National Quality Forum

Moderator: Scientific Methods Panel
October 16, 2018
2:00 p.m. ET

OPERATOR: This is Conference # 2898198.

(Miranda): Thank you for joining the follow-up call for the Subgroup Number Four Panel Meeting.

Karen Johnson: Let's stop and let (inaudible) take over. (Miranda) is – yes. (Miranda) is not doing well today.

Female: Sorry about that. So, thank you for joining the second meeting for Subgroup Four. We won't go through all the details that we went through on the first call since we just had that call. And, so, we don't think you need to really be reminded. But, we just need to continue with two measures.

We have already finished the review for 3452 and 3461. So, we won't be going back to those discussions unless something hugely different has occurred to you that you feel might change the result – only if it changes the result. The ones that we're going to be reviewing today are 3227 CollaboRATE Shared Decision Making Score and 3476 Communication Climate Assessment Toolkit.

As a reminder, we did discuss the other five measures in this grouping, and we did decide to pass them without further discussion. However, as Karen mentioned on our call yesterday, if something has come up since then based on your review that you feel has changed your mindset and potentially might change the result of these measures, you do still have the opportunity to pull

them for discussion. So, I will give you once last chance to pull them if you see fit.

All right. I think that's enough of a pause. So, I think we're good on those measures and we're going to focus on 3227. It will be the same structure where we will instruct you on when to vote. And as a reminder, the SurveyMonkey is a generic SurveyMonkey for all the measures. So, even though you may have questions in there that are related to something we are not talking about, we ask that you only vote on the items that we specifically ask you to vote on.

I think the next two will be interesting because, technically, we don't need to vote on them unless you think your vote would change based on the discussion we have. We did – both of these went down based on your votes. However, we felt that it'd be important to talk through these measures. So, you may not vote at all this round. But, it will all depend on how the discussion goes.

And then, again, with the SurveyMonkey, you do have to open it up each time not for each individual vote but for each individual measure. All right. I think that's all of the logistics. Karen, is there anything that I may have missed that you want to bring up?

Karen Johnson: I don't think so. I'll just make sure that I know who is on the phone. (David), can you …

(David): Yes. I'm here.

Karen Johnson: All right. Sam?

Sam Simon: Yes. Here.

Karen Johnson: John?

John Bott: Yes. Here.

Karen Johnson: Thank you. Zhenqiu?

Zhenqiu Lin: Yes.

Karen Johnson: And Joe?

Joseph Kunisch: Here.

Karen Johnson: Great. And I don't – we don't think Paul Gerrard is on the line. Paul, are you there? OK.

Sam Simon: Karen, this is Sam. And I apologize.

Karen Johnson: Yes.

Sam Simon: I am going to only be able to be on the first hour of this call.

Karen Johnson: OK. All right. We'll see how it goes. This might go really fast. If yesterday was any indication, probably not. But, we'll see how it goes. Before we delve into 3227, just a quick question for you, guys. Do you have anything you want to ask in terms of process or what we did yesterday, or are you ready to just jump into 3227?

(David): (Great. If I can), Karen – this is (Dave). Just – since in both cases – I'm just following on what we heard a minute ago. But, we may not even have to vote these. And on the original run-through, they didn't pass. What exactly, then, do you want us to talk about this time? It sounds just like in terms of where (the various forks in the road are) that these two are different from the ones we talked about yesterday.

Karen Johnson: Yes. On the first one, 3227, there is a little bit that I want to share with you that may make you want to reconsider your vote. So, I just wanted to air those things. And that's actually going to be true on both reliability and validity. So, that's the main thing that we want to talk about on the CollaboRATE measure.

(David): OK.

Karen Johnson: On the CCAT measure, the Communication Climate measure, it's not so much that we want to get into the nitty-gritty of the measure. There's just a

couple of kind of overarching things that I think would be useful to talk about just kind of in general for directionality of the (methods to panel) and that sort of thing. So, that's why those I doubt very seriously that you would want to revote. But, it could be that we'll bring up something that would make you want to change your mind.

(David): OK.

Karen Johnson: And we'll see how this goes. I mean, we had come up with this idea of we would definitely – where votes were split, we definitely wanted to discuss. And kind of our original thinking going through was if votes tilted one way or the other way, "Pass" versus "No pass," we wouldn't have to discuss those.

But, in every subgroup call, we as staff have actually pulled measures even if they have passed, for example, but just we want to point out something in terms of criteria or point out something maybe that is an overall question. So, this isn't uncommon with your group. So, we'll see how that goes. Any other questions?

OK. Let's go ahead and delve into the CollaboRATE measure. So, I'm going to do kind of the same thing that I did yesterday, maybe not quite spend as much time in the – in the introduction; but if we need to, we can. This is – it's a really interesting measure.

It's a patient-reported measure or shared decision making. And it's based on the CollaboRATE survey. And it only includes three items – so, not a big survey. We know that they have set up a minimum sample size of 25. The exclusions on this measure were a little unclear. The way the submission was set up, they say there were no exclusions. But, the calculation algorithm says that they would exclude cases where the response is one or more of those three items were missing.

Now this is something that could certainly be – we want exclusions to be clear and that sort of thing. But, it is something that we could verify with the developer and have them submit materials, if you decided that you wanted to change your vote.

