NATIONAL QUALITY FORUM

Moderator: Scientific Methods Panel
July 12, 2018
3:00 p.m. ET

OPERATOR: This is Conference #: 7697485.

Miranda Kuwahara: Good afternoon. And welcome to the July Scientific Methods Panel Monthly Call. My name is Miranda Kuwahara with the National Quality Forum. And I'm joined by my colleague Karen Johnson, Andrew Lyzenga, Poonam Bal, and Ashley Wilbon.

Ashley is new to our Methods Panel team, but she has been with NQF for several years. She is a senior director. And many of our Methods Panel members may have worked with her on the readmissions and the cost resources measure portfolios. We're really excited to have Ashley on our team. So welcome.

We'll begin today's call, with the brief roll call and a review of the meeting objective. We know that we have Dave Cella on the line with us. Thanks for joining, Dave. Is David Nerenz on with us?

Dave Cella: Thank you.

Miranda Kuwahara: How about Matt Austin?

Matt Austin: Here.

Miranda Kuwahara: Great. Bijan Borah?

Bijan Borah: Yes.

Miranda Kuwahara: John Bott? Lacy Fabian?

Lacy Fabian: Here.

Miranda Kuwahara: Thank you. Marybeth Farquhar?

Marybeth Farquhar: Here.

Miranda Kuwahara: Jeffrey Geppert? Paul Gerrard? Larry Glance?

Larry Glance: Here.

Miranda Kuwahara: Stephen Horner?

Stephen Horner: Here.

Miranda Kuwahara: Karen Joynt Maddox? Sherrie Kaplan?

Sherrie Kaplan: Here.

Miranda Kuwahara: Joseph Kunisch?

Joseph Kunisch: Here.

Miranda Kuwahara: Paul Kurlansky? Zhenqiu Lin?

Zhenqiu Lin: Yes.

Miranda Kuwahara: Jack Needleman? Eugene Nuccio?

Eugene Nuccio: Here.

Miranda Kuwahara: Jennifer Perloff?

Jennifer Perloff: Here.

Miranda Kuwahara: Sam Simon? Michael Stoto?

Michael Stoto: I'm here.

Miranda Kuwahara: Christie Teigland?

Christie Teigland: Yes, hello.

Miranda Kuwahara: Ronald Walters?

Ronald Walters: Here.

Miranda Kuwahara: And Susan White? For those of you that are on the webcast but perhaps not dialed in yet, just send us a note via the chat functions as you work to make your way on the phone.

On today's call, we'll review a few process updates that came out of the May in-person meeting. And then we'll dive into a methodologic discussion on reliability for insurance based measures. And with that, I will turn it over to Poonam.

Poonam Bal: OK. Thank you. So first we're just going to start from some updates on the discussions we had in the in-person meeting and the feedback that you provided to us about how to improve the process. So in regards to such a process, our reflection was that you suggested an option 1.5, something between the current independent evaluations and a full panel evaluation of all measures.

And the main take away, we heard from the meeting was that. Learning from each other is a key ask. Consensus from a larger group would be better. However, it would be fairly impossible to evaluate up to 50 measures in a two- to three day in-person meeting. It might work if we discuss only those measures where there is a disagreement. But there are cons to doing it that way.

Regardless, agreed we need earlier resolution. Waiting weeks afterwards is not working. And then also more information about final ratings, (place) agreement, basically creating that feedback loop.

In terms of evaluation form, your recommendation was to move to a more free-text option but still have some checkboxes, similar with that 1.5 option. So, next slide.

So after discussion internally, we came up with – at this point, two options. On this slide, we said most likely moving to subgroups. But after some more internal discussion, that may not be the best solutions. So we're hoping to get a little feedback from you on what you think would be the most feasible while still conducive to the benefit of the project.

So the thought would be for subgroup is that a subset of the panel members would evaluate a subset of measures, meaning kind of like now but a little different in the sense, instead of having three independent viewers who submit their reviews and then staff do their analysis and may be we jump on the resolution call. It would be that there would be a bigger group maybe five to seven people assigned to what we would call a workgroup. And then, that group would review all the measures assigned to them in that workgroup. And then come together for pre-planned meeting to discuss all of those measures. We would – so that would be the idea, we're doing a subgroup versus doing a full panel review.

Either way, it would require a series of webinars. Initially, we had thought the max would be 50 measures. And then, right now with the trends of those measures that were being submitted, we're roughly looking at 20-25 measures to be submitted. So it wouldn't really be possible to review that on a one webinar and so on.

