National Quality Forum

Moderator: Scientific Methods Panel
October 15, 2018
2:00 p.m. ET

OPERATOR: This is conference # 4393468.

Operator: Welcome, everyone. The webcast is about to begin. Please note, today's call is being recorded. Please standby.

Miranda Kuwahara: Good afternoon and welcome to the Methods Panel Subgroup Number Four Measure Evaluation call. My name is Miranda Kuwahara with the National Quality Forum and I'm joined by my colleagues Karen Johnson, Ashlie Wilbon, Poonam Bal and May Nacion.

We'll begin today with a roll call of Methods Panel subgroup members. Paul Gerrard, are you on the line? David Nerenz?

David Nerenz: Yes, here.

Miranda Kuwahara: Great. Sam Simon?

Sam Simon: Yes, I'm here.

Miranda Kuwahara: John Bott?

John Bott: Yes, here.

Miranda Kuwahara: Zhenqiu Lin?

Zhenqiu Lin: Yes, I'm here.

Miranda Kuwahara: Joe Kunisch?

Joe Kunisch: I'm here.

Miranda Kuwahara: Wonderful. So, we before we dive into measure discussions today, I wanted to make a few housekeeping remarks. So, this morning, we distributed a discussion guide. This document will guide today's discussion and we'll present measures in the order to respond that document.

Consensus wasn't reached for the first four measures and so we will be focusing on those measures today. All other measures received passing ratings. It will not be discussed on today's call unless a member would like to pull that measure for further discussions at this time.

If you choose not to discuss these additional measures, these decisions from your preliminary analyses will be made final. So, I'll pause here briefly and read out those measures that received the passing ratings.

The first is 0167, improvement in ambulation and locomotion; 0174, improvement in bathing; 0175, improvement in bed transferring; 0176, improvement in management of oral medications; 0177, improvement in pain interfering with activity. And those were the five. Would any member like to pull any of those five measures? OK. I think we can move on then.

So, in that same email that was distributed this morning containing your discussion guides was also a link in the body of the email to SurveyMonkey. We ask that you pull that survey out now to cast your votes on reliability and/or validity as we discuss each measure and the staff will prompt you on when to cast those votes.

I would like to mention that the link has to be – you have to re-enter the link to recast your vote for each measure. So, timing is limited on today's call. We have 25 minutes per measure. We'd like to come to consensus on all measures today, but if not, we have a follow-up call scheduled tomorrow afternoon.

Sam Simon: So, when was that email sent that had the link in it to the SurveyMonkey since we have received now I think three emails from you, folks, from the last few days.

Miranda Kuwahara: It was sent this morning, I want to say around 10 A.M., but I can double-check the time, Sam. It's 10:44 A.M.

Sam Simon: OK.

Miranda Kuwahara: Ands then the last thing I'd like to note is that this is a public call. There will be no opportunity for public comments and members of the subgroup panel cannot direct questions to developers. And for record-keeping purposes, we (would like you) to please state your name before providing any remarks. And I'll hand it over to Karen Johnson.

Karen Johnson: So, thank you, Miranda, and thank you Subgroup 4 for joining us today. I want to just reiterate a couple of things that Miranda has already said and maybe add a couple of things.

First of all, while it will be great if we can get through all four of these measures in today's call, in reality, we do and we have scheduled tomorrow's call as well. So, don't feel necessarily constrained by time. We want to make sure that we discuss these measures.

And these are probably really are most complex measures. Three of these are – actually, all four of these are based on patient report and/or there are some composite aspects, et cetera. So, they are hard to kind of get your mind around and there's a lot to talk about.

Second, the SurveyMonkey Web, I don't even know where it is. SurveyMonkey Web thing that you're going to fill out for us is not live. So, what that means is, we can't see your ratings once you cats them.

So, we – if we need to vote on multiple criteria, we will have to do those, just kind of not knowing what the score of the previous one was. So, just so you know that it's not live, we wouldn't actually know how things fall out until after the call when we look at the results.

We hope that this discussion guide is useful. We do apologize for getting it to you so late in the game. We really work hard to try to get it done faster and we just couldn't make it. So, apologies for that.

We hope still that you find it useful and we'd love to hear any feedback that you might have on the discussion guide as well as how we're facilitating the call, et cetera, with the idea of improving for next time around. So, maybe after the call, if you want to just drop us a line and let us know anything that you think went well or didn't go so well, that would be great.

I'm also going to put some apologies out for the discussion guide in general because as when I printed it out this afternoon to look through it in prep for the call, and there was a lot of typos and I don't know what was going on there.

We obviously wouldn't have put out typos if we had known that there were typos there, so somehow, (inaudible) reds squigglies were not showing up on the document, so apologies for that. I don't like sending out things (inaudible) (with them).

In terms of the discussion guide, just one other thing, we did try to be somewhat comprehensive in pulling out some of the main results of some of the main testing and also some of the main concerns that people talked about.

But we didn't note every little concern that everybody had, but it doesn't necessarily mean that we didn't think it was important although some things we thought were really more critical maybe than others. Feel free to add to things if you feel like that we didn't include them on the discussion guide that, yes, you feel like that that's something that you want to voice.

In terms of what happens to measures once you give your think and you rate things, it depends on where the ratings land. So, if you ultimately pass measures on both reliability and validity then those measures will go through the standing committee.

And along with those will go you PAs that you did earlier, the preliminary analysis forms that you've completed earlier as well as a summary that we, staff, are going to put together that summarizes not only the ratings but the discussion points that you, guys, talk about on the call. So, that will all go to the standing committee.

If a measure does not pass reliability and validity or in some cases composite criteria, the measures will not go to the standing committee. What will go to the standing committee is a notification that measure XYZ has been looked at by the Methods Panel and it did not pass and we will provide a very brief rationale as to why it didn't pass. So, for example, if you feel like the methodology used in one of the testing approaches was not appropriate, it will be very brief.

However, we would still provide to the developers both your original preliminary analysis that you filled out along with our summary of your discussion. So, (route we'd) go to the developers with the hope that they would be able to use that information to revise and update your submissions and hopefully bring them back next cycle to have them looked at again.

In terms of voting, we will be telling you what to vote on and when. In some cases, you don't – you wouldn't need to vote on both reliability and validity, so we will be very clear on what we want you to vote on.

And it may be a little confusing in a couple of places. There are a couple of measures, the last two are the four that's on our list, in both cases, those did not pass actually reliability. So, neither the two measures passed reliability.

now , according to our initial rule, we've said that measures that completely don't pass we don't have to discuss on the call, but we always put out a caveat that we can discuss any measure if – even if they didn't pass or even if they did pass.

If somebody from the Subgroup 4 is NQF staff, either one wanting to pull a measure for discussion. So, in both cases, those last two, we, staff, are pulling them for discussion. And we may or we may not need any votes on those and that will probably depends on how the discussion goes.

Let me stop there and see if we have any questions on process, et cetera, before we delve into our measures.

David Nerenz: Yes, thanks, Karen. Dave here. Sorry if I missed it just now, but I'm wondering if you could just do a refresher reminder. Since on our forms, we don't specifically say pass or don't pass, I'm just thinking ahead to what differences we need to resolve in the discussion we're about to do.

So, for example, do we have to worry about some of us say low versus some moderate? Is that a distinction between pass and no pass or is it we disagree between yes and no and whether the methods were appropriate or is an insufficient rating something that causes a don't pass? I just want to make sure I know how to focus on the key differences that we have in front of us.

Karen Johnson: Right. So, great question. So, pretty much, we don't differentiate between a rating of high or moderate. So, first of all, anything that's high or moderate is a pass. Anything low or insufficient is a not pass.

And then once we get everybody's ratings, we basically are looking at, there's a tilt to one way versus the other. So, for example, if four people said moderate and one person said low or insufficient, we're going to take that as basically the majority rules on that, we're going to put that through as being moderate. OK, so in other words, it passes unless somebody wants to kind of call that and discuss it and potentially revote.

David Nerenz: All right.

Karen Johnson: On the other hand, if there was a split, something along the lines of a three/two split where it seems it's kind of half and half, the majority is technically one way or the other, but it's so close, we call those consensus not reached. So, we want to discuss those.

David Nerenz: OK.

Karen Johnson: We hope that today's call goes well. We've done three of the calls so far and we've actually been quite pleased with how they worked out. And for the

most part, people got through the measures fairly quickly. I think one subgroup – one subgroup call, we don't need to do the follow-up call.

One, we might be able to resolve something via email or maybe not. Another subgroup were definitely having the call, so this is kind of all over the place, but I think in all the cases, people have achieved one of the goals that we were hoping we would achieve with this new design which different people are learning from each other not only in methodology sometimes but maybe more often along the lines of, "Oh, I didn't quite understand what the developer was saying and somebody was able to rephrase it and make it more clear," or "I voted low and this is why," "Oh, OK, I see what you mean," that sort of thing. So, I think it's been a helpful change in process. So, we hope that you agree to the work.

OK. So, we are going to split this up between staff and walk you though. So, the way that we're going to try to do this is give a few of the measure highlights, not everything that we have on this discussion guide. We'll talk a little bit about the ratings and the major things that we have found concerning and then we'll actually delve into the pieces that we want to talk about.