And we would just make that crystal clear in the submission. But, that's fairly – a fairly minor issue. The data source, again, is a survey. There's multiple modes of data collection – we've listed those there – and multiple languages (that it's) available.

And one of the things that was a little bit unclear or I wasn't completely sure about when I was looking at the submission materials is what is actually the level of analysis. So, I did contact the developer and asked them a couple of questions. And I did indeed verify that they are putting forward (the physician) groups – so, not the individual clinician level. So, only the group is what they are asking for our endorsement.

In terms of risk adjustment, that also is unclear because it was – there were some inconsistencies in the submission. So, I asked them about that so that we would have that clear for today's discussion. And the calculation algorithm actually says don't do any risk adjustment. But, later on, they talked about risk adjustment.

And, in fact, they are adjusting for mode, for age and for the interaction term of mode by age. So, those are the three factors that are included in the risk adjustment model. They actually examined the gender but, in the end, did not include gender in their final model. So, that is a clarification that we didn't have otherwise.

Now, in terms of rating through reliability, they were across the board. But, they did tilt towards the "Not pass" with one "Low," three "Insufficient" and one "Moderate." And, again, we pulled that mainly because we wanted to talk a little bit about this level of analysis and what was there.

The other reason that we wanted to pull this is because this one and the next one are such complicated measures, we wanted to make sure that, if indeed these measures don't pass, that we can be very, very clear to the developers of what the – the concerns were so that we can offer them very specific feedback. And I think, for the most part, you guys did that in the preliminary analysis forms. But, there might be something else or some other way that you say something that would make it more clear or so.

So, because this is an instrument-based measure, we note that both data element and score-level reliability are required. And, in fact, they did some internal consistency and intra-rater reliability analysis. They did some ranking analysis and a signal-to-noise analysis. So, they did do some data element and some score level. So, they meet those minimum requirements in terms of what they did.

I want to stop there for a minute. I would like to come back to validity and hit one issue kind of towards the end. But, let's go ahead and talk about reliability first. So, if you turn to page 11 if you have your discussion script open – the items to be discussed.

So, basically, in terms of what they said they did for data element testing, they said they did internal consistency and reported some Cronbach's alpha statistics there for inpatient versus outpatient. And they also talked about doing an inter-relater – sorry – an intra-rater reliability analysis. And you see the results there.

Now, the – there were two concerns with the second set of analysis, the intra-rater reliability – first, whether or not that Cohen's kappa was an appropriate test for intra-rater reliability and, then, just kind of a more general concern that the testing – since the measure looks at just the top rating, the feeling that that testing really didn't fully test the data elements.

Now, one of the things that I wanted to point out on this one – and we can talk about – we can certainly talk about, you know, what (methods) would have been appropriate if they were looking (to) intra-rater reliability. But, as I understand (the submission), this intra-rater reliability was actually looking at the clinician individual level.

OK. Now, I want to make sure that everybody would agree with that, first of all. But, if that's the case, they are not really putting forward – asking you to endorse at the individual level. So, you could actually just ignore the analysis and base part of your rating simply on the internal consistency piece of it, OK, again, assuming that I am correct and that the intra-rater reliability is actually looking at the clinician individually as opposed to groups.

In terms of Cronbach's alpha, I think in general we feel like that may not be the greatest test for this kind of internal consistency. I think people have stated before in other forums that looking more towards risk factor analysis type of thing is where we'd like to push the fields that we have …

Male: (Inaudible)?

Male: I don't know. (Inaudible) last bit of the (inaudible) (hope I can get only) …

Karen Johnson: OK. And somebody didn't mute your phone there. So, let me just finish with the internal consistency. We have allowed alpha – Cronbach's alpha types of analysis to go through. So, let me stop there and see if you guys want to talk any about the data element testing.

So, basically, what – let me rephrase what I am (pausing) on. It seems like the intra-rater reliability was done at the individual clinician level. If that's really the case, then you could ignore that because they are not asking us to endorse at the individual clinician level. That said, (we can talk about) …

(David): Karen, this is …

Karen Johnson: Yes?

(David): This is (David). Let me just focus on that, although it may not be the big issue. When I'm looking at what they say about intra-rater reliability, what I'm seeing is that they calculate it using these hypothetical scenarios or vignettes. It would seem to me that that doesn't really define any predicted level of analysis. Basically, what they've done is said, "Let's present people with a description of scenarios that vary widely in terms of their shared decision making, let the people rank them."

And, then, this intra-rater reliability, which also I think could be called test-retest, is they take a bunch of people who have done this once and, then, they let time go by and then they show them and have them rate them again and they find reasonable level of agreement. But, I'm not sure if I'm reading this correctly. But, that implies any particular level of analysis – it's sort of artificial vignette data no matter what.

Karen Johnson: OK.

(David): Now, did I miss something there?

Karen Johnson: Not necessarily. I am – I am going to leave that to you. I'm going to be looking for those documents. I should have them up and I don't have them up. But, you may be right. I may have misread thinking that it was some kind of a clinician level and, in fact, it may not be.

Zhenqiu Lin: So, this is Zhenqiu. I have the same reaction in the (MITA form). I think this is really (during) the early stage of development, you can use (similarly). Ideally, you want to see some – to use (extra) patient encounter and to (inaudible) what other tool is reliable. But, this is (not using simulator encounters).