And so, the concept would be though that we would do almost a consent calendar method where we would have measures, any kind of measures. The panelist weren't able to agree on the reading. That discussion would mandatory. But if the measures either fully pass or fully did not pass, those measures will only be discussed if pulled for discussion. They would be pre-slated for a recommendation. And then, only if people felt that they should be further discussed would they be discussed. That will be the same if we do a subgroup or a full panel review. And then, there would be discussions

assigned for the measures and either way there would be people assigned for in depth reviews and so on.

The ratings would be based on the subgroup vote. If we did workgroups and would be based on the full panel vote, if we did full panel. Either way, there would be no co-chair adjudication at the end. It would just be whether has been decided at that meeting would be the decision. And there would be formal vote instead of just ratings and then combining this together to see what the rating should be.

The webinars would be open to the public but developers would not be able to introduce the measures or provide feedback or provide answer any question. It would just be more for them to listen and learn. Discussion would really have to be between the panel members. We would still do an evaluation form in advance of the meeting. That would be how we would determine the consent calendar in where the measure should rest.

So there still will be a pre-meeting evaluation form that you would be filling out. But we would make it similar to what we discuss in in-person of a medium free-text with a little bit of bullet points. So that would be still required. And then the output would be a simple majority vote. Plus a staff generated summary. And then the evaluations that you do would still be provided to the standing committee. And then that's the public like they are now. Those evaluations will still be shared. So that will be changing. But now we will add a summary of if the measure was discussed in the meeting and what was the result and so on.

So majority wise, the subgroup and the full group is – would be very similar. The differences would be obviously in the full group. We would have more reviewers and potentially stronger consensus. Since everyone would be reviewing all measures, it would be – consistency would be more likely amongst the reviews. However, in order to do full panel review, panel members would have to attend more (meetings). So for – if we did subgroup such workgroup concept, each panel member would be attending one to two max meetings within a timeframe of two to three weeks in October.

If we go with the full panel method, then you could potentially be attending 5 to 10 review meetings in that two- to three-week October timeframe. So the list is much higher in regards to your time on a phone call.

Also, this could make it more difficult to keep quorum. So for voting purposes, we do require 66 percent of quorum, meaning whoever is supposed to be on the call, 66 percent of them are on the call in order to make a decision. And so, there is potential that that would be more difficult to achieve when the full panel is required to be on 5 to 10 meetings versus if they're required to be put on one or two meetings within that timeframe.

And then additionally, panel members will have to review more measures with the full group concept. So, you would be assigned the same number of measures to do a deep dive. If we do subgroups or full groups, you'll still have to do a thorough analysis and fill out the evaluation form for the same number of measures. However, if we have the subgroup, you will only review the ones that you did that deep dive for. If we do full group, you'll have to at least have an understanding of all the measures in order to have an enriched conversation on a full group webinar.

So those are the major points and kind of the differences between the two options. And then before we jump to the next topic, we want to take a moment to see what the panel's views were, what are you more leaning towards, would you prefer – do you think that asking for 5 to 10 webinars all two hours having to review up to between 25 to 50 measures at least on a – enough to understand the basis.

Do you think that that workload is achievable and something that you're willing to do in order to gain a more consistent and perhaps thoughtful discussions by having a full panel review all measured versus a subgroup reviewing a subgroup of measures. So with that, I wanted to open up to see if there was thoughts.

Bijan Borah: Poonam, this is Bijan. With regards to subgroups, do you – what's your view in terms of the number of measures that would be assigned to each of the subgroups?

Poonam Bal: Yes. So we are thinking about a range of about 5 to 15 really depending on how many measures we get. If we get a lower number, you're more likely to get five measures. If we get a lot of measures up to that 15 mark, it could be as much as 15 in the subgroup format.

Eugene Nuccio: This is Eugene. Several of us have been measure developers. How would we recuse ourselves when you group us in panels?

Poonam Bal: So we would – if we're to break you down to subgroups, we would make an active effort to not put you on a subgroup that has any measures that you're conflicted with. So that would also be a benefit of the subgroup one since everyone won't be assigned to everyone. We can make sure that you're put into a different one.

However, if you're only conflicted with say one measure and use of expertise for the other measures that are on the same topic area, you'd probably still be assigned to that workgroup. But then not be able to do evaluation or speak to the one measure you're conflicted with. So it's really case by case.