And if we need – if we have a consensus not reached, for example, on reliability, we'll have you discuss reliability, talk about that, get everything else in the air and then we'll stop and give you a minute to vote on reliability, then if we need to, we will go on to validity. So, we'll split it out that way. Hopefully, things will be fresh in your mind.

On this measure specifically, the first one, 3452, access to independence promoting services for dual eligible beneficiaries, Sam, you are actually recused from working on ...

Sam Simon: Yes.

Karen Johnson: So, what that means is you're not allowed to say anything. OK, so, I hope that's not too awful for you as you listen to the discussion, but it allows ...

Sam Simon: No, Karen, thanks. Yes. No, thanks for it. Now, I was just going to clarify what that all meant, but thanks for – that's for that.

Karen Johnson: Perfect. OK. So, let's go ahead and delve into this measure. This is a new measure. And I don't think that is such a big deal for this subgroup but new versus maintenance actually was a big deal in some of their previous calls. So, that's why we want to make sure that everybody is aware of new versus maintenance.

And for this measure, this is based on a survey. It's actually one of the CAHPS surveys – there's a ton of CAHPS surveys, I guess you guys know this. So, this is the one that goes out to MA plans and prescription drug plans.

And the item that this measure is based on is required only for the dual eligible beneficiaries. So, this is a composite measure and it basically is – it has three components. One is access to medical equipment, the second is access personal aid assistance, and third, access to counseling or treatment.

So, for each of the components, they basically compute (a mean) at their facility level and – I'm sorry, at the health plan level, and then the composite is just the average of those three components, so it's equally weighted.

The type of measure there, and I think David Nerenz is the one that pointed this out in your original P.A., and you're exactly right, it is composite that the data come from an instrument and we actually had in the discussion guide a PRO-PM, and correctly, David pointed out that these aren't PRO-PMs, there are actually structure measures. So, we do recognize that patients can report on structures (of payers) as well as process (of payers).

So, in terms of what that means with their evaluation criteria, it actually doesn't really make any difference whether it's a structure measure or an outcome measure, the criteria would be the same. This is risk-adjusted level of analysis of health plan.

So, in terms of ratings for reliability, we had three responses, two voted moderate and one insufficient. So, that is a consensus not reached by our MAP and we need to discuss it.

So, testing was done both at the score level and at the data element level. And just a note that because it is a composite measure, one of our criteria is that score level testing is absolutely required.

Now, there is an assumption in this measure that we need to confirm before this would go forward to the standing committee, and that is what is being put forward for endorsement is actually just the composite rate, not the individual component. So, we will make that – we will verify that and make sure.

Otherwise, if they wanted the composite overall as well as the three components to be endorsed, they would need to show testing for the three components individually as well. So, they didn't seem to do that, so the assumption, again, is that it's only the composite that's being put forward for endorsement.

Now, our numbering system and really our naming conventions for the measure types worked pretty well most of the time, but sometimes, they don't work as well as we've liked, and this is one of those times.

So, a composite – this came through correctly as a composite measure, but because the underlying data come from an instrument, our rules say that we expect testing of – or demonstration that the instrument itself is reliable and valid, that's included for any kind of instrument-based measure.

However, we recognize that that may not be as clear to everyone as it is to us, so what we have set for this measure and it's completely up to you, I don't think that they've really presented what we would call data element reliability and validity or at least they didn't present reliability and validity of the instrument.

Now, for this measure, because of the nomenclature and potential confusion, we are willing to relax this requirement for this measure only, for this evaluation cycle only if only that it is agreeable for you.

The expectation would be, if it does go forward to the standing committee and it's endorsed then that additional testing of the instrument would be presented either prior to going to the standing committee or no later than next time

around when it's up for endorsement. So, that's kind of up to you. (Move point) is if you, guys, don't pass the measure, right? So, it's obviously, that's something for you to think about if it – if it passes.

In terms of validity, again, the ratings were right across. So, score level testing was conducted by correlating to two other measures that there were concerns with the results in the testing, with the risk adjustment approach, meaningful differences and missing data.

And then finally, in terms of the composite construction that also consensus was not reached, and I think that was more along the lines of a desire to see a little bit more about not only how the composite as presented works but some discussion about alternative weighting and aggregation rules.

So, with that, I'm going to hop straight into the reliability things that we need to discuss. So, first of all, there were some questions about the specifications and specifically, how are the dual eligibles identified for the purposes of this measure.

So, again, I do note here that this is still good with these questions only to patients in (MMP) plans and these items are meant to be filled out only by duals. That may not be quite yet enough to identify for you how duals are identified or maybe it (is).

In terms of testing, they did the score level testing, as I mentioned. They did basically the inter-unit reliability, that kind of global or average reliability, I've heard people call it that. And they also use signal-to-noise analysis. And you see the results there. On average, IUR, 0.71; signal-to-noise ranged really from 0.33 to 0.8, and it, as expected, got better with the plans that have more patients.

The developer reported what they considered to be data element reliability testing and we could have a discussion about that. Basically, they looked at the internal consistency of the three component measures against the composite measure using Cronbach's alpha.

Again, that's – they still would need to show reliability of instruments, but I think there was a disagreement as to whether Cronbach's alpha tells us much, if anything, about the data element validation.

So, the questions for the subgroup, first of all, are the score level reliability results adequate? So, they got pretty good as the sample size went on. Kind of the next question as a segue to that one, given the fairly low score level reliability particularly for plans with small sample size, should there be some kind of a minimum sample size required for the measure?

And then finally, this data element reliability analysis, is it really appropriate, first of all, for showing anything along the lines of data element level? And maybe it's really a better analysis and maybe not, I would let you answer this, would it be actually more helpful as part of the composite analysis?

So, let me stop there. The way we'll do it is I'll just kind of encapsulate the major questions that we want you, guys, to discuss and just open it up for you, guys, to talk about?

John Bott: Well, this is John Bott. I was the one – I was the outlier noting insufficient. And just please, correct me if I'm wrong, but the big reason I did was I noted that testing was not conducted at the data element level for the composite not the three individual measures, but later on, it was clarified by NQF. They weren't going for endorsement for the three measures, just the composite.

So, it seems to me the questions on their form where they should have indicated what tests they were doing for data element level and the results are 2A2.2 and 2A2.3. In all – and in those, the responses to those two questions, all of their headings, that the bold headings they used are in regard to a score level reliability.

I only had so much time to review this this morning, but I'm not seeing quickly the – their response to providing data element level reliability results.

David Nerenz: And John, this is Dave here. I'm inclined to agree with you and I feel a little guilty for not having noted that in my own preliminary review. I – even in

terms of looking in the checkbox on the form that they're really are claiming to do scoring level of reliability.

John Bott: Yes.

David Nerenz: Now, as I try to reconstruct my thought process, I might have assumed that this comes from a very sort of well-known standardized survey, but back in that development, there has been testing, but the measure developers didn't bring up forward here.

It strikes me that it's really tricky in a situation like this because you can bring forward all the evidence you want about reliability of the whole CAHPS survey, but what they've drawn here are three very specific elements that apparently are only answered by dual eligibles. And so, I would agree with John that I don't think we have in front of us information on data element reliability because it would have to refer to the three individual items.

Joe Kunisch: Yes, this is Joe. And I did – and I apologize because this was one that I have at the last and had difficulty getting to do in the full analysis on this one. But I did read through it. I did also notice that.

But I was also wondering that maybe somebody on another committee member can kind of help me as – I was wondering why they made it a composite measure to begin with being that each of these indicators wouldn't necessarily apply to every patient or every person that filled out the survey instrument and did they address that in the analysis? Again, I didn't have time to go through every page of the testing analysis to see if they address that. But did anybody else have any insight on that piece?

David Nerenz: Well, it's Dave here. I'll do a quick answer (to see if anybody else reads this) differently. My sense of how the survey is structured is that you first ask your question, did you need something, and then the follow-up question, which is directly (germane) here I, how hard was it for you to get it?

So, it seems like if a person doesn't need one of the three things then they don't answer the question. And then I'm looking right at the nature of – in

their analysis that the reporting results of the reliability and validity, they had to discard nearly 70 percent of respondents because – for incomplete response.

And I took that to mean that somebody just didn't need one of the three things therefore didn't answer one of the three questions therefore we couldn't generate a composite score. So, what we've got in the end is the score is based on only the subset of dual eligible plan numbers, we need all three of these things. That, to me, doesn't necessarily create (a fatal flaw) but it's interesting certainly.

Zhenqiu Lin: And this is Zhenqiu. I read it differently. I have – the way they construct it and they're just so – identify respondents who have the need for one of the three items, it could be all three.

But I thought what we are trying to is to derive an indicator. For each (inaudible) first not based on patient level. So, for indicator one, you get a number for a particular plan and then give is the same – get a similar score for indicator two and then get an indicator (there and) it's called (inaudible) three and then composite (that). But I don't know. I mean Sam is on the panel. I – this is for clarification. Can we – Karen, is it OK to ask Sam to clarify?