Karen Johnson: OK. So, I guess – so, I think maybe I – my notes were just wrong here and this isn't really a clinician level. So, apologies. I kind of led you down the wrong – the wrong thing there. There was concern that the intra-rater reliability that they have presented not only is, like you said, based on these scenarios but it may not even be the right method.

I think what I'm trying to say is, you could just say, "Hey, we (don't have) (inaudible) (method)" or – it's simulated data kind of thing. I'm just pointing out that you could just ignore that and base part of your rating only on the internal consistency analysis. Does that make sense?

Male: Yes.

Male: (Yes. Although) …

Male: Yes. Just – as we just heard it, I think this is a reasonable thing to do when you're in the early stages of developing a survey. You know, you want to see if you're going to pick up the difference if it's there. So, you make up some plausible vignettes.

Now, this kind of thing has been done in other scenarios and that I think it's part of the package. I wouldn't stop there. But, in this case, I don't think they

did stop there. So, I – this, by itself, wouldn't sell it to me. But, I don't think it's bad.

Karen Johnson: OK. Does anybody want to talk at all about the Cohen's kappa not being an appropriate methodology for doing this kind of test? Is that still a concern?

Male: (Inaudible).

Joseph Kunisch: This is Joe. I …

Male: Go ahead.

Joseph Kunisch: I was just going to say I'm – I wasn't so much concerned about that. But, just the design of it didn't really make sense. And, maybe, again, somebody might have more expertise on these survey-type instruments.

But, that – the reason our – that the fact they had it on the 1-to-10 scale but they were only testing the positive (of the) responses, you know, basically – they scale one through nine, you'd make the assumption that there is no difference between those scores. You're only looking for the highest score and to test the intra-rater reliability against just that. And that didn't really make sense to me why they chose to do that.

Male: Well, I …

Joseph Kunisch: (Inaudible).

Male: For what it's worth, I read that it's what – I guess is often called this top box method. You see this fairly frequently in analysis of patient satisfaction surveys or other similar kind of surveys where the distribution skews to the high end of the scale.

And for better or worse, owners and developers and users of measures say, "Well, let's calculate the metrics and group based on how many of the respondents choose the highest possible score," which basically means the highest score is a "Pass" and everything else is a "Fail" and then you sort of go from there.

Now, we might feel that that's not the best way to do it. But, I guess once they make that decision (to revise that) (inaudible) (to do this), OK, once you make that decision, let's see how it plays out there from there and do you get reasonable reliability and validity once you've chosen to do it that way. But, it's not weird, I guess, sort of in a bigger domain of survey measures. You know, we see it (in other places).

Male: Yes. I agree. This is not uncommon.

Karen Johnson: OK. Anybody want to talk about it being an inappropriate method – using of the Cohen's kappa – or is that something that something that you are willing to live with?

Sam Simon: Well, so – this is Sam. I did raise that. I don't think it's the biggest issue here. My understanding of Cohen's kappa is that you use it for inter-rater reliability. They were using it in an intra-rater – so, the same rater. And my understanding is Cohen's kappa is used for two raters. But, I don't know that that's the biggest – the biggest issue here to raise.

Karen Johnson: Right.

Sam Simon: And it doesn't sound like that would – if they changed that, it would – it would change anyone's mind about this reliability testing.

Karen Johnson: OK.

(David): Yes. Karen, (Dave) here. If I can just suggest – looking at the distribution of scores, I'm the outlier. I'm the one who said it was "Moderate." Three of my colleagues said "Insufficient" and I am inclined to think on that basis (I'm finally) the one who is wrong. I'd be interested in hearing why it was deemed "Insufficient" because I – again, I am probably the one who (read or) misinterpreted something here.

Karen Johnson: Yes. And I'm going to let you guys talk about it here in a second. My guess – and I'll hand it over to all of you, guys. My guess is the biggest concern with the testing was the values from the signal-to-noise analysis. And

thinking about the score-level testing that they did, they did two different things. One was the signal to noise.

And, then, they also did that clinician-level ranking. And that's the one that I should have been saying. You could just ignore that clinician-level ranking. They are not putting it forward as clinician level. So, that was the concern, that it was at the wrong level of analysis. So, we could just ignore that and focus our discussion on the signal to noise.

And in this case, it could very well be that you guys will not want to revote. I think where we might need a little help is maybe a little bit more direction to the developer on what they – did they miss a step in her somewhere in the analysis? Or what went wrong with that analysis, in your mind?

Zhenqiu Lin: So, (inaudible) (ranking), I don't think this is – I think we are not looking at the clinician level. But, the analysis is not about (reliability) at all. Right? That's – (what they are trying) (inaudible) whether you see a different survey mode, you will get a different clinician ranking. The analysis they – there is no significant difference (on this) (inaudible) (from reliability).

And that (inaudible) (out provided) group profiling (is understood between variant) 0.0028. And, then, they found the reliability average is 0.7. I mean, they (inaudible) (25) (inaudible). I just can't – I have trouble understanding how you can get – is it based on (a hierarchy model between) (inaudible) (of variant, you try to) apply their formula. It's very difficult to get that high reliability. (You probably take thousands of respondents).