Eugene Nuccio: Thank you.

Christie Teigland: This is Christie. I have a question about why the opening up the webinars to the public. I mean under the current process, we didn't even necessarily know what our other people who were reviewing the same measure said or thought much less share it with the rest of the panel. So why are we now switching to a public forum giving visibility into these arguments? I think it might have some impact, but.

Poonam Bal: Yes. So, the thought process behind that is that we are a transparent organization and we want the public to have the opportunity to hear the discussions if we are having free-scheduled meetings. And I think before, the idea was that since they're individual reviews and eventually whatever is put into those reviews is put in – given straight to the standing committee, it's still transparent and open.

It would be very difficult to obviously make e-mails or those kinds of interactions available to the public. But since we are doing a webinar where individuals could join, then we should do our best to make it transparent.

Joseph Kunisch: Hi, this is Joe. Just a quick question. What do you think on the subgroup would be the turnaround time expected if you're talking between to 5 to 15 …

Poonam Bal: So we would keep the same structure right now where you have four weeks to do your evaluation.

Ronald Walters: This is Ron. So I have one of those dangerous things that always goes through my mind. Do you anticipate there's going to be two rounds of measure submission during the next year?

Poonam Bal: Yes. So every year we'll have two cycles so there's always be two rounds of measure review.

Ronald Walters: Yes. So here's my dangerous thought. First of all, you have to decide what's your outcome is. And obviously the outcome is the quality of the review. That's hard to measure, can be measured and so on and so on degree of consensus or something like that.

The second outcome I think you referred to is speed. And so, part of this is also geared towards getting these done more efficiently. This is starting to sound like I'm a quality measure for most of medicine.

We want to do it efficient. And we want it to be of high quality. There is nothing to say. For example, and here's the dangerous thought, without too much extra headache, we could do the first group by one methodology and the second group by the other methodology. And at the end of year, we'll all would be a lot smarter about which one worked better for everyone. That's just a thought.

Matt Austin: This is Matt Austin. I guess I would find it difficult in say a two- or three-week period to be on 8 to 10 webinars lasting for two hours each, just in terms of other commitments that I have. I would slightly vote or in favor of the smaller subgroups.

Jennifer Perloff: This is Jen. I agree.

Larry Glance: This is Larry Glance. I agree that if the two options are as presented that we should go with the subgroups. The other option that we had discussed in our in-person meeting would be to have the entire group meet, and in this case, it sounds like it would be twice a year in person in D.C. and spend two days doing this.

And I think that there are two advantages to this. One, is I think when we are all sitting together in one room, we learn an awful lot from each other. And the second thing is that it would allow each measure to be evaluated by essentially every measure – every person on the Methods Panel. And I think that there's some value to that.

One of the key things is that it would make over time for much, much more consistent evaluations. The other thing is by concentrating this into two in-person meetings. We sort of, you have like a Saturday and Sunday, you kind of you do it and then you're done as opposed to having multiple phone calls over several weeks.

Poonam Bal: Thanks for that, Larry. And we definitely did consider that during our own review to see what would be the most feasible. Unfortunately, it would be actually closer to three in-persons because we still have our in-person to talk about more conceptual concept.

So, considering funding and feasibility from getting everybody in three times a year, in-person meetings didn't really seem feasible so we strategized, well, what if we did in-person cycle. But in the end, there would be have to – even if we went with that route at some point, we would have to do a series of webinars at least for one of the cycles in order to balance out all the in-person meetings.

Larry Glance: Karen, what about combining the method of evaluation with the conceptual stuff? In other words, spending 80 percent of the time talking about the measures and maybe 20 percent of the time talking about more conceptual

issues, in that way that would wiggle it down to two meetings a year, which might be maybe a little bit more cost-effective for NQF. What are your thoughts about that?

Poonam Bal: Yes. So we could potentially do that. But it would really depend on how many measures were being reviewed, so if we had fewer measures that would definitely be feasible to squeeze in – reviewing the measures in more conceptual concepts. But if we have high number of measures to review then that would basically take out any time for a conceptual review. It may even result in us not being able to review everything during the in-person as well.

Sherrie Kaplan: Yes. This is Sherrie. I apologize for missing the in-person meeting. We had an emergency. But I'm – we're focused on workload and timing, and I'm still a little bit confused on what actually NQF wants out of the Methods Panel review. And here's why? I was on the Patient and Family Experience Committee yesterday. There's confusion about how to use the input from the Methods Panel and their deliberations.