Karen Johnson: I'm sorry, we're not allowed to let Sam talk, but I would say Zhenqiu that I would agree with the way that you described it that each – it's rolled up for all the patients who need the first one, the medical equipment and then so they roll up the actual value for access and then they do the same for the second one and the third one and then they're combining there. So, I don't – I personally did not read it that it has – it's limited to the intersection of the people who actually would need all three of those.

Zhenqiu Lin: Right. And it's how I interpreted it. And also, I thought now using the (item) already validated by CAHPS so that's why I didn't get into it, but it would be helpful if they can provide whatever validation done by CAHPS. (I focus) on the composite measure score reliability. I think that's what we were asked to focus on, right?

Karen Johnson: So, again, Karen from NQF, and yes and no, Again, this is one where I would agree with you that they didn't really provide the data element level testing that we would expect to see.

This is – but because of the – a little bit of confusion of it being instrument-based measures rolled up into a composite, we're willing to give them a little bit of benefit of the doubt if you are.

And again, that goes back to what we're saying that if you're willing to give them that benefit of the doubt, we would really ask for that testing of the instruments as needed for these items either prior to sending to the standing committee or no later than next time around when they bring it back.

But normally, we would not do that. Normally, we would be saying, "Hey, if it's not there, it doesn't pass." Again, we're making a special exception for this measure if you're OK with that. If you're not OK with that, we will send it back and we will say, "Bring that to us and we'll look at it again when you bring everything to us."

Jim Barry: I'm a little – this is John Bott. I'm a little confused. In the instructions on our evaluation forms, Page 1, instrument-based measures, it says in the table both data elements and score level testing required. I'm not sure what the confusion is. This is clearly an instrument. So, what's the confusion? Can you restate it?

Karen Johnson: OK. I think the confusion is, well, it is based on instrument-based measures. So, that part isn't confusing at all. It's being put forward in the way that NQF types measures as a composite.

So, somebody could have looked at it and said, "Hey, it's composite. So, I don't have to show that for composite measures. So, it's really both. And you're right, it shouldn't be confusing but it is a little bit, and quite frankly, NQ, we need to think a little bit more about how we display types.

Right now, a composite is just another type of measure, and that's probably also a best way to think about composites because composites, as you, guys, know, could be composites of outcomes and processes and structures and

instrument-based or not instrument-based, whatever. So, it's a little bit our fault there. So, we didn't want to penalize the developer because (we had a kind of non) clarity on measure types.

Zhenqiu Lin: So, this is Zhenqiu. So, Karen, actually, this is how I understood it and when I tried to follow instructions (inaudible) that's why (inaudible) composite (inaudible) we focus on the measure score.

Karen Johnson: Yes, yes. It's confusing and, again, that's why we're willing to give a little bit of leeway, and again, only if you, guys, are agreeable to that. If you're not agreeable or quite frankly, if it still doesn't pass, anyway, it's kind of a midpoint but I think maybe ...

John Bott: This is John.

Karen Johnson: I'm sorry.

John Bott: OK.

Karen Johnson: The only thing I was going to say is, they did say that they did some data element testing with their Cronbach's alpha analysis. While it's not what we want to see, I don't even know if that would get us anywhere either (inaudible) might want to weigh on – in on what the Cronbach's analysis might be for us.

David Nerenz: Yes, Dave here. I mentioned in my comments and I repeat it (briefly) but once you start with the premises, the three individual questions don't have to be correlated or conceptually related.

Cronbach's alpha, to me, is essentially uninformative, whether one kind of support was available, it doesn't necessarily have to correlate with another one even if you want to make a composite measure (inaudible) composite measure on completely independent things. So, I read the Cronbach's alpha but I felt it was essentially irrelevant.

Zhenqiu Lin: And this is Zhenqiu. I agree with David. I thought from experience, it's kind of (inaudible), they don't have to be correlated, right? It's still OK to

composite them. (Inaudible) consistency, yes, I agree, this is not relevant here.

Karen Johnson: OK. So, to finish the discussion about reliability, I think the questions that we really have to deal with are, number one, they didn't provide the data element testing that we would really like to see. Again, NQF is willing to give them a little leeway there if you, guys, are, so that's one question.

But even beyond that, they did provide score level results. So, are those score level results enough to get this measure through or would it, regardless of the data element stuff, the score level results, are they strong enough to get it through? So, do you, guys, want to talk any about the score level results that they provided and what you think about those?

David Nerenz: Dave here. I talked to those with (Eric), so, I think maybe all of us are seeing the same thing. And I would re-enforce and agreeing to John, that my sense, they've presented two statistics about score level reliability, both of which at least, if you take the number and face value, it looked OK to me, which I think led to my rating of moderate, but they just don't have the item level of reliability.

If the policies (inaudible) accurate for composite measures then I – at least in the reliability domain, I (inaudible), but others may have a different view of the score level (inaudible).

Zhenqiu Lin: So, this is Zhenqiu. I – yes, I always thought I'd vote for moderate in terms of reliability, and that's what I focus on the measure score. And I mean in this case, I mean the idea is if you can have the results from CAHPS in terms of the item level, but that's OK with the measure score as a composite score.

John Bott: This is John Bott. In my cryptic notes which, again, I didn't have time to look through this morning, if I were to have made an assessment on the score level, it would have been definitely in the low bucket. I can explain why. I'm looking at cryptic notes that I have.

And as far as do we give them a pass because of some confusion, it would be hard core especially, I came from being a measure developer, I say no, it's –

we feel it's important to have data level reliability information here and it's unfortunate NQF could have – sounds like better provided that but it's – we feel that's important to evaluate the measure.

It's not here and we don't – we don't have enough to go on to evaluate this measure so it's insufficient to evaluate it. It's unfortunate they would have to go through the process again, but it's just the way it is.

David Nerenz: And Karen, maybe to the extent that you can a little guidance here is good because it seems to me as we work through our individual ratings, the choice between low and moderate is kind of a subjective judgment call. I don't remember there being numeric cutoffs or any other very explicit sort of guidance.

But then when it comes to this point, moderate is a pass and low is a fail, we could create kind of a quandary because it's – two of us might look at a set of numbers and so (inaudible) that was moderate and the other one looks at the same numbers, and so, I think that's low. We're – I don't think that's fundamental disagreement but it all of a sudden means we can agree on the pass/fail outcome.

Karen Johnson: No, you're exactly right. I mean that's why everybody would love to have some threshold of some hard numbers and we haven't, as you know, been able to come up with what would be this low was too low.

So, it really is kind of your comfort level. Do you feel moderately confident that this measure has enough reliability to be used? If you have low confidence in that based on those numbers then you would say low, and that's really (inaudible) so.

David Nerenz: So, Karen, if I could just ask you a question. It looks like for the score level, we've got two things. There's that statistic and the reported (net) value of 3.42 and a tiny little P value. And they report reliability of 0.71 and then they've got a signal-to-noise ratio of 0.61.

I will confess freely knowing even that this is a public call, I am not familiar enough with that range of measures to be able to make a confident judgment

between what's low and which moderate. I'm willing to be instructed on that point.

Jim Barry: Well, this is John Bott again. Here are three – four factoids that I highlighted in red, and to me, I highlighted it in red to myself, is saying, "Hey, there's something – there's something up here."

And one, a self professing moderate to poor performance and signal-to-noise. Secondly, the signal-to-noise as the average signal-to-noise across the plan was 0.61. And thirdly, they say the mean composite performance score was 0.69. Fourth, I highlighted in red, and so the coefficient of variation was 0.09.

I – if I highlighted those in red, typically, what I do is I simply go out to the Internet and look up what kind of values are desirable and undesirable, sometimes I go back to their form, and sometimes in the form they say, this is the rating in good range and in a poor range but – so, I wound up highlighted those in red unless I felt pretty confident that those four instances jump out as (being pretty poor). But I can certainly stand to being corrected. I'm not – I'm not great at this – at this – in this sort of instrument of evaluation.

Zhenqiu Lin: And it's kind of tricky, right? (Inaudible) same group are (inaudible) all groups that we (inaudible) we had consensus. And then a couple other measures at the facility level, reliability, it's one, I put that 100 percent.

I remember one goal (and the 75) percentile is one. Do I really believe that? A risk-adjustment – well, risk-adjustment measure of reliability score at 100 percent? So, sometimes, there's more than (inaudible) instead of score, you had to look at how they calculate that.

I mean I can see how that happened. You're using sort of – using his observed result without accounting for risk-adjustment when you have that P equals zero or P equals one or you have N really big and that's where you can get one. But we know this isn't sort of exactly reliable, I mean when you have the (inaudible).

Karen Johnson: So, this is Karen again. I don't think we have much else that we can offer you in terms of how to kind of weigh those values. I mean they did some kind of

bootstrap bias corrected confidence interval. Actually, never mind, that's the data element stuff, so just ignore that.

I think you're really looking at this average signal-to-noise at 0.61. The next (board there) a quarter of the plans have a reliability of greater than 0.7. And those with plans with reliability less than the 0.7 had 102 (beneath) on average anywhere from 45 to 201, so, obviously, at least somewhat a function of the sample size. So, I think I'm not sure that there is much else that we can deal with that other than what you just kind of (inaudible) on whether you think that's strong enough or not.

So, let me pause here and see if anybody else has anything else to say on that. If not then I'll give you some instructions on how to vote.