Karen Johnson: So, just FYI, I did contact the developer. I found out a little bit more about the number of groups. That was one of the questions that you had. They did have 153 clinician groups. And they told us a little bit more. And you can see that in the bullet where NQF verified with the developer. The sample sizes in the groups range from 31 to 1133 with an average of 204 per group and the majority had fewer than 300 responses. So, it probably doesn't make you feel any better, Zhenqiu.

Zhenqiu Lin: If that's the case, it's making me feel not any better. Yes.

(David): Now – (David). I mean, that would seem to be a really crucial point. I just – I'm not that good with the (intuitive) math on this. But, if one statistic they present renders another statistic implausible to impossible, then part of the response is that we just don't see this matching up, that we don't think you can get this kind of reliability given what you've said about the 0.0028. That would be a reasonable response.

Karen Johnson: So, what do we say to the developer? They suggested – it seems like that they used (Adams) beta-binomial approach, which maybe they did or maybe they just used the overall equations. I'm not sure.

Zhenqiu Lin: So, if you (look at the form), they list the equation. It's using a (inaudible) beta-binomial approach. But, then – but, between (inaudible) (there getting from the hierarchy model) (inaudible). So, I think when you use (inaudible) measure, (right), it's not (research adjusted). And, so, I'm not sure how (they apply that formula).

Karen Johnson: So, if they were to ditch the beta-binomial – let's just hypothetically say ditch the beta-binomial but do a signal-to-noise analysis, would it just be the hierarchical and they would take this random effect variance and plug that into a formula somewhere?

Zhenqiu Lin: Yes. If you are using – (inaudible) (also. When you kick at ATC, it would be) sigma squared over sigma squared plus (pie squared) divided by three. Right? And, then, (inaudible) you can (inaudible) (you can derive reliability) and you can found out (fundamentally most of the hierarchy model vote).

And that's why – but, here, you see they have – when they had – I think they are using the (John Adams) beta-binomial approach. It's that – I mean because their model is (the largest regression model) – right? So, I would expect to see the (pie squared over three instead of key times one minus) (inaudible).

Karen Johnson: OK. (Pie squared) over – OK. All right. I think that was the gist of what I wanted to get from you, guys. It doesn't sound like you are interested in re-voting this. Am I correct? This still seems like we have kind of two different statistics that don't mesh.

Male: Yes.

Male: Yes.

Male: Yes.

Karen Johnson: OK.

Male: Yes. I would endorse that.

Karen Johnson: OK. All right. So, what we're going to do is we're going to continue. We will put this down. It's not passing. So, we will make sure that we try to explain what we think the problem is and maybe a potential solution of another analysis (to run).

Something we won't talk about today but something that maybe we will come up on the monthly call – this clinician-level ranking. I found it a little interesting, and I wondered if that could be considered almost analogous to the split test – the split-half test that we do. So, anyway, put that on the back of your mind. We will – we will come back to that at some other time.

Sam Simon: Hey, Karen.

Karen Johnson: Yes?

Sam Simon: This is Sam.

Karen Johnson: Yes?

Sam Simon: This may be out of scope for this discussion – so, that's fine. I struggled with this measure in the sense that it really struck me as a – and I understand why we weren't evaluating it as a composite. I understand that it uses patient-reported data.

But, it really didn't seem to meet the spirit of combining two or more component measures, each of which more individually reflects quality of care

in sort of a single performance metric. That I'd be curious to understand sort of what – why NQF wouldn't consider this a composite.

Karen Johnson: A quick question for you, Sam. Are you getting it – this one confused with the Communication – the CCAT measure that used the patient and the staff?

Sam Simon: Well, I think – I actually thought both could be composite measures. But, I am – I am still thinking about this CollaboRATE tool since it's looking at the – the CollaboRATE tool has three questions and they're – and we're looking at the – what was it – the average percentage of people how had the top score on all three items. So, we're combining three items into one.

Karen Johnson: Yes. And is it are they combining three at the patient level or are they rolling – are they rolling each one up to the clinician group level and then combining? In my mind, that's the difference?

(David): (David). Let me try one other thing that – certainly in any (multi-head) of scale, you have some combination of multi-items to produce a score. So, simply doing – using multi-items, at least in my mind, doesn't make it a composite.

It seems to me the characteristic of a composite is that you, first of all, establish two or more measures as legitimate measures in themselves and they have to have certain properties and, then, you put them together. Now, am I off-base on what we mean by the term composite? And that's the essence that I've been thinking about. And, then, in this case, it didn't occur to me that that's what they are trying to do.

Karen Johnson: Yes. (David), the scenario that you just described – that's one of the types of composites that NQF recognizes. We also recognize and call for all-or-none measures. We treat them as composites as well. So, those act a little differently. They are not the individual scores rolled up at the level of analysis and then combined. They are done at the patient level and then combined at the end. But, that's kind of the exception. So, those are the two that NQF recognizes. And (inaudible).

(David): OK. But, in that …

Male: Yes. And …

(David): In that case – just to press the point, in the all-or-none scenario, we are talking about two or more measures, right, not two or more survey items?

Karen Johnson: In the all-or-none, it's definitely not two or more survey items. A good example of an all-or-none is we have a measure actually that another subgroup was looking at. It's called Optimal Diabetes Care. And, basically, it looks at five different things. And it's, obviously limited to diabetes patients.