And so, can somebody from NQF actually say what they actually want. What's the product as a Methods Panel because it can – (other) issues blah, blah, blah, that's different meeting, if it's this actual nitty-gritty review of each of these measures, then the – how this is going to get used and how you're shaping it for use in standing committee, it seems to me the first question to answer.

Poonam Bal: Yes. So, really the goal of the Methods Panel is twofold. So we do have the higher level conceptual goal. We want to use this panel to get to some of those more conceptual ideas and making sure that we're reviewing things the way that we should and that are providing guidance to the developers and the creator community on how that's received. So that's one aspect.

And that's really what the goal of the monthly calls we have with you is suppose to be. And then on the other side, in terms of what give to the standing committee, the goal was that the Methods Panel would take the weight away from the standing committees of having to review the scientific susceptibility for difficult measures or complex measures.

So it was supposed to be that methodologist who are experts on how to review these measures review them and provide guidance and feedback. The standing committee who can then choose to either take your recommendation and say, the standing committee reviewed it. And we think that they did a great job. We don't even talk about this anymore. We have anything else to add. And just accept your recommendations.

And on the other end, they could take it and say, OK, we agree with what they said. But here's some conceptual issues that we've found based on the topic area that we're dealing with. That may not have come up in this review. And I think that's a good example for where the standing committee would might want to talk about measures where risk adjustment is involved.

The Methods Panel is only (to sort of) review for the actual method. (If the) methodology correct, what and how they came to the result, if that makes sense. Not review if the right things were used and if you agree with the decision that was made. So that would be more of a standing committee decision of they did not use of right elements or knowing the field, this is not the right decision to look forward with.

And so there is that place with the Standing Committee and a place for this Scientific Acceptability Methods Panel. So that's really what we're aiming to do. Does that help clarify?

Sherrie Kaplan: Well, no. Here's why. There was – we disagreed in the Methods Panel and the feedback we gave. And there were – it went to some input by the chair, I believe, or the co-chairs. And there were still a lot of confusion. So it actually caused more discussion and confusion than it helped.

And so I'm still kind of interested in exactly, are we – is the goal of the scientific review panel to methods review panel to reach consensus, to give feedback to the standing committee? In which case, we need to thresh out what the issues are and have lengthy discussions for the measures level. Or if we're providing both guidance in the conceptual – at the conceptual level and application at the individual measures level. I think those are two very

different tasks. And they – how you approach them and how much time they take is going to be conditioned by what you want out of the process.

And maybe it will get better. But I suspect that we'll still have differences unless NQF make a specific point of we must reach consensus on how to give recommendations back to standing committees or not, I don't think I yet feel like I have enough guidance to sort of vote on this structure.

Poonam Bal: Yes. So that's what we're really hoping to solve with this new process. We agree that maybe it wasn't so clear when there was a disagreement. There wasn't always an opportunity for the groups to talk amongst themselves and then the co-chairs would provide their feedback. And then that was all given to the standing committee without any sort of summary or description of, what was the final decision/recommendation.

So the goal of this new structure is that we would not lose individual review where people can provide their own thoughts and disagree because that's always – everyone comes with a different point of view. But then come into the webinars and have a discussion and come to consensus.

And when we say consensus, we don't mean 100 percent everybody agrees that this is the takeaway. But we will be using strong a majority. So that's a 6 – greater than 60 percent of the committee has agreed – I'm sorry, the panel has agreed that this measure is good enough to pass and here are some things that they brought up as concerns.

And then the standing committee hopefully will then have – and then we'll be writing a summary of the discussion or writing a summary of what was said in the pre-meeting evaluation survey – form. So, the goal is that with the summary and having open dialogue and having everyone speak to each other about the measures that that confusion will go down. And then there'll be more of a clear statement of majority of the panels would say that this measure is methodology sound.

Karen Johnson: And this is Karen. Just to kind of underline a couple of things, we will ask whether we do this subgroup idea or whether we do full panel. There will be a formal vote on measures.

To Larry's point, to that the in-person, we would love to do in-person. We are funded by CMS to do this work. And in the past, they've really been pushing against funding as many in-persons as we used to be able to do. So, we just are not sure that that would be looked upon favorably even though we agree and we would prefer the in-person.

Larry Glance: Yes.

