becomes much more costly and time consuming if we're relying solely on what we would call clinician rated measures-- we use a lot of those in rehabilitation medicine-- or performance based measures. And interestingly, some work out of Boston University suggests that that, in certain contexts, certain clinical populations, using a PRO is superior, gives you superior discrimination, but better psychometric data than using performance measures or clinician rated measures.

Quality of life is the holy grail, increasingly. We want to deliver value to our patients, and that's conceptualized in many different ways, but an important one is enhancing their quality of life. And function is one of the, if not the main among the main determinants of patients' quality of life. And so being able to assess patients functionality over time is our North Star, it's our clinical target in rehabilitation medicine. And patient reported outcomes allow us a means of inexpensively and patient centrically longitudinally assessing the parameters that we're most interested in; their mobility, their applied cognition; their ability to execute what we call daily activities.

So this was a secondary data analysis. I received a career-- a federal career Development Award to look at the performance of one of the very few CATs-- actually, to the best of my knowledge, there are only three Computer Adaptive Tests for functional items that are widely available and being used. And I'm very interested in the care of chronically ill, particularly cancer patients, and one of the challenges we face is sending them to rehabilitation at the right time. If you send them too early, they wonder, why am I here? If you send them too late, you've missed your window of opportunity to preserve their functionality, and too often, when rehabilitation services are involved, it's late in their care when they've already had significant functional decline and been living in potential a needlessly disabled state for quite a while.

So we're trying to find that sweet spot when patients need rehab and can appreciate their need for rehab, and are able to participate and gain maximum benefit. Which is no small feat when, for the most part, the medical specialists that are caring for chronically ill patients are focused on the disease. They have a lot on their plate. Assessing patients functionally is yet one more thing, and often, it's not done.

So we became interested in Computer Adaptive Testing as a way of serially monitoring quickly, efficiently monitoring these patients over time, but nobody had looked to see how sensitive are these measures? How do they perform? How responsive are they in detecting subtle levels of disablement?

So that's what we looked at, 311 patients with late stage lung cancer. And our original publication, which was the main point of the study, we found that the CAT performed exceedingly well. It was very discriminating. It was sensitive to very subtle, small decrements in our patients functional capabilities. We really honed in on their mobility.

Having established that, it opened up the door to conduct some secondary data analyzes and ask other questions, and some of those results were reported in the paper published in the American Journal of PM&R. And we were looking at behaviors. One of the neat things that you can do with CAT that you cannot do, for the most, part with pen and paper testing, you can look at test takers behavior. How long did they take for a specific item, did they skip a specific item, did they change their answer, because all of these things are electronically recorded.

And so we started looking-- first we are interested in symptom intensity, and does symptom intensity change test takers behavior. And the answer was, yes it does. Fatigued patients, particularly older fatigued patients, dramatically reduce the amount of time they're willing to spend on a question than when they're less symptomatic.

And that was actually women more than men were willing to take-- we didn't see the difference with symptoms as much with women, but they were much more willing to spend time on a question. And we assumed that's what we want. We want them to think carefully. We asked them about their ability to perform a specific activity. We want them to consider carefully when they've last done that activity, and really make an informed summative judgment when they answer the question.

So we did find gender differences. We found differences by symptom intensity, with symptomatic patients spending far less time. They spend less time answering the questions, and they skipped fewer questions, meaning that-- suggesting, at least my inference was, that they want to get this over with.

And so the other thing we found was that the consistency of their responses-- and this is an opportunity afforded by IRT, because we've now calibrated all those items. We-- and we do that with the responses of thousands and thousands of patients, and based on their responses, we model and we create this calibration along that unidimensional continuum. So if a patient is giving inconsistent responses that just don't make sense, it becomes much harder to hone in and to estimate their ability. And so if you're only giving 10 questions-- the CAT also estimates what we call a standard error, which is how certain can-- how confident can you be in this estimate. And so after 10 questions, highly symptomatic patients still had large standard errors, meaning their responses just didn't make-- weren't as consistent as the non-symptomatic patients.

And what our group took away from this is that really two things. One, perhaps the reliability, the faith, and the credence we place in the information we're gleaning from PROs may be less when patients are symptomatic. They have less cognitive ability to invest in the process of answering. But also, when patients are giving answers that just don't quite make sense-- as reflected in a large standard error, the uncertainty-- then it may suggest that clinical attention to other parameters, ie their symptom burden, their cognition may be indicated. so that the information we're gaining by looking at test taking behavior is multi-dimensional and may have important and actionable clinical implications.