So, they say, "Does Mary have blood glucose under control and high blood pressure control and taking an aspirin and not smoking?" And there is a fifth one that I forgot. But, basically, they look at, does the individual patient have all of (ever) how many components. And if they do, then they get a check mark. And if they don't, they, you know, get an X. And, then, after they do that for all the patient, they roll it up and that's the composite measure. So, that's the all-or-none.

(David): Yes. But, the key thing – just to clarify the point, the things being combined are themselves measures in their own right. That's the key thing. Right?

Karen Johnson: They are not performance measures in their own right. So, just saying that Mary has her blood glucose under control is – would not be a performance measure. You would have to take Mary and Joe and whomever – all the patients at a clinician group, for example, and roll that up and then maybe somehow another take the averages. In that particular measure, they don't do that. They are just looking at the individual patients.

And so, basically, there are – there are looking at a patient – did the patient get this, this, this, this and this? If so, then the numerator is a yes or a one. Right? And they do that for all the patients and then roll it up. So, it is a special case that NQF considers a composite. And that is a different animal than the scenario that you described where each individual measure – each individual component is a standalone performance measure.

(David): OK. Well, if that's not the case, then I will back off and say now I don't understand what a composite is either. I thought I did.

Karen Johnson: Yes. We will – we will come back to that and make sure that you understand. We are very clear, though, that having multiple items that make up this domain of interest – we do not consider those composite measures at NQF, even though the field calls that a composite, right? Multi-item (scales) can be rolled up into a – some kind of a construct. We do not call those a composite.

(David): Right. And I wouldn't either.

Karen Johnson: Yes. But, I think that was Sam's question. Right, Sam? This is three different questions. And to get back to answer your question, it would depend on whether or not they are taking these three items and kind of rolling those up together and then aggregating, in which case we would not call this a composite.

If they were taking the three individual items, rolling up each one – so, you would have a performance measure about understanding your health issues, a separate performance measure about things that matter and, then, a separate one about what matters most – and they roll that up at the clinician group level and them combines those three rolled-up values, then we would call it a composite.

Sam Simon: OK. That's …

Karen Johnson: Does that make (any sense)?

Sam Simon: I'll admit to being – I'll admit to being a little confused about it. I did see this as kind of meeting a description of an all-or-none. I do recognize that each of these items don't sort of stand on their own as a performance measure. But, I don't want to take up more of the group's time. I will – I will – maybe we can chat about this offline.

Karen Johnson: OK. So, we'll definitely come back and devote a little time to talking about composite measures. And another thing I might do is just kind of share with you the report that we did a few years ago when we pulled a group of people

together to determine a composite measure. So, we will – we will come back to that and make sure everybody is on the same page there.

One quick question for you, though, before we get off this measure. And I don't want to beat this one to death, especially since, Sam, you have to leave in a couple of minutes. I do want to draw your attention very quickly to the validity testing that they did. Now, you guys did pass this measure on validity, and it was passed as a "Moderate."

My only question for you has to do with the concurrent validity that they did. They compared the results from this measure to two other CAHPS measures. And to be honest with you, I was confused because of the way they reported the results. Since this is a group measure, I would – I would have expected two correlations, you know, once correlation between their measures and CAHPS measure one and a second correlation with their measure versus CAHPS measure two.

But, they talked about in 92 percent of the measured groups, correlations were greater than 60. So, I was just confused. It almost sounded like they did correlations at the clinician level, not at the group level. I don't know if any of you have any insight on that. Am I just misunderstanding what they reported? (Inaudible).

Male: I'm trying to catch up. I'm trying to read their text. Is it under the heading "Concurrent Validity"? Is that where we are?

Karen Johnson: Yes. We'll see. Let me go …

Male: OK. I'm just trying to read quickly.

Zhenqiu Lin: And I – actually, the concern on the (sensitivity) section is just right above the concurrent validity. So, they say (in the online simulation service) study, they only captured 39 of the clinical scenarios where all three dimensions (of SDIM) were present. So that, to me, is a little bit troubling.

Karen Johnson: And, (David), if you have the – this – further down. This was actually some fairly new data that they just received from California, their medical group

survey. So, this is kind of hot off the press data. Again, it's just the way that they phrased it.

Male: Yes. I see the issue. You know, when they talk about their methods, it seems pretty clear that what they are correlating the group score is particularly in the (VA) context. But, then you go down to the results. This may have this thing about weakest correlation, medical group, level (inaudible) one of 153 medical groups are – equals 0.58.

So, if it's just within one medical group, then I guess – I don't know. Either we're talking – (patient) is the level of analysis or clinician. Yes, it's a little messy. I mean I was – (I love) to think that somewhere across all this, they had at least some plausible evidence. But, it's not as clear as it could be to say that what they are doing is looking at the validity at the measure score level. It could be stronger.

Karen Johnson: OK. So, I think – again, I don't want to spend too much more time on this. I think another one of the feedback that we'll provide them is to really make sure that they are correlating the two medical groups and presenting that data. And, of course, we will give them the preliminary analysis that you provided that also gets to some of these other things that you had noticed.

OK. We've got 20 minutes before Sam has to take off. So, we'll see how far we can get – are you guys OK with stopping on 3227 and going to the next one?