Zhenqiu Lin: I mean in some cases related to this, not exactly relevant to this one, but I think for the future, maybe you can just (simply just_ ask for all the measures to be reported not just point estimate, also uncertainty, right?

I mean even so you report a reliability really high, that depends on the data you use for the testing. If this will happen, the data you use have huge (inaudible) facility variation and then you are more likely going to get a higher reliability.

(Where exactly) in the measure apply to the real setting and then you may not get a sense between facility variations. In that case, right, the reliability will be small though. So, it's not (inaudible) will be there forever. Now, if you now get 0.8 or 0.9, it will always be 0.8 and 0.9.

So, instead, I'll just – I thought maybe (inaudible) report something (inaudible) statistic, you account for uncertainty, you report that, right? Now, you have a score that's 0.8, is it reliable or not? You can quantify that.

Karen Johnson: So ...

John Bott: This is – yes, this is – oh, go ahead.

Karen Johnson: No, go ahead, Joe.

John Bott: This is John Bott.

Karen Johnson: John Bott.

John Bott: Going back to David, and I'm always badgering NQF about definitions, but going back to what David said is, I didn't realize before that medium and high was, in NQF's mind, pass and low was fail.

So, when I was – so my color-coding in my form of red, red was telling me that if I have to sum this up as the end, this is – this sways me towards low. But at the time of doing that, I wasn't thinking low meant fail necessarily. Low just meant low out of the spectrum of – from low, medium to high.

So, for me, to really go back and say, would this mean pass, I'd have to do more (inaudible) because I think if it was – if it was pass, so, again, maybe a call (inaudible) can we define low, medium and high with – in those terms that – I don't know people's names, but the NQF person said.

Also, just maybe a housekeeping comment, I see we only had three people weighing in on this, two moderate, on insufficient, but we have five people on the panel apparently. I'm guessing Sam didn't (fill it) because of conflict of interest or whatever, but that means there was another person who didn't vote to that. Is that a problem for NQF then in tallying this?

Karen Johnson: It's not a – that particular thing is not a problem because we're still discussing it on the call. If we were making just a final decision based on only three, that would be a little bit tougher, but we specifically wanted to discuss on the call and we have – we have – we have four people on the call who can vote, so which is reasonable. Sam can't because he's (inaudible), but otherwise, we would have five.

Going back to the low versus insufficient versus moderate and high, apologies for that, that's just something that is still engrained in us that, you're right, maybe we never told you that moderate and high, when standing committee see this work, we are looking for a more than 60 percent. Is that what our cutoff is? More than 60 percent of the standing committee has to vote moderate or high to push a measure through and call it passing.

So, that's just kind of always been our – the way that we operate. So, you're right, we may not have specifically told you, guys, that. There is a difference, and I think you, guys, know it, but it probably (bears repeating), there is a difference between low and insufficient.

Insufficient basically means you just feel like you need more information before you can make a reasonable rating. Low typically means that you though that results were just low or potentially in kind of the opposite direction or trying to prove one thing and they kind of maybe for the opposite or something along those lines. I think I got all your things, John.

John Bott: Definitely.

Karen Johnson: So – yes, and going back to Zhenqiu's comments about variations and showing variations, I agree with you and as a matter of fact, very soon in the next month or two, we're going to put forward some potential recommendations that we think you, guys, as the Methods Panel, would agree with, and that's one of them.

When you show signal-to-noise, don't just give us a mean or a median, but give us some other kinds of statistics. And to be fair with this one, I think they did that. For the signal-to-noise, they did provide a range from 0.33 up to 0.86.

So, they did give you that range and then they also told you how it varies a little bit based on size. It's not completely like a decile or anything like that, but at least a hint that those numbers are, to some extent, the function of the sample size that they have.

But we will come back to that variation question and get a formal kind of recommendation from you, guys, that we can put forward so that soon, we'll be able to expect that always. Any other comments before we go on to vote on this one?

John Bott: Well, this is John again. And it is just in the spirit of a comment and tell me if I heard this wrong, but there's one person who did not fill out the form other

than Sam, but you're going to allow them to vote. I would question whether that's appropriate.

I mean the evidence – the sort of evidence we have that you really poured over the measures, in fact, you filled out the form and you (that might be your) tally in and do not know (to be the creative which) the person weighed in and allowing to vote, which seems to be the final upshot doesn't seem fair.

Joe Kunisch: Yes, and John, this is Joe. That actually was me. This was the measure I mentioned that I didn't – it was the last one and I didn't have it to the point where I was filling out the form at the end and I had some questions about it.

But I'd be – I'd be perfectly fine from abstaining from the vote for this. I did read through the measure and know what the ratings that I would have put forth and I didn't have some questions around why they didn't do the individual data analysis kind of along the same lines as you have been discussing. But I don't disagree with what you're pointing out, and from that perspective, I can abstain from voting on this particular one since I didn't fill out the whole form.

Karen Johnson: So, this is Karen from NQF. And we would look at it as – Joe, if you – if you say that you've worked with the measure, you understand what the measure is about, you see what the numbers are, et cetera.

If you feel comfortable based on that even though you didn't fill out the form then we're perfectly happy especially since you were able to be on the call and hear the back and forth. We're perfectly happy for you to vote. We don't feel like that you should abstain unless you just feel like that you don't know enough about the measure (to do that). We wouldn't want you to vote on the measure if you happen to attend on the call. So, I'll leave that for your decision.

In terms of the vote, I think it really comes down to the results of the score level testing, first of all. So, if you feel that the results of the score level testing are kind of good enough then you may want to vote moderate on this – on this measure.

I don't think it would be eligible for high but I – my sense is that nobody would come close to voting high. But we still have that question of the data element testing, which was not provided. And if you feel strongly that you want to see that before pushing it on to standing committee then we would want you to vote insufficient.

So, basically, if you vote moderate, you say that you feel like the score level results are adequate enough, if you feel like you really want to see the data element testing before, you see – before it goes forward, you would vote insufficient.

Potentially, you have a third option which is that if you feel like that the score level results are not good enough, you could vote low even if you'd be willing to give them leeway on the data element testing. So, you really have three options. Hopefully, those make sense to you. I can go over them again if you'd like.

David Nerenz: And Karen, Dave here. I know we spent a lot of time on this, but I think it's time well spent because we're explaining some important issues. And I'm trying to reconcile, so this moderate versus low, in full respect to John's (color) coding which I imagine more important than mine.

When I look at this signal-to-noise ratio in the reporting of the average value of 0.61, part of what was in my mind I think that led to a moderate rating was it seems to me that this is going to be my (inaudible) that's actually quite a bit higher than a similar statistic for other measures (inaudible) quite widely used and they're also already NQF-endorsed, I'm thinking of the hospital readmission measure, for example.

Karen, anything else who's on the phone from NQF staff, is there anything you can tell us or something that popped in your head about what ranges of signal-to-noise ratio you see and is there any pattern you observe between what flies and doesn't fly in that particular statistic because I was inclined to think that 0.61 is not so bad?

Zhenqiu Lin: Hey, just because you brought up this, this relate to how we automate it because – and so, that's the tricky thing (different). I mean there are many – so many different way to calculate reliability.

I mean the way (inaudible) test like recently the reliability and what you get is always the lower bound of any reliability you can get. I mean John, actually, I had a paper on that. So, it's not just the numbers. Sometimes, we have to look in exactly how they were calculated.

David Nerenz: Yes, that's right. I'm just trying to do an accurate and clear job with the idea of trying to see where we can find consensus and taking what's in front of us and slating it into this rating level.

Karen Johnson: Yes, it's a little tricky. I think it is fair to say that we have endorsed measures that have signal-to-noise ratios that are less than 0.7. Often, we've said kind of 0.7 is kind of the rule of thumb, but knowing that not everything gets to 0.7, that sort of thing.

So, from that perspective, yes, things have gone through with less than 0.7. I can't really give you a lower bound because I'd just be making it up. I don't really know off the top of my head what those would be.

I – personally, I think the thing that I would also consider here is not just the overall average but knowing that it's greater than seven, that might tilt it a little bit one way or the other or maybe not. So, I can't help you as much as I'd like to be able to help you.

David Nerenz: OK. Well, I would have thought that the range was more extreme. I thought – and this may be just my own way of paraphrasing of this interpretation that (any help of this) article during analysis and the readmission measure at least in my own interpretation (appraising it was) something like 5 percent signal, 95 percent noise which is a ratio that gets you nowhere near 0.7. And that's part of the context that's in my head, I may be wrong.

Karen Johnson: You may be right because I think I've heard that too. I just don't recall – and ...

David Nerenz: (Inaudible).

Karen Johnson: I mean we can Zhenqiu might know because those are – those are measures out of his shot, so he might know a little bit better than I would. I don't know if you know off the top of your head, Zhenqiu what your ...

Zhenqiu Lin: I don't know the exact context but I think (inaudible) point seven in the document (inaudible), it's really referred to the internal consistency (inaudible). In the context, we talked about PRO measure. I think Dave also – Dave Cella also mentioned that, right, it's really more – I mean that's (inaudible) use for – to evaluate the adequacy on internal consistency of the scale.