As a palliative care and rehabilitation physician, I'm most intrigued by the symptomatic piece, because symptoms are so prevalent. Intense symptoms are prevalent among chronically ill populations. And if we're going to leverage the potential of PROs, we really need to understand that interaction of symptoms with the cognitive processes that are required to generate the answers we hope patients will generate that are accurate and good summative judgments of their functional capabilities.

When patients are suffering, when they're experiencing intense symptoms, they-- they're less willing to invest their cognitive energy, or they simply may have less to invest in the process of answering the questions. And so the utility of PROs may be lessened in highly symptomatic patients. Again, to the best of my knowledge, no one to date has looked at this, and we only looked at lung cancer patients. We only looked at lung cancer patients with fairly advanced disease, and whether these are generalizable findings, very much remains to be seen.

The second paper was published-- actually, it was published first in the archives of physical medicine rehabilitation, and it too capitalized on information generated by IRT and CAT that would not be available to us with the same degree with pen and paper administration. And what we were looking at, we used two data sets, actually, in this case, and looked at discordance between clinician and patients rating of their functionality. We know that in-- and this has been demonstrated across very diverse clinical populations-- that in about a third of cases, patients ratings disagree with their professional caregivers relatively pretty significantly. And what we found is when discordance was present, the patients were much more likely to endorse higher levels of pain. And which-- and when they were experiencing pain, their responses were less consistent. This was with the IRT.

So that, in part, it may be that-- and this is speculative, because, again, these are relatively preliminary and novel analyzes. It may suggest that intensely symptomatic patients have trouble with PROs, making the summative judgments, and so perhaps, in some of these cases, the clinician is more accurate. We didn't compare either the clinician or the patient's ratings to a gold standard, so we're not sure who was right, if anybody was right. It also may suggest that when patients report on their ability to perform a function-- an activity, they are considering their internal state, and the likelihood that the intensity of their symptoms will increase, which, obviously, the clinician does not have access to that information, because it's internal and subjective.

It's interesting that we've also found that, when there is a clinician patient discordance, there's more likely to be additional testing, there's more likely to be an appropriately prescribed treatment. So it's a very important clinical construct, and more work up, more costly medical work up, may not be the answer. It may be we need to delve a little bit more into the cognitive and symptomatic state of that individual, because those are the things we should be honing in on.

We're just starting. We were awarded what's called an R1 grant from the National Institutes of Health to pursue work very related to this. And along-- well, my interest is really the functional preservation of patients with chronic disease. And one of the great challenges they face is, when they come in the hospital, they're usually sick, but in addition to that, they remain immobile. So they might have a pneumonia, but a majority of patients, actually, particularly elderly, come into the hospital, lose significant functionality during their hospitalizations that cannot be explained by their medical conditions.

And this is a robust finding across institutions, countries. You can't single out anybody. When patients come in, they spend the overwhelming majority of their time in bed, and for ill patients, particularly elderly patients, they lose muscle bulk very rapidly. It doesn't take too much to preserve it, but in the absence of some exertion, they lose it.

And so what we're trying to do is leverage IRT and CAT to develop a very rapid screen that would be administered to patients on initial hospitalization that would stratify them into ability levels, and these ability levels would link directly to mobility or functional care plans that would be very simple. It might be doing leg kicks at the side of the bed. For someone who doesn't need to be supervised, marching, just weight bearing exercises.

But the goal would be to capitalize on the precision, the efficiency of IRT and CAT to integrate such an assessment in clinical workflows and start an individualized care, one tailored to their specific ability level as soon as they hit the ground, in the hospital. And to continually administer it maybe every two days, so that as their functionality changed, our interventions, our function, would be appropriate and keep pace with them. You know, after patients come out of an ICU, their functional status is often quite different, and we need to match that. So that's where our research is going now.

We have hundreds of measures, functional measures, but historically, very few have been integrated into clinical practice. And I believe part of the problem is we have measurement experts who are incredibly skilled in the development, the validation, and the vetting of these measures, and then we have clinicians, on the other hand, who are very busy, have other needs, and have not-- well, now they're being pushed, but in the past have not really seen much point in integrating these measures, and they haven't been. So we have lots of measures, their clinical application limited.

And where I see a tremendous opportunity for IRT and CAT, given its efficiency, given its precision, it really can be tailored to help clinicians in areas they need, without disrupting or clogging up their workflows. So I think it could be, really, a very revolutionary way of integrating the patient's voice into their care, and caregivers, because they play an increasingly important role.