Male: Yes.

Male: Yes.

Karen Johnson: OK. So, 3467, the Communication Climate Assessment Toolkit. This is one that, again, did not pass. And I just wanted to bring out a couple of things to you – some things you mentioned. So, this is a patient-reported structure measure. So, (David), you mentioned that in an email, and we actually agree with you. So, we would really call this an instrument-based measure of a structure. And, so, we will work with them to change their measure type there.

So, Sam's point about should this be considered a composite, you know what, to be honest with you, it never even occurred to me but, yes, I think you are right. It should be considered a composite because they are rolling up the patient side of things and they are rolling up the staff side of things and then – so, those are two individual measures across the nine domains – and they are combining them. So, Sam, I agree with you there that we should ask them to present this as a composite measure as well.

Sam Simon: Yes.

Karen Johnson: Yes. So, I don't know why I hadn't caught that before. But, I agree with you. This one – the measure actually came through I think last fall. And I want to give the developers a lot of kudos because I think the last time they submitted it, the submission wasn't as clear as it could be.

And they have done – it's very obvious to me. For those of you who did see it the first time around, you won't really appreciate it. But, they did a phenomenal job of explaining things much better than they did the last time. So, it looks great.

And they were very clear which was – I think the major stumbling block last time around is they are submitting the – what they are calling the nine combined domain scores as measures for NQF endorsement. So, one of the things that I wanted to just make sure that everybody is clearly about for the next time around when we do this is this is one of these measures where it is one NQF ID – it's under 3476 – but they are actually putting forward nine individual performance measures.

Now, what that means is that it's perfectly OK if you had (leaned) this way. You could have said, "I'm willing to endorse five of these but not the other four." So, just so you know, that is absolutely an option. Now, right now, I think the answer is it all needs to go back and a little bit more work done. But, it could very well be that you'd be willing to endorse a subset of these nine coming back through. So, let me stop there and make sure everybody is clear on that.

Joseph Kunisch: This is Joe. It's clear now. But, it wasn't clear definitely when I was reading this. I was reading it all as one submission.

Karen Johnson: Yes. It's a little tricky. It's another one of those kind of vagaries of our numbering system. If I have my way, I'd just ask everybody to submit them all separately. But, developers really don't want to do that. They want to put all these things under one ID number. And there is – different people have different kind of opinions on whether they should or not.

But, I know in some of your comments, there were two or three especially that were of concern to some of you. And, so, when we bring them back – and, hopefully, some of you will be able to look at this one again – just remember that you have that option of maybe passing some but not all. So, that was one thing I just wanted to bring up.

The other thing that they suggested – and I didn't know if you would have some advice for the developer or not. And that was their score-level testing. So, on your discussion guide, on page 14, basically, this is an instrument-based measure or really nine instrument-based measures.

So, our requirements are they need to tell us, you know, about the reliability and validity of the instrument as well as the reliability and validity of the nine individual performance measures. And what they did for score-level reliability was just show the variation in the overall scores, and they gave some standard errors.

And, basically, they said they weren't able to calculate meaningful side-level measure score variances because it's coming from two different patient scale. So, my question for you is do you have any specific advice for these developers. Just showing the variances with the standard errors, it sounds like you guys aren't willing to take that as a reliability analysis that you would accept.

And that's fine. What would you suggest they do? I mean – or are they misunderstanding the signal to noise and what that would – can they actually do a signal to noise and maybe they just didn't know they could?

Male: Well, (inaudible) here. The – I think that's kind of my main problem. They clearly have done some work here. But, (they kind of find out) we can't do any statistics. And, then, my reply is, "Well, if you can't do any statistics, then I can't judge your reliability." You know, the fact that there is variation is sort of something. But, everybody else working on measures starts with that observation and then translates it into something like signal to noise or an (interclass) correlation or something.

Now, I don't feel like I'm sharp enough in this area to say, "You do should X specific test." But, they seem to claim they can't do anything because the (Kaplan) 2009 tutorial only talked about dichotomy. So, it comes down in the dichotomist – or dichotomist measure (inaudible) (dichotomist) measure. But, there hasn't been a barrier to other people.

There are plenty of ways of doing statistical tests for reliability with continuous measures. ICC might be one of them. I cannot – I don't feel like I can say, "Do one specifically." But, (I graded them well) because they didn't offer anything. I mean they understood the issue but then they said, "We can't do it."

Karen Johnson: OK. So, doing an ICC may be …

Male: It – it gets into this funny issue, you know, back to the composite. This is really a tricky one because they never claim to add all these nine things together. So, it's not a composite measure singular in that sense. They are nine separate measures.

But, where the composite might come in is that they have two very distinct data sources. They've got a patient survey and they've got a staff survey and they put those together. Now, I'll just defer the NQF definition whether that makes each one of these a composite or if it doesn't.

But, they essentially say that's the sticking point. I guess I don't understand why exactly that's a sticking point, that however they choose to combine the two data sources, it seems like then that that produces a score. And maybe their point is that, well, once you've done that, there is no variance at the site level with which to do any statistics.

And I guess maybe I'll grant that. But, I don't think that means, "OK. We say, 'Fine. You get a pass. Go ahead.'" (And we somehow – they still have to) try to address the mathematical challenges of showing reliability.