Karen Johnson: And then finally, the 8 to 10 webinars, probably worst case scenario, but we have to tell you that worst case scenario. And that would happen if, number one, we did get that closer to 50 measures per cycle that we had originally thought. Again, the last few cycles that didn't happen we ended up getting closer to 25. So that's one thing.

The other thing is, you would probably be doing – Sherrie, that threshing out on the measures that you don't – there wasn't kind of pretty much agreement on. And so the 8 to 10 calls would have give you room to do that threshing out.

The more measures that there's consensus on by the small group, you may not need to call them out and there might not be anything to thresh. So we might be able to drop a couple of calls. Again, as we get, as time goes on and we get better at kind of knowing what we need and knowing what we're looking for, that sort of thing, hopefully that number that needs threshing out will shrink.

Dave Cella: So this is Dave Cella. And Karen, thank you very much for saying what you just said. You said a few important things there. And I was going to comment on them. But we're still waiting for you to say that because I thought it was more appropriate for you to say them.

And one of them was the issue of the funding. I think we did agree that it would be good to have face-to-face meetings. And I don't think anybody on the call disagrees with that, that the funding is not there for it.

The other – and this may be a way to address Sherrie's question. Unfortunately, Sherrie, I think we have to do both tasks. We have to make

decisions or make judgments on submissions and serve a review function, at the same time we have to figure out how to do it well and come up with standards for that. And that latter half of the task is where the white papers, the white papers come in. And I think many of us are eager to see that process start.

The other thing I would say in this was an analogy that David Nerenz came up with. And I know David is on the call but I'm not sure he's unmuted. But I'll say something and you can maybe amplify Dave.

But we're thinking that, as Karen mentioned, yes, every decision is going to be a decision of the committee, not a subgroup. But in any particular submission, just like what NIH reviews, for those who are familiar with NIH review process and practice. There are going to some that are obviously not possible. And there could be a way to triage them. With full support of the committee but probably was very brief minimal discussion and maybe not even any discussion. But the opportunity to call those – call anyone forward for discussion if anyone on the committee thought it was important to have a conversation about it, but most of them would be triaged and not discussed. And then there will be also a few that are unanimously by the subgroup approved to go forward and would likely not take much discussion.

And so the need of a discussion would be on those controversial ones in the middle that might represent the whole less than half the submissions, whatever the number of that is. And that would be the vehicle by which we would try to align our standards and definitions. So that is a maybe summary of what we've been discussing since our last meeting recognizing that we have – and Sherrie pointed out two very different jobs to do at the same time.

Poonam Bal: So, we are at the – past the halfway mark. And we do want to talk about a couple more things. It isn't really seem – I know there are some comments about the benefits of full panel and there are some benefits of a subgroup. I think at this point, if you could just give us an e-mail with what you feel would be the best route to go either subgroup or full panel, we can go from there and give you a final answer before we start the new review process.

So August 1st is the intent to submit deadline. And at that point, we'll have a greater idea of how many measures will be going to Methods Panel. And we'll start to start our review process and give you dates and such like that. So, if everyone could just give us an e-mail, it can just simply be the word subgroup or full, just so we know which way before our leaning. And we can finalize that and give you notification by e-mail about which route we'll be moving forward with.

And if you have any additional questions that weren't answered during this call, also give – you can send us e-mail with those questions and we can respond to them the best we can.

So with that, I'll just jump to what Dave just mentioned, the toolkit/white paper progress to date basically, what we need to do and want to do.

So creating a methods toolkit, again, the toolkit is one thing we're using but the name up in the air. It's beneficial to (all). So staff, panel members, public, developers, that will outline definitions, methodologies, thresholds, best practices, really the idea behind it is that we will come to consensus on what's appropriate and demonstrate that to all of our stakeholders and the general public so that this can become the standard.

And so it will be easier to come to consensus on measures and hopefully result in less disagreement and maybe eventually no need to discuss measures because there's so much agreement on consensus, and not – and having those. So, we're definitely working towards creating that and finalizing that.

Along with that, we are trying to submit peer-reviewed journal for key methodological information based on Methods Panel discussions. The goal is that those journals will produce information that becomes a subset of the toolkit but in different format. So obviously, the way that we structure things may not be appropriate for peer-reviewed journal but taking the major points from that and using it to build our toolkit is the goal.

But while also publicizing this work and getting it out there and gaining traction. So, you received an e-mail a couple of weeks ago about three papers

that have already been conceptualized will over the overview of the Methods Panel work. The second paper will be on clinical outcomes. And the third one will be on patient reported outcomes.

