NATIONAL QUALITY FORUM

Moderator: Sheila Crawford
March 14, 2019
1:42 pm CT

Sam Simon: Hi, this is Sam Simon.

Deanna Hayes: And I would like to know …

(Miranda): Hi, folks. We are going to give one more minute and then we'll kick off the call. Hi, everyone, this is the NQF team. We know that some folks are still dialing in who are members of the panel, so we are going to start in just about two minutes. Thank you for your patience.

David Cella: Hello?

(Miranda): Hi there. Thank you all for joining today's Scientific Methods Panel Subgroup Number 2 Measure Evaluation Call. We will kick off the call with a roll call of subgroup members. First off, do we have David Cella on the line?

David Cella: Yes. Hi.

(Karen): Thanks for joining Dave. Joe Kunisch?

(Miranda): Hi, I'm Miranda.

Ashlie Wilbon: This is Ashlie. We should combine the roll call with the disclosure (unintelligible). So, maybe if you just want to start to just give an overview of the call and kind of discuss the webinar, I'm sorry of the conference call.

(Miranda): No problem.

David Cella: Is there a webinar by the way? Is there anything online that I did not get that?

Ashlie Wilbon: No, sorry, I misspoke. There is no webinar.

David Cella: Okay.

(Miranda): All right. So, a discussion guide was sent out to subgroup members Tuesday afternoon. That document will guide today's discussion and we'll follow the order presented on that document. Consensus was not reached for 3461 and NQF 40005, so both measures are slated for discussion today.

All other measures will not be discussed during today's call unless a subgroup member would like to pull a measure. If you choose not to discuss any additional measures, the decisions from your preliminary analysis will be made final. At the end of the call, we will pull the group to see if any subgroup members would like to pull a measure for discussion.

In that same e-mail containing your discussion guide, there was also survey linked to a SurveyMonkey. We would like to ask you to pull up that survey now and cast your votes on either reliability and/or validity after the conclusion of each measured discussion.

David Cella: And where's the link? Sorry. Where's the link?

(Miranda): It's included in the e-mail that went out Tuesday afternoon around 3 PM from the Scientific Methods Panel inbox.

David Cella: Okay.

(Miranda): And when it's time to vote we will prompt you to do so. We expect to review all measures during today's call but we do have a follow-up meeting on the book for Monday, March 18th, from noon to 2 PM to discuss any outstanding items and if we do address all items today then we can go ahead and cancel that meeting.

And finally, I just want to note that this is a public call and developer representative are on the line to answer questions from staff or panel members. However, there is no opportunity for a public comment and for record keeping purposes we asked that each subgroup member or developer to please state their names before speaking. And so, Ashlie I'll pass it to you for the live roll call.

Ashlie Wilbon: Okay. Thank you, Miranda. Hi, everyone. This is Ashlie Wilbon, MS the director of NQF. I will be leading the disclosure measure as well as the discussion for 3461 which is one of the measures that does not reach consensus.

So, we do Disclosure of Interest -- oral disclosure of interest for all of our measure evaluation calls so there was me as the kind of we get through the preamble for this but essentially you received Disclosure of Interest Form when you are named to the panel and then we send you manuscript of that

disclosure of interest form of each measure recycle and as you got your relationships and the measures that you reviewed or any related orthopedic measures that we have identified through the process.

So, between these two forms, if you have a lot of questions about your activities or degree of involvement with the measures and so in the interest of transparency today, we are going to ask that you orally disclose any information you provided that you believe may be relevant to the panel's work today specifically the measures that you reviewed as a part of this review cycle.

You don't need to name specific things from your resume unless they are directly related to this work. This would include plans, research, consulting and both paid and unpaid development activities related to the work of this group.

So, on your agenda you can see the list of measures that were reviewed by this subgroup, so 3461 which is a functional status change for patients with neck impairment, and CAHP clinician group survey, consumer assessment of health providers and system (HCAHP survey), the HCAHP hemodialysis survey, CAHP from home health care survey, and then several functional change measures that are around change in scores for various aspects of functional status, self-care mobility and then also a series of measures around in-patient rehab facility functional outcomes as well, as well as the collaborate measure for shared decision making and the Child Hospital Consumer Assessment of Healthcare Providers and Systems.

So, again based on that, as we call your name if you could just let us know that you are here and then also whether or not you have any interested exposed. Also, to point out that just because you disclose does not mean that

you have a conflict. This is all -- again, this is in the interest of transparency. So, I will call your name and then please let us know again your name and any interest exposed. David Cella?

David Cella: Hi. That's me. I don't know if this is a conflict but just for the sake of caution I am using the collaborate measure in the study and Glyn Elwyn is a consultant on that study.