Zhenqiu Lin: I agree. I think they should be able to do something. For each facility – I mean they can come out with a score. Right? And (that's – well, that score) (inaudible) or – and, then, you can (final) the variance (among the facility). Right? You can (take it as a ratio). So, I mean there are many ways you can do it. So, I just – it's kind of puzzling why they said they couldn't do it.

Karen Johnson: OK. All right. So, our feedback is "You really can do it." So …

Zhenqiu Lin: But, that's separate from the issue in terms of how they combine, right, the – like – saying (the same) domain and then they combine patient response with the staff response and then (inaudible) the internal (consistent or other psychometric) properties. Even the same domain – the staff version is very different from patient version in terms of (psychometrical) properties. And, then – so, that calls into question whether it's reasonable to combine them.

Karen Johnson: And I think maybe that gets to Sam's point that if it's – conceptualizes the composite, then we can talk about their quality construct of kind of combining those two. And that might actually help a little bit there, I think. Right? They would have different psychometric properties of the patient side and the staff side and they could talk about those and then kind of separately talk about, "Hey, these are individual measure that we are combining into this composite measure of, for example, leadership."

Male: Yes. Although, again, just technically in semantic picking point, I am not sure they ever claim that the patient responses represent one standalone measure and the staff represent another. They're just two data flows into the measure. So, let's not worry about that.

I guess what I am curious about that I am not enough of a statistician to have the answer is if you have multiple patient responses for a given provider, you've got variance. And if you've got multiple staff provider – staff responses for a given provider, you have variance there.

Then, what they do is they add those two things together. And, to me, now they go into some real deep water about how you – how you add variances. And I – somebody must know this. But, I don't know it.

Male: Yes. And I thought that was (inaudible) your – the point we were making (rightly).

Male: Yes.

Karen Johnson: So, Sam, does that off the – kind of off-the-cuff bother you, that they are kind of adding (inaudible)?

Sam Simon: Well …

Karen Johnson: (Inaudible) from a – I guess from a composite standpoint?

Sam Simon: I – yes. I mean I definitely share that concern about that. Well, we just don't – we don't know a lot about it because they haven't provided a lot of the information we do want to see.

Karen Johnson: OK. I think that was – now, I wanted to mention one other thing, the construct validation. So, it actually did go through for the most part (inaudible) (split on) validity. But, I think one of the major concerns of their validity testing was their choice of measures that they did their construct validation against.

And I think – I feel like we talked about this yesterday. Just, at some point as a methods panel we'll have to decide are we – are we going to say, "Hey, they did something and the results were reasonable and we're going to let them through this time" or we start kind of drawing a line and saying, "Not only do you have to do something, but it needs to be a reasonable something; it can't be kind of the trivial solution or something that sort of doesn't matter" or maybe it's just we need more description about why you think this is a good measure to compare against.

So, I think that will be just a discussion point at one of our monthly calls that we can maybe as a full panel come to some kind of flavor so that some

measures aren't being held to a really strict interpretation of the construct validation and others kind of get through on, you know, somewhat trivial things that check the box that maybe don't really do much.

So, that was just, you know, things to come. I think that's all I had on these two measures. Did you guys have anything else that you wanted to bring out on these? Let me be clear. Nobody wants to re-vote on reliability or validity of this measure. Correct?

Male: Right.

Male: Correct.

Male: Yes.

Karen Johnson: OK.

Male: Correct.

Karen Johnson: OK. Do you – do you want to bring out anything else on these measures?

John Bott: This is John Bott. Just a question about how NQF views the need to risk adjust or not. I hope I got this right. My crib note says the measure steward found a relationship between the domain score and race, gender and language but they didn't factor it into the risk adjustment consciously because they didn't think it was appropriate to factor in demographic factors. So, is that – is that an OK position for them to have after they demonstrated that these variables do influence the score?

Karen Johnson: That's a tricky question, John. Yes and no. I mean they could certainly do the analysis. So, basically, the idea is that they've shown conceptually that they feel like there is a reason to even look at those. And, then, they have taken the next step and they have looked at it and they found that there actually seem to be some relationship there.

But, then, they kind of – and I don't remember if this is what they did. But, I – from what you said, I guess it is. They kind of decided anyway that they

didn't want to include that. That could certainly be their decision. And, then, the question is, is that something that is a reasonable thing to have concluded?

And we have directed you guys is methods panel to do is to point out any concerns that you might have with that sort of decision making but, then, not let that be the only factor that would take a measure down. So, in other words, if everything else looks great but you just didn't agree with their decision not to include something in the risk model, we don't want you to fail the measure because of that. And the reason is we want that to go on to the Standing Committee and let the Standing Committee have that conversation and talk about it from that perspective.

Male: Yes. I was sort of looking for an update on NQF's policy on factoring in demographic risk factors.

Karen Johnson: Yes.

Male: (Inaudible) (I know). I know it's kind of in the pipeline right now and I thought this was a shorthand way to get an update to the policy. But, it sounds like it's a bit gray. The other thing that I thought was odd about their statement regarding that – they deemed gender in that bucket of demographic factors.

But, as having been a measure steward and developer for a long time, you know, us in the community of measure development typically don't put gender in that bucket of demographic factors and then weigh whether we ethically should factor in gender into demographic factors. This is just carte blanche if it – if this is a risk factor, lo and behold, they factor it in.