Karen Johnson: OK. I mean eventually, I think the Methods Panel will have to grapple a little bit with what are reasonable ranges that we can accept. And to support you in that, we will try to go back and look at the various measures and kind of see what method would you use and kind of what some of the statistics were.

Unfortunately, that's all in the submission forms. We have to basically go through all of that. We don't have it kind of separated out in the databases that we can search through. So, when the time comes to do that, we will do it, but it will be a very labor-intensive thing to do.

David Nerenz: Just a suggestion, we spent about an hour and we haven't even got to validity (inaudible) measures yet. Maybe we should – maybe we should do vote and at least move ourselves along (there).

Karen Johnson: Yes. I think we pretty much squeezed as much juice out of this one as we're going to get. If – it's going to be how comfortable you are with the score level results and would you be really upset if it didn't go forward versus are you reasonably OK if it does.

And then kind of along those same lines with the data element testing, if you just feel like you really shouldn't go forward without that testing that you, guys, can look at then you would vote insufficient.

So, with that, I think we'll actually go ahead and open your link and select your measure, that would be 452 and vote in reliability.

John Bott: So, this is John. Just a process question, so then the question under that is about validity, so we check the box that we want to under reliability. We just leave this page open until we're done talking about validity before we get gone.

Karen Johnson: Yes, yes.

John Bott: OK. Thanks.

Karen Johnson: Yes. And you're right. I mean, we have (spent) a long time. I think it's been useful because I think it might help with our other measures and there are certainly points that we have to kind of grapple with the across method panel.

But let's see if we can maybe go a little bit faster with validity. Again, the rating or split between the three folks who answered the (TAs).

So, the score of our testing, they did it by correlating with two other measures and they explained what those measures are, the user experience, rank order, correlation. You see the correlation values there. The first one with care coordination measure fairly low correlation and not statistically significant according to the (confidence level events) there. The other one was a much higher correlation and statistically significant, again, according to the CS.

The meaningful differences where they did some (RBS) test and some (fatigue) test and showed that at least some plans had more than a quarter (hedge) scores that were significantly either above or below the mean. In terms of comparisons with other measures, that's not so shabby. So, that's actually quite good.

In terms of risk adjustment, this was probably one of the bigger concerns. The risk adjustment was done using age, education, self-reported health status, self-reported mental health status and use of a proxy response. But they basically followed the overall CAHPS methodology rather than developing risk adjustment models for the composite itself or in this case, for the

individual components because they would risk adjust the component and then take the average to get the score for the composite.

So, they show you a little bit in terms of the adjusted (squares) but the concerns were (some rejection) coefficients were quite small, meaning not statistically significant, and two, actually had difficult signs, positive versus negative, in one or two of the three indicators.

(So, with) some concerns with the missing data and the analysis that were done there. However, the developers did – there's been quite a bit of effort telling you about analyzing valid and nonvalid responders and make – trying to make the case that the nonvalid responses were infrequent and they felt would not bias the results.

So, in terms of thinking about validity, you're back to the score level validation, are the results adequate? So, you have the two Spearman rank correlation coefficients to look at.

We're relying on that overall CAHPS methodology. Is that a fatal flaw? What do you think about that? There didn't seem to be (a mention made) that we saw a mode of administration.

So, if they are – if there are various modes, how was that taken account or should it have been then kind of probably a little bit less – well, I shouldn't say that. Is there a concern about the meaningful differences and similar performance across plans, as I've said, the 28 percent being higher or lower than the average actually compares fairly well with some of our other measures that go through? Are (missing) survey responses handled appropriately or do they bias the measure?

We also have that question of the data element validation. So, again, that's not here. It should be are you willing to give them a little bit of leeway or are you not?

So, let me stop there and see what you guys would like to discuss on that.

David Nerenz: Well, Dave here, just to get the ball rolling. I was (going to rate this) low and I didn't realize why I did that, that that was essentially (kiss of death).

But I guess I'm – just in the sense of that, it seems to me that what we've got in front of us, if we (peel away – peel the) extra things is that the argument for score level validity is based on correlations with two other things that maybe are related, maybe aren't related. I understand between the two comparison points, one is likely to be more than the other and that is, indeed, the pattern of correlation.

But I just felt across the range of the kind of validity testing, we see another measure. So, at least, (those are in our set) that we had in front of us.

This was just kind of (weak) and I'm not sure that's any fault of the developers because they only have certain datasets available to them and it's reasonable to go on and say, look, if we calculate this composite and then we look for something external, what can we find and what would we expect.

The trouble is if we then try to just go in the context of other measures running through it that we've seen, is this a strong evidence of validity? Well, I don't think so. And as I ended up calling it low just because the contextual foundation, I don't think, is as tight as we've seen another measures where we say, OK, here's the – this is the measure we're focusing on, It should correlate strongly with these three other things or it should, I don't know, I don't want to ramble too long here.

But they have something. It's clearly something (than) nothing. But it – I ended up calling it low.

John Bott: Yes. This is John. The reference prior work done with the CAHPS but then they – but this isn't necessarily CAHPS. So, I don't know why they keep on leaning on CAHPS. I'm just reading it now, just reading my (current note in).

Empirical testing wasn't conducted, that these risk factors were appropriate for these measures. They're just leaning on CAHPS. (Then they go on) to say that their testing form says, quote, "no testing was conducted to validate the adequacy of statistical model whether risk adjustment approach." I just think

that's – it's unacceptable. We don't – we have insufficient information here then.

And I was the one that made the comment about the mode of survey administration was not considered. It is a big risk factor for (age gaps), I know, and they didn't even consider it. So, to me, it's a significant flaw.

Zhenqui Lin: This is Zhenqiu. I think in term of validity testing, I think, obviously, (they only have) very limited testing, right, maybe due to the limitation on data. But they were positive, somewhat positive. So, some evidence, I don't know (whether I can say very strong).

My main concern is more with (them) trying to use CAHPS methodology now in terms of (inaudible) in terms of risk adjustment. CAHPS has (inaudible) about why they include this (very important) risk adjustment.

The key (example) use of proxy responder. (This is the one of the) most significant predictor for all three indicator. And the interesting thing is for (one) indicator it's positive significant. For (one), it's negative (inaudible) and another, it's nonsignificant.

So, this is sort of a (core into the) question (whole regime) underline the reason why CAHPS include proxy responder. But, I mean, for this particular setting, right? So, that's a – that's where I wish the developer has start to deal only exploration instead of relying on CAHPS methodology.

Karen Johnson: OK. Does anybody else have anything to say about either concerns that we haven't already voiced or – I think it is a little tricky and the score level results and correlating with other measures. I will tell you that, David's right. It's not nothing.

And they did a better job in terms of their submission than what we used to see. We used to, sometimes, people would just do a correlation that tell you basically nothing. They just say we correlated this with this and here's what we got.

And at least they told us what the measures were and the hypothesized relationships. So, it's moving in the right direction. Whether or not those are good enough measures to really validate, I think, is yet another one of these. You have to weigh it yourself and figure out.

We have had, in the past, measures that have gotten through when they've kind of done the very trivial solution, the obvious measures that probably don't really validate much but if they turn – if the results were wrong, they might invalidate it. If that makes any sense.

So, we have a little bit of that to give you some context. I think eventually and maybe now's the time and it's kind of up to you. Eventually, we do need to start saying, hey, these measures that are being correlated in this construct validation analysis need to make a really good sense and need to be the right now.

Now, again, whether today is the day that we start being more strict on that or whether we say, hey, we've got to win because they – we're at least hearing some hypothesis and getting a little bit of interpretation, again, way more than we used to have, is that good enough? I think that's completely up to you.

Let me pause real briefly. Is there anything else you want to discuss or are you ready to vote? OK. On this one, I'm trying to think of what your options are for voting.

So, number one, if you feel like that you're not willing to let this go through without seeing the data element testing, you would vote insufficient.

If you feel like that – and I think this sounds reasonable. If you fee like that the risk adjustment methodology not kind of doing it from scratch with this kind of studying with tax measures, if you feel like that that is basically a fatal flaw, I think we would ask you to vote insufficient on that as well. OK? So, in other words, they did give you a sufficient model (team) to consider.

If neither of those kind of fit in your thinking, then think about the score level testing results that were provided if you feel like those are probably reasonable enough and then you would vote moderate. Otherwise, I think you

would vote low because you would be saying that what they've done actually did not validate (your measure).

So, hopefully, that's clear. I'll give you a second there to push back or ask anything. OK. If not, go …

Zhenqiu Lin: Karen, I …

Karen Johnson: Go ahead. Sure.

Zhenqiu Lin: So, this is Zhenqiu. Just a question, I think this is not the only measure (use) some item from a (poor or) existing instrument, right? Are we – I just want to be consistent (across, right)? I remember I have seen other measure also using some individual question from CAHPS or other source.

Karen Johnson: Yes. I mean, any other measure, we would say, hey, if you didn't have it, then go back, come back when you have that in your submission. Again, the only reason that we're even (positive) that we might allow some leeway is because we feel like there some confusion with this being a composite (and if) patient reported an instrument-based measure suggest because there's potential confusion there – and to be honest, I think there's a little (inaudible) confusion amongst the channel members (inaudible) should you be looking for both or (it won't get enough).