Ashlie Wilbon: Okay. Is it paid or any involvement with the developer …

David Cella: We are paying him as a consultant on a study that we are doing here. It is funded by a foundation, Peterson Foundation in New York.

Ashlie Wilbon: Okay. But you did not participate in the actual development of collaborate…correct?

David Cella: No, no. I'm not involved in the development of a measure at all but we are using it.

Ashlie Wilbon: Okay, okay. Thank you.

David Cella: I don't know if it's a conflict but just you know. And by the way, could somebody resend the Tuesday e-mail? I apologize but I have looked in my deleted docs and my inbox and I don't see any mail that came Tuesday. Could you resend it so I can get to the SurveyMonkey?

Deanna Hayes: We will send it now.

David Cella: Thank you.

Ashlie Wilbon: Okay, thank you. Thank you David. Marybeth? Are you there?

Marybeth Farquhar: Yes. I'm here. I don't have any conflicts of interest in the last five years. The only thing I need to disclose is that I was originally working with the CAHP Team back in 2005. So, it's quite a while.

Ashlie Wilbon: Okay, thank you very much. Joe Kunisch?

Joseph Kunisch: Yes. Hi, this is Joe. I do not have any interest to disclose and I also didn't - or I can't find the survey e-mail.

Ashlie Wilbon: Okay. We'll make sure to resend that to everyone. (Unintelligible)

Man: Yes I am here. I had no open measure.

Ashlie Wilbon: Thank you. David Behrens?

(David Behrens): I'm here. No disclosures.

Ashlie Wilbon: Thank you. Sam Simon?

Sam Simon: Hi. I'm here. Nothing to disclose.

Ashlie Wilbon: Thank you. So, again due to the interest of transparency, if you feel at any point in time that there is any bias during the review please feel free to reach out to NQF staff and send us a message and we will go ahead and jump in to review.

I did want to find out - again, as Miranda mentioned that this is the first time with this new setup of doing the subgroup call that we are also engaging

developers who will be -- who can be invited to ask-- I'm sorry respond to questions based on the panel's questions or need to understand the measures, so keep that in mind.

I do just want to make sure that the developer for 3461 is on the line.

Deanna Hayes: Hi. This is Deanna Hayes from FOTO. I'm here. I need to let you know that our lead science who found the project was not able to attend the meeting today so I'll do my best to answer questions that are passed by me. I may have to take some questions and head back with you when Daniel is available.

Ashlie Wilbon: Okay. Thank you. I do also just want to now reach kind of disclosure to know who are we going to call, to ask the panel if there is anyone who would like to pull any of the measures in addition to what already is slated for discussion?

(Karen): Hey, Ashlie. This is Karen. We will ask them that question again at the end after we have the conversation about CAHP.

Ashlie Wilbon: Okay.

(Karen): For CAHP.

Ashlie Wilbon: Okay. So, you have another opportunity. So, okay that sounds great. So with that let's go ahead and jump into 3461. So, I'll start out with an overview and then of actual measure itself and then highlight some of the voting results that came back then we'll dive into some of the specific discussion items that we identified based on the individual review that was submitted by methods panel members.

I just want to point out this measure probably was familiar to some of you. If you reviewed this measure last cycle it was submitted in the fall cycle of 2018 and there were issues identified by the subgroup at that time that resulted in the measure of not passing reliability and validity.

So, for those of you that are able to access the discussion guide it is segmented out by the concerns that were identified in the previous review as well as additional concerns that were identified for this review.

So, keeping that in mind, this is to refresh everyone on this measure if the patient reported outcome performance measure, PROPM, consisting of risk-adjusted change, its functional status for patients or seniors of Asian older with neck impairment, the change in functional status is discussed using the Neck Assessment Performance Measure and the measure is adjusted to patient's characteristics known to be associated with functional status outcomes and it is risk adjusted and it is classified for individual clinician and clinician group practice level.

Based on the level of analysis that is specified for the threshold for the clinician level is 20 plus patients per year and the clinic level for threshold is for small clinics 10 plus patients per year per clinician, for larger clinics 40 plus patients per year per clinician.

So, again it is risk adjusted and it uses a statistical model. In terms of the reliability evaluation, so we have one high score, two moderate, no low and two insufficient which puts the rating for reliability into this consensus not reached.

So, again just a little bit of review of what was done for reliability testing. The testing that they conducted was filled at this four level and mandated

element level and some of the results here for high level for the testing that was completed is here in the discussion guide.

At the patient level, stated element testing, they used the (test) via internal consistency using Cronbach's alpha and Item Response Theory and Cronbach's alpha was 0.98 and the IRT base person reliability score was 0.96.

At the clinic level, they looked at the average reliability of clinic meeting FOTO threshold of the number of patients per clinic and that score was 0.79 and then at the clinician level the average reliability for clinicians with 10 or more was 0.64 and for 20 or more was 0.76.