But, they – which I have never seen before – they put it in that bucket of demographic factors and questioned whether – we don't want to – we don't want to risk adjust for that because, this is masking a variation based on demographic factors et cetera, et cetera. So, I didn't know if NQF had a position on which factors are in that discussion of demographic factors or not.

Karen Johnson: Yes. When we talk about social risk factors, we generally are not thinking about age or gender. So …

Male: Yes. That's what I thought.

Karen Johnson: Yes. And going back to your other question about, you know, our policy, as (David) mentioned I think on our last monthly call, we had the panel to help us think about social risk factors and adjusting or not adjusting. And at that time, NQF listed its prohibition against including social risk factors in the risk adjustment (as a policy).

And we ran that trial for two years. In a nutshell, the findings of that trial was that, number one, developers actually cooperated with us quite well in many cases in that they did try to start thinking about the conceptual rationale. Many of them actually were able to find some data and do some analysis.

For the most part, the measures that came through – the eventual decision was not to include social risk factors in the measurement. So, that was kind of like what we saw at the end of the first – what we call our first social risk trial. And, of course, one of the takeaways that we knew before we started the trial is getting the right data is hard and sometimes it's not possible.

What we have done is – since is we have begun yet another trial. So, basically, our prohibition against including social risk factors is no longer in play. So, again, basically what we are saying is, is there a conceptual rationale? If there is, use the data that you have that represents that conceptual thing that you're expecting. Do some analysis. If it looks like that a particular factor is associated with the outcome of interest, then serious consider including it in the risk adjustment model.

Where the – some of the controversy is now is some of the rational – some people just (uprightly) say, "Hey, we think it will create a unfair or two-tiered or whatever system. So, we just – (our priority aren't) going to include it anyway." Some people do that.

Other people do things like, "We included it and it didn't change the (C) statistics" or, "We included it but the ranges of results were very similar and people didn't change around ranks very much, so we think we're not going to include it."

And what we are trying to do I think especially this time around is ask for different types of analysis. So, in other words, you know is it surprising that a (C) statistic doesn't change very much, for example, if you just add one extra variable? Is that the right question to be asking? If that's not it, what should it be?

So, one of the things that we are suggesting that people do is look at kind of the – look at rankings with and without risk adjustment for providers who have the kind of extremes of the social grouping. So, if your grouping is poor versus not poor, you know, look at providers who – look at a lot of poor people and look at providers (who look at very few) and see how those rates change if you adjust (versus unadjust).

So, I think it's yet to be determined. But, there is some dissatisfaction and saying, "Hey, it seems to be working in the – in the risk adjustment model just in general that things are statistically significant." But, you know – but we're still not going to include it. So, that's kind of where we are.

(David): Karen, thanks. (Dave) here. If I could just – if we have a couple more minutes, sort of on behalf of the panel that did this work in 2014 – one of the key things that we've included in the report is this idea of trying to distinguish whether any effect or any empirical association you find with race, ethnicity, poverty, (inaudible) (whatever the factor is), is it or is it not mediated by quality of care?

Now, this is a little more easy to think through when you're talking about an outcome measure where, on the one hand, you have outcomes that are somewhat (distal looking in time and place) from the (care division) hospital readmission. And 30 days is one. Hemoglobin A1c control is another one.

It is quite plausible that the effects that you see related to a demographic factor have to do with things that occur out there in the big world as opposed to things that relates specifically to what providers do. So, there is a very plausible case there to do adjustment.

On the other side, if you're talking about process measures like administration of beta blocker in heart attack where everything that matters is directly under provider control and there is no clinical indication to treat people differently by demographics, then there is a very strong rationale that says you don't adjust even if you see an empirical association because you're probably adjusting (a way or quality effect).

It's tricky in a case like this. It's – (structural) measures are actually pretty unusual in our broad domain. And it seems like it's both up to the developer and then up to us as evaluators to say if there is an empirical association, is it likely to be mediated through quality of care, at which point – meaning, in this case, bad communication or the structure leading to bad communication.

And if so, you probably don't adjust for it. Or is it likely from something not quality of care, at which point it would be quite defensible to adjust. It is subjective. It is a judgment call. Sometime you have relevant data. Sometimes you don't. But, it ends up being tricky in a case like this.

Karen Johnson: Yes. Thank you for that. That helps me too. I hadn't really thought about it too much in this case of communication. So, yes, I think the long story short is we are still working it through. Just one more thing that I will note on another one of the call but I don't remember which one.

We actually had some developers who took the results of some of our readmission measures where the ultimate decision was not to adjust for social risk factors and the Standing Committee still went ahead and endorsed the measures. They interpreted that erroneously as suggesting that we think it's OK just kind of across the board not to adjust for social risk factors.

So, on those calls, we were very clear in saying that we do not agree with the interpretation that they put forward on that work. So, kind of an interesting interpretation. I don't know where they got it. But – OK. Do you guys have anything else you'd like to bring out? If not, we can end the call.

Male: Thanks, Karen.

Karen Johnson: All right.

Male: Thank you.

Karen Johnson: Thank you so much. Have a great rest of your day.

Male: Thanks.

Karen Johnson: Goodbye.

Male: Goodbye.

END