And again, NQF, a little willing to be more flexible and very much willing for you guys to say no, we really want to see the data element testing before it goes forward.

Male: (Inaudible).

Karen Johnson: (So, go ahead) and put in your votes for validity and we are almost done with these measures. So, there's a little bit more to talk about in terms for the composite construction. In this space, they did Spearman's rank correlation coefficient between the components themselves and between the compliance and the composite.

And you see the value of care. It looks like from the most part, the correlations are fairly moderate to high and (inaudible).

Male: (Inaudible).

Karen Johnson: Yes. It sounds like there's an airport …

Male: (Inaudible).

Male: (Inaudible).

Karen Johnson: Yes, maybe – can you mute your phone, Joe, just until you need to talk. I think it's Joe. OK. There we go. That's better. Thank you.

Karen Johnson: So, they did the correlations, again, between the components themselves and between the components and the composite. I think the concern was that while they did talk quite a bit about their weighting methodology and why they ended up choosing (waiting), they didn't do quite so much in terms of their overall aggregation method.

So, possibly more a question of I'd like to have seen more as opposed to I didn't like what you did tell me. I hope that's fair. But let me stop there and see what you guys had to say about the composite construction.

Male: Yes, I think there's – I really – I thought it was basically OK and was pretty straight forward. You say two or three things that represent elements of some larger concept. So, it do not have to be correlated with each other highly necessarily.

I mean, often these things are three reflections of the same (concept if you do do that) but you could just say, well, there – they're three separate things and you take them together. They give us this bigger picture of this bigger concept. And in the absence of some technical reasons (that weighed other than) equally, you can (inaudible).

So, I thought basically, just in the pure compositeness, it was OK.

Zhenqui Lin: I agree.

John Bott: Yes, this is John. Go ahead.

Zhenqui Lin: I'm sorry. This is Zhenqui. (I agree).

John Bott: Yes. This is John Bott. I think you need to defend if you're going to use straight weighting. I mean, when I was (at ARC) and we we submitted measures, we submitted composites. It was a big discussion of why did we use the weight we used and just say, just we're going to go to the default of the straight weighting. I think you need to defend that and I – they just (were silent on it).

Joe Kunisch: Yes. And this is Joe.

Karen Johnson: Go ahead, Joe.

Joe Kunisch: Sorry about that. I was just going to agree with that because I think that's where I had the most kind of struggle with this and as – they made (the two) composites that I didn't see anywhere where they were (used) addressing that. I mean, (we are waiting) being that each one of these isn't going to apply to every patient.

Karen Johnson: OK. I'm not sure there's too much else to say about that. Is there anything you want to (bring up)?

John Bott: (This is John Bott).

Karen Johnson: Go ahead.

John Bott: For housekeeping on the SurveyMonkey, there's' – correct me if I'm wrong. There's not a question to change – to weigh in on this, right? There's just, really, the scientific acceptability question and – sorry. There's just reliability question to weigh in on and the validity question, right?

Karen Johnson: Excuse me. The composite section has just been added. So, when you're ready, you can submit your votes for reliability and validity, reenter the link to this measure and then there should be an option for you to weigh in on composite measures.

John Bott: OK. I see.

Karen Johnson: Sorry about that. I think we missed it. It was composite when we built the SurveyMonkey. So, kudos for getting it fixed so fast.

So, we have spent way more than our allotted time on this one. Hopefully, we'll find some time going forward and we'll see how we do. I want to go ahead and I know you guys are probably still finalizing and opening and voting and stuff and I do want to go ahead and at least begin the next measure.

I don't know that we'll be able to finish it but we might. So, Ashlie, was this your or was this one mine?

Ashlie Wilbon: Hi, Karen, this is Ashlie. I was going to tee up the panel discussion on this one, but feel free to jump in on anything.

Karen Johnson: Okeydoke. Thanks. So, are you ready to go ahead?

Ashlie Wilbon: Yes. I'm ready.

Karen Johnson: OK. Thank you.

Ashlie Wilbon: So, this is Ashlie Wilbon and I'm going to tee you guys up for discussion on the next measure which is 3461, Functional Status Change for Patients with Neck Impairments. This is also a new measure. It is the patient reported outcome consisting of PRO, Patient Reported Outcome measures of risk adjusted change and functional status or patients 14 years of age and older with neck impairment.

And functional status is the step using the Neck FS PROM which is an item response theory-based computer adaptive test or CAT. They also have a handwritten or static paper-pencil 10-item short form that is available for the measure as well.

So, the measure focus is the average residual score which is the actual change for minus the risk-adjusted predicted change score. They did offer some threshold values for clinician versus clinic level of analysis. And the

thresholds for clinician were about 20-plus patients per year. We will discuss that a little bit, I think, in the reliability testing and then for the clinic level of analysis for small clinics of (1-3) clinicians of 10 or more patients per clinician per year.

And then for large clinics for more or more clinicians, the threshold was about 40 plus patient s per clinician per year. So, again, we'll talk a little bit more about that later. They do allow proxy responses and I already mentioned that the data gathered via this computer adaptive technology or computer adaptive test for the 10-item short form.

The level of analysis, again, is the clinician group or practice level in the individual clinician level. They did also mention other level of analysis, so we'll need to, if this measure makes it through, and some of the feedback that we'll give to the developer to clarify what other level of analysis that may be that wasn't clarified.

It is risk adjusted. And in terms of reliability, there was not consensus among the panel members that submitted their PAs. There's two moderate, one low, one insufficient and one undecided and this measure was pulled or a request to pull this measure came through to us which is another reason why it is up for discussion today. Some of the key issues that came up for reliability were around data element level testing.

And, let's see, they did provide data element level testing to demonstrate their reliability of the instrument which was assessed via Cronbach's alpha and item response theory, person reliability analysis. In addition to the individual level scores, they looked at the standard error of measurements and analysis of the minimal detectable improvement.

They also did signal – I'm sorry – score level testing using signals from noise analysis and there were some concerns expressed by panel members around the specifications and the methods that they used in testing. So, I'm going to just skip down here to the items to be discussed and we'll just focus on reliability for now.

Some of the issues around specifications were around kind of the lack of specificity around the data sources and some definitions for data elements around which patient count with – as having neck impairment. The conditions that they included seem to be not really specific and they have some language about these conditions are not limited to but wasn't clear on what that full list of conditions would be that would qualify folks for having – for being in the measure.

The numerator statement was not described as a change score. There was no description, again, of how proxy responders are used and unclear of how an episode is defined or how the discharge was determined.

In terms of the testing for the data element reliability, just a kind of in summery of what they did here, they used responses from the full item. The Cronbach – they had a Cronbach's alpha score of 0.98 and an IRT based person reliability of 0.96. The scale of reliability of the score was 0.91.

The median error of measurement of individual scores were stable approximate in that continuum that range from 3.7 to 3.9 which as the functional status points which corresponded with 7.2 to 7.6 of the functional status points within 95 – 95th percentile profit interval.

The minimal detectable improvement for the overall score range from 6.8 – sorry, there's a – I'm missing datapoints there. And I'll just kind of – I'll move forward from that one. And the – move on to the score level reliability.

So, the clinician level of analysis and the clinic level of analysis was also submitted. The results are there. I won't read through them in detail but we can refer back to the testing attachment that's needed for results on that.

Some of the concerns within methodology for the score level reliability were around the formula that was used and there was some concern around whether or not the forms that they use versus the description of the method actually align that the formula that they use may not be correct or may have been misaligned with what they were describing that they were actually doing.

It was also unclear whether the method was based on misassessed model or whether there's an (error base) for each of the providers and also unclear if the analysis was based on raw change scores or the residuals. And then also, relatively small overall variance was explained by the signal.

So, I'm going to stope there with reliability. And open it up for discussion. And we'll vote on – discuss and vote on reliability before moving on to discussion of validity.

Joe Kunisch: Yes. This is Joe. I'll get this one started because this one, I really struggled with and you don't know if we're allowed to say where we were on that scale of the voting.

But I had extreme difficulty with the measure specifications and even discerning what the population was that they were trying to measure was very ambiguous. I think I put an example there, health problems, dash, allergies listed in the code back but what is that? Is that an exclusion? Does that mean an allergy to anything?

There was no definition around these data elements and using the code book. I wasn't sure why they didn't use ICD-10 codes or CPT codes to define some of these. So, even reading through just that specifications, I couldn't determine a lot of this and just to make sure that it wasn't just something I wasn't able to see, I even sat down with a couple of our expert abstracters and said, OK, if this was the specifications we received and we knew we had to review and validate this measure, could you (reset) down there and even define this population and go forward with this?

And they were, at this time, the same place I was. So, being that, it's hard to go on to the other sections when you can't get by that because if you get in to reliability and validity testing and don't understand what they're actually even testing or the population that they're testing, that's where I kind of struggled on that.

John Bott: Yes, this is John Bott just to jump in after Joe because I would just second what Joe has to say. In their NQF format they completed, the S& where they're supposed to give denominator details, underlying details, they (rattle)

off a couple of conditions and they say including but not limited to, that's a no-no.

In all my work with admin, I had 50 plus measures endorsed by NQF, you don't say that. And then in the two Excel files that they provided with some – a but more specificity, they label – they list the number of conditions but they don't define the conditions and they don't say the process by which that data has captured.