Thank you for everyone who has put in – who has volunteered to help with the writing process for those journals. And then given their preferences, we'll reach out to you soon, letting you know which one we'll be asking for your help for which paper. Also, if you have not e-mailed yet and you would like to be part of one of the writing groups, please e-mail us by the end of this week so we can get all of that finalized and get the writing groups off the ground.

We are going to have different leads identified for each paper. So we'll have those leads also reach out to you as we get closer. And I think that's really it. Were there any questions about the toolkit or white paper before we jump to our methodological discussion?

Sherrie Kaplan: This is Sherrie. I volunteered for the patient reported outcomes thing. But I don't know if anybody ever got them. We had trouble with our – we converted to a new e-mail system. And so I was worried it got lost. So, just so you know, I did volunteer for the patient reported outcomes paper.

Poonam Bal: OK. We'll double check the list to make sure you're on it. Thank you.

David Nerenz: Poonam, thanks.

Poonam Bal: I'm sorry, what was that?

David Nerenz: Dave Nerenz, just saying thanks.

Poonam Bal: Thank you. Anything else? OK. Well, with that, I'll pass it over to Karen to start our methodological discussion. Karen?

Karen Johnson: Thank you, Poonam. And I realized that our discussion today on process took a little bit longer than we had originally hoped. But thank you for thinking about that for us.

Again, when you send us to your e-mail on your preference, really what we're asking you to do is differentiate and tell us which one you really think would work best for you and for the panel. So, if you would say for workgroup, that would mean that basically every measure would have kind of the stamp of the full Methods Panel. And the potential difficulty is there may be a lot more calls and measures that you'd have to look at.

You would – no matter which option you would do, whether full or the subgroup option, you would still be asked to do the deep dive on the same number of measures. So, I think we've beat that one to death.

Let's talk a little bit about our methodological issue. And I don't think we'll solve it today. But maybe we will. And the – and I'm really glad Sherrie is on the call because I think this was something that Sherrie kind of floated to us early on and to be very transparent. I didn't really get what you were saying, Sherrie. But it did come up and we talked about it quite a bit in the in-person meeting. And I think we have a handle on where your discomfort.

So, first of all, our expectations to date for reliability, so this is reliability and this is instrument-based measures only, what we use since thus far is that we need reliability to be demonstrated for the instrument itself. We call that the data element level and for a performance measure score.

And when we say data element, we were actually talking about the items or the questions on instruments. Score level again, focusing on distinguishing differences, where you're looking at data that have been aggregated up to the provider level.

Just a real quick note there, we are still having – we're still not real sure about this distinguishing differences thing and that's going to be the subject of maybe not the next call because we probably won't get through this today, but very soon.

And also, as part of the writing of that second paper, I think we're really going to thresh this thing out, so more to come on that. But if you go to the next slide, thinking about instrument-based measures only and thinking about those out where performance measures are based on multi-item scales only. So,

we're limiting our discussion today to multi-item scales that feed into performance measure.

Our question is, and Sherrie I hope I have it right. Is there some kind of an in-between analysis that we should be expecting, another level of testing if you will? And that would show the reliability of the patient scores that result when combining multiple items.

So, again, before, we were saying, well, show us after you've aggregated up to the provider level that would be the score level. And we have said, data element, that's really all about the questions of the instruments so things like Cronbach's alpha or factor analysis, things like that where you would show that sets of questions you'd hanged together for lock of a better way of saying it.

So, that's what we've been focused on up until now. Is there something in-between where we need to actually look at the scores and how the data look once those have been rolled up and look at the actual patient level data?

So, the question is do we need this other kind of level of analysis, if so what is that, if not Cronbach's alpha? But what kind of kinds of things would that entail? If we're interested in that, should we still be interested in knowing about the actual items and questions and making sure they hanged together?

And, again, if so, then, what language do we use to describe? What I have on the slide here may or may not be kind of the right way to put this forward. So, I'm going to stop there. I want to hand it off to Dave Cella who – Dave, you may need to reinterpret what I just said. Hopefully, I got the concepts down. But Dave will help us walk through maybe let's try at least 10 minutes today.

If we don't solve it in the next 10 minutes which I suspect we will not, we will keep going with it on the next call.

Dave Cella: OK. That was great, Karen. I think we should go right to discussions since it's just 10 minutes. This is – we really want to make sure that we are at least clear on our charts. And you're talking specifically about multi-item scales.