So, I think, so again, in addition to the reliability of individual score they also looked at the standard error of measurement and analysis of the minimal detectable improvement. In terms of the high level some of the concerns that came back of this reviewed cycle again there were still some I think concerns or lack of understanding in terms of the signal-to-noise analysis and understanding which level of analysis that specific results were applicable to.

So, with that I think we'll stick to reliability and kind of jump down to the items to be discussed. Again, this measure does not pass some reliability or validity and some of the items that were identified and shared back with the developer in the prior cycle that were specific to reliability of the specifications were around the lack of specificity of the definition and some of the key data elements in the measure around copy responses, the episode definition for how discharge is determined and captured around the clarity of the numerical statement on the descriptor and whether or not that was being calculated as a change for and then clarifying the definition around neck impairment.

So, our first action item is to just kind of check the pulse of the committee to see based on their review whether or not you feel the developer adequately addressed these issues from the prior cycle and their follow-up submission for this cycle. And we can also pause to see if there are any kind of any high level concerns about the measure or questions for the developer before we kind of delve into this discussion.

David Cella: This is Dave Cella. I don't know. I'm one of the two insufficient graders and if it's possible that the committee can help me with my problem maybe I just missed it, but I didn't feel I was able to determine the signal-to-noise ratio. I was fine with the data element reliability but in terms of the measure I couldn't find information to determine signal-to-noise ratio. So, it's possible I missed it. Someone can point me to it and then I would change my vote.

Ashlie Wilbon: Are there other panel members who might be able to point to him. I'm going to try to find it myself and in the meantime if the developer on the phone if you are able to point in the submission clicker then maybe we can get to it on where that information could be found that would be helpful.

(David Behrens): Dave Behrens here. On Page 24 the measure testing form is at least where I thought I saw it additionally it states insufficient but at least that is where the heading exist and where the data is here as far as I can tell.

Ashlie Wilbon: Yes, I agree. Was that you Deanna?

Deanna Hayes: Yes. I was going to say the same thing, signal-to-noise used the Adams method that was illustrated on and described starting on the bottom of Page 24.

(Karen): This is Karen from NQF. It looks like I am seeing on a table on Page 27 the test results. Hopefully I'm not looking at a different version than you guys have.

Ashlie Wilbon: Table 2 2A2, correct Karen?

(Karen): Correct. No.

Ashlie Wilbon: Okay.

(Karen): 2A2 III …

Ashlie Wilbon: Yes, Okay.

(David Behrens): That's right, method on 24, findings on 27.

David Cella: I'm pulling up the document.

Ashlie Wilbon: So, it looks like I'll just kind of recite what I'm seeing as others are looking for it. At the clinic level it looks like the average reliability score was 0.79. At the clinician level for 20 less patients per clinician the average was 0.76.

Sam Simon: This is Sam Simon. So, I was the other insufficient and what I couldn't find in this was an understanding of whether or not these results were derived using the Risk Adjusted Score for this measure and I get it might be something that I missed but I didn't see it.

Ashlie Wilbon: Is that something that you can clarify Deanna or point us to I will look as well but there is someone in the submission that you can direct us to or respond to that question?

Deanna Hayes: It was and I'm looking to see where that was stated as well. I do apologize. Daniel would have been here but he had a very extraordinary situation. He is the one who wrote this.

Man: I just echo what Simon just said. I actually wrote about it. The reliability change we base on HRI and model output so I was wondering that the model specification that relates to what Sam states whether the 11-risk (unintelligible) was included when they (unintelligible) model because if the model is different from the measure then that will be a concern.

(David Behrens): Yes, I guess - Dave Behrens here again. I would defer to the others in the group who understands the Greek symbols better than I do. But in looking back on 24 and in the third paragraph from the bottom, we are talking about hierarchical linear models, there's a ton of vertical functional statuses that charge adjusted fall variables used for risk adjustment. That's the basis on which I thought the statistics and the results then were based on risk-adjusted.

David Cella: That's great Dave. Thank you.

Deanna Hayes: The other page that I point out is on Page 27 the first paragraph under Results and Reliability of Providers, a clinician and clinic level. He states in the middle of that first paragraph that is completed the PROPM though it perhaps should be described better but you know PROPM means that risk adjustment was used, otherwise, we would have said PROM.

We would be happy to clarify when Daniel is back in the land of the living.

Ashlie Wilbon: Does that address the concerns? I think it was David C. who queued about that particular item or whether or not risk adjustment was used or do you think that you need more explanation?

David Cella: You are not calling on me. This is Dave Cella right here. I think it's (David Behrens).

Ashlie Wilbon: Yes. I think it was (David Behrens) to have that concern or maybe I might have mixed up. Sorry.

(David Behrens): No sorry. It was Dave Cella in this case.