So, that's why I rated it a low and – but for other reliability testing results, the testing result seem OK. But given the definition of low which you should rate low if you believe the specifications are not precise. So, to echo what Joe said, yes. The specifications are definitely not precise.

Sam Simon: Yes. So, this is Sam. I had a similar concern but I was more concerned and maybe this was something I just missed but I was the one that tell if the analysis was based on sort of the raw scores, if the reliability results were based on that or if this was the risk-adjusted score.

And I don't know if others sort of had that same concern or maybe that was just mine. But I – so I rated this as insufficient.

Karen Johnson: And this is Karen. Just to be clear, the measure is about the residuals, correct? So, we would expect analysis to be based on the residuals. Am I saying that correctly?

Sam Simon: That's – well, that was my expectation, but I didn't – that wasn't clear at all from the writeup, at least, to my reading.

Karen Johnson: OK.

Davide Nerenz: And Dave here. I thought that was clear when they're talking about validity and some of the things deeper on and their testing document, but not so clear upfront. And I'm agreeing with that. I'm trying to scan it right now and I see pretty clear estimate and display of residuals later but not in the reliability section.

Sam Simon: Right.

Zhenqiu Lin: And (this is) Zhenqiu. So, in different location, they talked about reliability based on (inaudible) tech model. If there's one (inaudible) (tech model) you get one, right, error variance (that may seem to imply) and one error variance (for each provider), as I remember, the (inaudible) about that.

Joe Kunisch: So, this is Joe. Just to clarify, when you say this measure was pulled, does that mean they – the measure developer retracted it from submission?

Ashlie Wilbon: No. That means that someone from either a staff member or one of the panel members requested that this measure be brought up for discussion on the call.

Joe Kunisch: So, is there an actual ask then of the – of us, like, are we supposed to resolve something here or this is just for discussion? Sorry, if I'm not …

Ashlie Wilbon: Yes. Right. So, the – for reliability, there was one person who actually didn't cast the vote on reliability. It was undecided. So, we ended up with five people that reviewed the measure but one person was undecided. I was going to try to go back and forgot who that was.

But just to – if we can have everyone cast the vote that's either, high, moderate, low, insufficient, that would give us what we need to kind of complete the tally of where this measure lands, so – on reliability. So, that was part of the reason for bringing it forward as well.

Karen Johnson: And just to be clear, no matter what the undecided person is, if the undecided person had casted a vote one way or the other, it still would have been consensus not reached because it would have been a three-two split one way or the other and with three-two splits, we always would bring it for discussion. So, yes, we will ask you to revote on this.

Joe Kunisch: OK. Can we …

David Nerenz: And actually, if I can just take half a step back. Sorry, I thought (that) was going into a pause there, I'm – again, just getting (to the form) near bottom of

Page 17, they're talking about the signal and noise analysis under the bigger reliability section.

Those dependent variables, functional status change, discharge to physical therapy, adjusting for all variables used (by follow up from) risk adjustments. So, they don't quite use the word residual but I think that to mean that effectively, that's what going on. I guess we (can quibble about) residuals versus risk adjusted scores, but (I think conceptually), they're using the adjusted scores, not the raw scores.

Joe Kunisch: That's helpful. Thank you. Yes. No, it does. Thanks.

Ashlie Wilbon: So, I'm not sure – sorry, go ahead.

Karen Johnson: Sorry, Ashlie. I just wanted to ask (the queue) a little bit having the – (saying) they did the mixed effects model but then also saying that they had the error variance for each provider. Could you expand on that just a little bit? (Or is it) like a typo in just what they said and how they stated it or does it feel like maybe they did something like their method wasn't quite right or do you have flavor about that?

Zhenqiu Lin: It's had to tell. I was – when I read it, it should be only one. But in the way that it's described, it seem to imply multiple ones, multiple – say for each provider, you get an error variance on the (inaudible) to model. I just – I'm not sure I can see that and that's why I am a little bit confused on that.

Karen Johnson: OK. Thanks.

Joe Kunisch: Just from the interest of time, can we move to a vote? I don't know when we know we're ready for a vote but I'm feeling like I'm ready.

Karen Johnson: Yes, I think …

Ashlie Wilbon: Yes. Yes.

Karen Johnson: … nobody else has anything else to say, yes. Sorry, Ashlie. I'm going to be quiet so you can take this one. I'm sorry.

Ashlie Wilbon: No, it's OK. Feel free. I was just going to, yes, say that feel free to go ahead and vote. And I actually, Karen, tap you for to kick us off for validity as they're doing that.

The votes that are submitted for validity, at this time, were one moderate, three insufficient, and one, again, one undecided. So, with the one moderate and three insufficient for validity at this point, the measure would not pass. We did have the one undecided, so to Karen's point again, we needed to have an understanding of where the undecided vote would lie, so we would know whether it would be a consensus (site breach) or whether the measure would not pass validity.

So, we did want to just kind of bring this forward to you guys to see if there's any further discussion on validity and I will just highlight some of the issues that came up with validity around testing. And I think there was some concern or lack of clarity around whether testing was done at the score level and whether or not it was adequately presented to determine that it was done for both the clinic and the clinician level of analysis.

And just a note here from NQF, because it is an instrument-based measure, both data element and score level testing is required, also some concerns around the threats of validity, there were some concerns with discerning meaningful differences at the clinician level particularly if the measure is going to be used in accountability purposes and since the results seem to indicate that (the significant) differences were noted primarily at the upper and lower end of the distribution, lack of clarity around how incomplete surveys were handled, concerns with the application of the risk adjustment approach and possible instructional bias, also concerns with the limited testing of the risk adjusted (change) score and with the lack of the differentiation of the clinic versus clinician level validity.

So, let's see. There's one other concern here around that what some of the providers may have patients (look for) scores at discharge, providers may have the same score but for different reasons and, for example, like if all the patients improved a little versus some patients improved a lot while others didn’t improve or got worse, whether or not those were well differentiated.

So, I'm going to pause there and see if there's any discussion on validity and call for a vote at the end just to make sure we have that undecided vote results and we can determine the final (blow) for validity.

David Nerenz: Dave here. Let me just raise a quick question to my colleagues. In reviewing what I put down here and then also going to their testing (form), I may had been a little too generous or too kind here because what I'm seeing under validity is really what consider an (impressive amount of) validity information on the data elements.

I'm not sure I'm seeing it, perhaps, even at all for the measure score. And they have this very pretty grants near the ends about, variation of performance, so that's actually more of a reliability concept, I think, than a validity concept.

So, I guess, I'm going to turn to my colleagues. Do you find evidence here about testing for validity at the measure score level at all?

Zhenqiu Lin: I agree with David. This is Zhenqiu. I think most of the validity testing (are done) at a patient level. I mean, what's missing is really a provider level. That's a little bit on Page 32, (maybe 41), on that Section (2D4). I saw a tiny bit of it but I wish I've seen more because this is about measure score at provider level.

So, we did a lot of testing on patient level. So, that's one thing.

The other thing I'm a little bit concerned about is so in the way that they construct (the score) for each patient is residual score. So, HS score minus risk adjusted score. So, when you get (where at) t a provider level, so two provider – one provider, OK? If someone improved a lot, someone not at all or even get worse.

And another provider – everyone improve a little and they (these two) provider all look the same. So, (instead of) – (you submit), well, your average then (inaudible) average really high or really low and you get a little. But there's so many different way to get there.

Sam Simon: Yes, I mean, this was – I didn't have a lot to say here on my from but I just didn't see any evidence that clinic or clinician level PRO-PM scores were validated. So, I agree with everything that my colleagues just said. This is Sam.

A part of me wonders too if this is something that NQF staff can pick up on in the future in the sense that this may not have even needed to be reviewed if it was determined that – that's not a judgment call necessarily, the information is either there or it's not. I didn't really – I did not see any evidence that there was this clinic or clinician level validation and this is the measure that maybe didn't – perhaps, there could have been some back and forth with the developer and they have this, they just didn't provide it or didn't know to provide it.

So, this is going – I guess my comments going back to NQF staff seeing if that's something that they can include and there – when they go through the measures.

Karen Johnson: Sam, we actually do a fairly in-depth, what we call completeness check really early on before you guys get these measures and we send things back to developers and say it looks like you could probably expand on this or you got to add this or whatnot.

To be honest with you, I don't know if we caught that particular thing or not. I'd have to go back and check and see if we did. With the timeframe being what it is, we – even though we did that, we weren't able to take the time to see if people actually did and provide the things that we advise them to provide.

So, for what it's worth, we did try and we …

Sam Simon: Yes. No. And I understand. There's a lot of measures.

Karen Johnson: Yes. Yes. We actually got pushback because we only give developers 48 hours and they didn't particularly care for the quick turnaround either. So, yes.

Zhenqiu Lin: Karen, I think if you look at page – testing (form) page, at the bottom of Page 41 and top of 42, there are tiny – two tables (somewhere), a little bit testing, but obviously, I think we wish to have more (inaudible) doesn't seem to sufficient.

David Nerenz: Well, and that's what I was looking as well. And this may help excuse Karen and staff. When I reviewed this, I looked at these gorgeous-looking beautiful graphs in 41 and 42 and looked at their texting that supports the validity and I thought, OK, this looks pretty good.