So, this is something, what else would we want to see with regard to anther level testing for a multi-item scale?

David Nerenz: David Nerenz here, if I can just say things just a little bit. Karen, if you mentioned that this came up originally from a cause or concern raised by Sherrie. I wonder if we could turn to Sherrie just that this framing of the issue actually capture your concern with or the (cause) you brought up, or is it not quite there yet?

Sherrie Kaplan: I think it does, but can I reframe it a little bit?

Dave Cella: Yes. Go ahead.

Sherrie Kaplan: Because what I was concerned about and it came up in the context of when you use something developed and you never – only scores in my world only scores are ever evaluated for psychometric properties, you can't say something as ever reliable and valid for all purposes for all times.

But, at the patient level, if questions consistently measure the concept – the construct that you're trying to measure, that's OK. That's great. But if now you're using that patient level evidence of a reliable and valid construct to estimate one level up, say, the providers performance, then the error that you get at the patient level becomes part of the error at the physician level.

And so, at the patient level …

David Nerenz: Right.

Sherrie Kaplan: … something like Cronbach's alpha. But now at the physician level, you've got between provider variations. And then over between provider plus within provider, so there's not a strong thumbprint of provider behavior within their patient population, then the denominator is going to be big and that (interclass) correlation or the error term around it that is going to be large. And therefore, it probably doesn't really do the job at that level.

So that's kind of what I was concerned about. I don't know if that's a between in-between analysis. But it certainly, and I think I've sent along some splines

of what would constitute a really not physician, good physician – (Tim Hoefer's) paper, a not good physician level measure where there's a lot of variation within physicians.

And then another paper that actually we wrote that shows a fairly narrow error bars and fair amount of between physician variations. So that's kind of where I was headed. Does that helps?

David Nerenz: Yes. And Dave Nerenz back again. Thank you. To me that's perfectly clear. And I actually resonate with that thing as the issue. What I'm hearing though, there are only two levels of analysis we're talking about. We're not talking about an in-between or a third. We're just basically saying, the questions about reliability of the measure at that patient level. And then we're talking about reliability of the measure as a measured performance at a provider group or hospitals and other kind of level. But I don't think I heard anything near about an in-between or third level.

Sherrie Kaplan: I didn't see one.

David Nerenz: OK.

Karen Johnsons: So this is Karen, when I said, in-between, again, we have mostly, well, we've only ever kind of expected people to tell us about, I guess, the – Sherrie, you may have talked about the reliability or construct that the Cronbach's alpha kinds of thing tell us. So, but it does sound like you want something in addition to that. And that's why I was calling it in-between. But is that correct or am I still missing it?

Sherrie Kaplan: No, no. If now you're using that measure that you used to – that you tested at the patient level for another purpose, you're using that estimate provider performance, you have to test it for reliability at that level. And that's what – that's the second level of analysis that I personally would like to see if everyone's going to be use to estimate my performance, I'd like to make sure that it's reliable at the level that you use it.

Michael Stoto: So this Mike Stoto. I think maybe the confusion is what's the data element level? I think some of us may be are talking about the data element level is.

One element in a – and what gets put into a scale or others may think the data element is the scale itself. Is that where the issue is coming up?

Karen Johnsons: Yes. So, this is Karen again. So, up until now we have said data element just talk about the actual questions, right? So, make sure if you have three items that make up your construct, make sure that those three items are working and there's not a fourth that you don't need and all that stuff, right. So, that's the kind of analysis that we've been expecting and looking for.

We also do expect score level testing. That's when you roll it up to the provider level and you do your in-between versus within. So we also expect that. But I think, Sherrie is saying that there is a little bit more that we need. So, again, I'm having trouble kind of verbalizing what that is. So, I'm trying to get a better understanding of how to verbalize it and what that is exactly.

David Nerenz: Got it, Dave Nerenz. If I can jump in? Again, I think we're using some words in different (phrase) and …

Michael Stoto: Yes.

David Nerenz: I didn't hear Sherrie say anything about the data element level referring to a specific item. So at least in my background in this field as well, (you) understand the reliability of individual items in a month item scale. If you do a test like Cronbach's alpha or any other reliability test, you're basically establishing the reliability of the scale.

Yes. Cronbach's alpha, you look at how the data elements perform relative to each other. But Cronbach's alpha is fundamentally a property of the scale, not of each individual item. Now, so that means (you're really standing) of two levels to talk about. You've got the level of toolkits and reliability scale to measure some property of a person.