But now, as I look at it a second time, I don't think this is about validity. It basically says there are observed differences and performances and there's a few that are outside the range of constant intervals but that's a reliability concept. It's not a validity concept. We don't why these differences exist.

Zhenqiu Lin: I mean the Table (2D4), the (inaudible) (pro forma) (inaudible) level and then another one, (pro forma) (inaudible) level. You can be generous (inaudible) a little bit (taking into) the validity.

Male: Yes, but …

Zhenqiu Lin: … the graph is not about the validity, as part of that validity (inaudible).

Male: Yes. No, as I look at those now, the tables, they're essentially just a categorization of the figures above them, right, just how many below – fall above or below a certain level?

Zhenqiu Lin: Not exactly.

Male: Well, I mean, I know that they're capturing here in the tables the percent who achieved this certain degree of improvement. But again, fundamentally, the concept here, as I'm reading it now, is variation of performance. It's not about the reasons for the variation or the fact that the variation actually reflects quality of performance.

I mean, I think we're agreeing with each other. We just may be saying it in slightly different ways.

Zhenqiu Lin: Yes.

Karen Johnson: So, one thing – and I've been struggling to bring up the testing attachment and I'm still (trying to) bring up, but when we were looking at it, when we see nomenclature like known groups (construct) validity, sensitivity to change, those kinds of things, our automatic assumption is that that is at the score level. But when I was reading the text that went along with it, I wasn't so sure because it sounded like it was just kind of (grouped) by patients, not by clinician groups or clinicians.

So …

Male: Great.

Karen Johnson: So, from that – so, I think possibly, what we may have done is just look at that verbiage and say, they've got the right words there, they must have done it. And only in – and this is up to you guys. It kind of does seem like perhaps – I just it again, I'm sorry – perhaps they – that might not be aggregated at the clinician level.

And May is trying to find the section that I'm referring to. She's probably going to have a better luck than I am.

May Nacion: (Inaudible).

Karen Johnson: Yes. These are the same ones that I think Zhenqiu was talking about, Page 41, the bottom of Page 41 and then the question is if you look a little further down when they – either when they were first describing it or maybe later on, it sounded to me like maybe it's at the patient level.

So, and maybe I'm wrong. And if I am, I apologize.

Ashlie Wilbon: So, this is Ashlie. I wonder, just in the interest of time, we're about seven minutes before the end of the call. If you feel – if panel members feel that you're ready to cast your votes for validity and does – so, just to recap, it was one moderate, three insufficient, one undecided. It sounds like if there was not core level testing but you're satisfied with the patient level, I'm sorry, data

element level testing that they provided that insufficient might be the vote there or depending on your level satisfaction with what they provided or not are low.

So, Karen, I don't know if you have anything to add to that if that's accurate. Do you agree that's the accurate characterization of the options?

Karen Johnson: Sorry. Ashlie, did you say that score level was not (proof) needed or did we mishear you?

Ashlie Wilbon: No, not that it wasn't needed but that it did – it's the data element level – sorry, let me just (inaudible) a little bit. That – it says that it did not appear to be provided that the vote would be insufficient if they feel like it – all of the testing was not provided or adequate for them to make a decision.

Karen Johnson: (Inaudible). Yes.

David Nerenz: Yes, Dave here.

Karen Johnson: Go ahead.

David Nerenz: Let me try to make a positive suggestion on this one. I was the outlier going into this discussion. I'm no longer an outlier.

Karen Johnson: So, we'll go ahead just for our records and ask you to vote …

David Nerenz: Was that cryptic enough?

Karen Johnson: I think I got what you said. I'm still going to ask you guys to click that button and vote for it just so we have it here. And that …

David Nerenz: (Inaudible).

Karen Johnson: Yes. That gives (you) a chance to weigh in. Go ahead, David, I'm sorry.

David Nerenz: I was already jumping ahead. Since now you've put the composite question, are we going to be able to close this out without answering the composite question?

Karen Johnson: This one's ...

Joe Kunisch: I was able to.

Karen Johnson: … composite was it? Was this …

Ashlie Wilbon: (Inaudible).

Joe Kunisch: Yes. I just – I skipped that. I skipped the composite question. I was able to (close it)..

Karen Johnson: OK. Yes. You're looking at the surveys, actually. Yes. Skip the composite question.

So, yes. You all have already voted earlier. (And the call) on reliability now on validity and then hit submit.

Miranda Kuwahara: Yes. In order to give you one survey instead of 10 million surveys, we do have to get those other questions. And so, yes, you can just skip them and not (a reply).

Karen Johnson: OK. Well, congratulations, Ashlie. You got folks through a lot faster than I was able to, so that's great.

We're actually not (going to do that) because we still have two measures that we have to talk about. We have another two-hour call scheduled tomorrow afternoon, OK? So, this would still be fresh on our minds, that would be good.

Just one other thing, the five measures that did pass, again, we don't have to discuss this unless one of you would like to. So, (We'll) be thinking about that. We'll give you another chance tomorrow to prove.

There was one question that made some people a little uncomfortable, just kind of with that group of measures and it was the question about verifying that the score level testing for reliability was done with the risk-adjusted results.

Since that was a question across the board, I did email the developers to ask them that question and their response was we can confirm that risk-adjusted values were used to conduct the reliability and validity test. So, that – I don't think it would actually change your vote anyway but it might make you more comfortable about them.

Joe Kunisch: That's helpful. Thanks, Karen.

Karen Johnson: Sure.

Male: Karen, I do have one question about the remaining two measures. Is it the same situation for the remaining two? I did kind of skipped ahead and look where we have general consensus but there's just one person who didn't vote or we need to kind of get all the votes in or the others are actual split votes for the remaining two measures?

Karen Johnson: It's actually neither. The remaining two measures, both of them failed according to you guys. So, it wasn't a split, though. It was a complete no-go.

And what we would normally or potentially do is just send it back to the developer now, not insist that you guys discuss it. But we actually pull these measures because we'd like you to discuss a couple of that specific things. And we can get into those a little bit more on the call. But I think – I think …

Male: OK.

Karen Johnson: … that it was in your discussion guide but we used our prerogative to vote and not necessarily because we think you should or would change your measures or your votes but I think there's just a couple of things that for consistency and consensus, going forward and across the panel, we'd like to bring out and talk about it a little bit.

Male: OK. Thank you. That's helpful.

Karen Johnson: OK. We have two entire minutes left in our call. Is there anything anybody would like to mention, discuss, anything we could do differently on the call tomorrow to make it go smoother for you?

John Bott: This is John.

Sam Simon: So, I'm going to say …

John Bott: Go ahead.

Sam Simon: Sorry. Just very quickly, I was just going to say providing sort of the overall context, like what our charge is for each of these measures at the outset of the discussion, I think is help – going to be helpful.

(Now), you've just done that for tomorrow's talks. So, (certainly, you don't need) to do it necessarily but I think that that context or (table setting) is helpful.

Karen Johnson: OK. Anybody else?

John Bott: This is John Bott. I was just going to ask what was different with the other groups that they got through all four measures in one shot and – which I believe you said – and versus us. Were we unruly? Could we have done a (bit better behaved) or something or …

Karen Johnson: No. I really think, to be honest with you, these are harder measures. For the first one, one of the things that we kind of kept circling was the results were fair, what do you do with those? And then that NQF is willing to be more flexible but maybe we shouldn't be. That was a different (animal).

The functional status, I think, that we just discussed had its own kind of difficulties. So, these were just harder measures and to some extent, I think that's a little bit reflected in our tardiness in getting you the discussion guide.

We try to look at these in detail and write some stuff that was useful for you. And it just took a while to go through it.

John Bott: Yes. OK …

David Nerenz: I found the process very helpful. I found the discussion really interesting and helpful.

Karen Johnson: Good. I think that's been the case across the board. And I think it is helpful and there's absolutely no – if you guys don't mind your kind of the world and your subgroup colleagues knowing I voted low or I voted insufficient or whatever, if you don't mind telling that, then we don't mind you telling that.

Because at the end of the day, those votes aren't the ones necessarily that counts. It's the ones that as you guys are voting, that consensus vote as a subgroup is what's counting. So, that's why we want to give you plenty of opportunity to pull anything that maybe you did agree on. You could still pull it and discuss it and it's the consensus vote that we go through

And I think the other good thing that I'm really pulling out of your discussions and other discussions as well is I kind of have an – my agendas, if you will, for the next several monthly calls because we're running into things that I think we ran into a little bit in the last two cycles, but it's really hard to get our minds around because people were doing things individually and kind of in a – kind of their own (level).

And now, when we have more people kind of talking about it and coming to consensus about things, it makes it easier to spin off a conversation that I think we can make some real headway in the next few calls.

So, yes, you guys weren't unruly at all. All right. I'm going to let you go. I talked a little bit past 4 o'clock. So, apologies for that. We'll talk to you again tomorrow at what time?

Female: Two o'clock.

Karen Johnson: Two. OK. Thank you, (all). Thank you so much.

Male: Thanks.

Male: Thanks a lot. Bye.

Karen Johnson: Bye.

END