And then with Sherrie question which I agree with fully is, the second level is the reliability of the score of – a provider score that's inevitably made up of multiple of these (inaudible). So somebody might solve (the new) terminology, multiple individual patient scores. But I also think the (phrase) in-between is very (nice) here because that implies there's one thing and

another thing and then a third thing in between. I think there are only two levels at least as I'm thinking to about it.

Sherrie Kaplan: Yes. Let me – can I jump and try one analogy? So, if I were trying to estimate all of your math ability with three questions, I wouldn't – first of all, I wouldn't be able to do it because it's a complex construct. And that'd be crazy. I'm trying to ask, which sub-dimension do I leave out? Do I leave out arithmetic or calculus or geometry or whatever?

And then the – so for me, the individual questions are, you would not do that because you're trying to estimate a complex construct with a bunch of different things that we think reflect that multidimensional construct. So, I don't want to get lost in jargon. But to me, that kind of the individual item level for these kinds of measures isn't really where you want to look.

Now, you can do Spearman-Brown prophecy formula and say, did I have enough of those things? So I have enough math questions to estimate this complex construct. Three is way too few. Maybe I need 50. But that, you can do kinds of analysis that get you there, but I didn't want to bring that into it because I think that might add too much complexity and people go – can go nuts trying to get methodological rigor mortis. You get death by psychometrics and you don't want to make people look completely crazy.

On the other hand, this is a kind of thing where if you're going to now use it at the next year, exactly what David said, you have to evaluate it for use of that tier. So that means you have to see how much error is a function of, did you sample too few patients? Did you sample enough patients to get a stable estimate of that doctor's behavior?

So, this is – I think there is no in-between is what I'm saying. I think there's just two levels. And I don't think the individual item level is meaningful.

David Nerenz: Agree.

Dave Cella: This is Dave Cella. I'm mindful that it's three minutes until closing the hour and I actually to have to be in another meeting at 4 o’clock.

You know, I think we're really – this is a (inaudible) questions (folks) and multi-item scales. But if that multi-item scale has a good internal consistency, it's almost by definition – at least their experience going to do a better job in a single item measure of the same thing.

So I don't know that you need more than internal consistency. And what we're not talking about here is stability, which I think from a reliability perspective, it's the more important component of reliability that is if something doesn't change overtime, will you get the same score. Because that's probably going to contribute as much or more to the ability to differentiate providers and practices in a real way or based upon what's really different.

So I think the writing of the toolkit or the white papers and the wordsmithing around that is going to help this process and discussion, because I don't hear a lot of disagreement, just a struggle to find the common acceptable terminology.

Karen, I wonder if you want to make sure we get anything else in, in the next two minutes before we have to clearly – I have to go, I don't know about others.

Karen Johnson: Yes, thanks, Dave. I'm sorry, I was on mute. And so thank you for that.

I think what we'll do is we'll just kind of keep working on this. Keep working on our language. And I know Larry, you're probably very interested in this because you're heading up the writing of that second paper.

We'll decide whether we need to bring this one back next month or if we'll have a different discussion next month. So, Sherrie, I may tap you to make sure I really understand and we'll go from there.

And our next, well, we have about a minute. I think I want to handed to over to Miranda to tell us about our next steps and I think that's all we need to do.

Miranda Kuwahara: Sure, so before doing so, we'll take this opportunity to hear from any members of the public who would like to offer comments. If you're not

connected via phone, you can submit your comments via the chat function in the lower left hand corner of the screen.

Operator, could you please provide instructions to the participants.

Operator: Yes, ma'am. At this time if you would like to make a comment, please press star then the number one.

Miranda Kuwahara: And while people are queuing up for comments, I will review our next steps.

We'll reconvene next month on August 9th for our next monthly call. But in the interim as a reminder, we ask that as your mini homework assignment, please submit and e-mail to us with your preference indicating whether you'd like to move forward with a subgroup review or a full panel review. Just a really quick note, I know we already have some members who have submitted their preferences. So, thank you to those.

Operator: And there are no further comments at this time.

Poonam Bal: Perfect. And then also if you would like to be part of the white paper group and have not already submitted your preferences, please do so by the end of the week so we can finalize those groups and get the work started.

Miranda Kuwahara: And there are no comments via the chat. So, with that, we can adjourn today's meeting. Thank you all for your time and engagement this afternoon.

END