Ashlie Wilbon: Dave Cella. Sorry.

David Cella: What? I didn't ask about risk adjustment. Yes, I asked about the signal-to-noise.

(David Behrens): Okay, then I'm sorry. It was not David Behrens either. I was the one that spoke and said that here is in the document I think it describes it but I didn't have any concern.

Sam Simon: Yes and this is Sam Simon.

David Cella: It was Sam Simon. It was Sam Simon.

Ashlie Wilbon: It was Sam Simon. Sorry.

David Cella: Yes.

Ashlie Wilbon: I wasn't writing down. I apologize, sorry.

David Cella: Sam Simon did it.

Sam Simon: I did it.

David Cella: Yes, thank you. That was helpful. So, this is Dave Cella about the signal-to-noise. Thank you for pointing me to it. I don't know if this is where you want to go just yet but I am uncomfortable changing my ratings and perhaps even to high but certainly to moderate and I am just wondering should we discuss these 10 patients versus 20 patients issue?

Ashlie Wilbon: Sure.

David Cella: I mean, where are we with that?

Ashlie Wilbon: That certainly is within the realm of the reliability discussion, so…

David Cella: It seems like the below 20 was problematic. I just want to make sure that we are clear on that.

(David Behrens): Yes, agreed.

David Cella: Yes and that was something I raised in my review as well. I don't know you know how NQF use it as that is something has to be in the specs that the measure is not reported for less than 20 in this case, but that was something that was concerning.

Ashlie Wilbon: I made a clearance on that but I do see how they highlighted that particular row for 20 plus range so I am wondering if that was the intention. I don't know Deanna if you could clarify that?

Deanna Hayes: I'm looking. He did that to emphasize that that is the acceptable level of reliability.

Ashlie Wilbon: Okay.

Deanna Hayes: For 20 for clinician.

Ashlie Wilbon: Yes.

(Karen): I'm actually looking at the specifications and in F15 which is maybe Page 6 of the specifications form but anyway turn to F15 sampling. It does look to me like they are saying at the clinician level for clinicians with 10 or more or 20 or more patients per year that are reporting the reliability, I thought somewhere that I actually saw in the specifications the 20 or more but my understanding and Deanna correct it if we are wrong, but is that what you are suggesting and putting forward and if it is and if it is not still spelled out in the meet we will make sure in your specifications that it is spelled out?

Deanna Hayes: Thanks Karen. Yes that is what we are suggesting.

(Karen): Okay, thank you.

(David Behrens): There is just one observation but this doesn't I don't change our decision that in going down the steps we are basically saying that this 0.7 has the status of a magic number essentially the sort of like it is high although what's underneath it is continuous for saying okay to 0.71 you're good, if you're 0.69 you're bad, the 0.64 levels for 10 plus wasn't awful. They just fall below this 0.7.

I don't mind to specify that it should be 20 or more, I'm okay with that, but just as we take a head down the road in our work particularly in this people we are talking about on reliability for those involved with that this is one of the things that if we have to get into more detail, is there a specific rule of thumb that we are still counting it in that we are going to say that if it is 0.7 passes and below 0.7 fails?

Deanna Hayes: This is Deanna. I just want to say we've had some more discussions and I would like mine …

David Cella: Yes, I mean, this is Dave Cella. That's a good point Dave. I think a lot of the discussion is probably off topic for a specific submission like this and when we should definitely have it, but related to this submission and I guess I'm wondering it's kind of in the sense because if we say well only people with 20 patients or more then what happens to those who have fewer than 20 patients and we may be denying the opportunity to evaluate their quality just because of the magic number issue when actually there is going to be more error in the estimate but not so much so that they shouldn't be compared and held to the standard.

In other words, it's kind of throwing an advantage to providers who have fewer than 20 patients, if I'm thinking about this right? To get you off the hook, is that the case where this would play out?

(David Behrens): No. Let me check in the sanction that that the people who belonged to and managed and used this database and registry essentially is free to do whatever they want and they've done very careful analysis here I think and if they want to show data and use data for people of 10 I think that's up to them.

I think the NQF endorsement though suggests that in essence they are going to use it in meds or something else is going to use it and they choose to follow NQF endorsement. At that point, they would specify a minimum sample size. You know, it is always a tough call. You are basically saying if you have 20 cases then you are good to go, if you have 19 cases then you're not good to go.

This is true with any of these arbitrary cutoffs but to move things we are recognizing that in the real world there has to be such cutoffs. At least here I'm uncomfortable calling at 20.

David Cella: You said you are comfortable or uncomfortable?

(David Behrens): I am. I have discomfort and slightly elevated blood pressure but we got to decide. I'm uncomfortable.

David Cella: Okay.

Ashlie Wilbon: Okay. So, I think that was helpful and certainly something that we are keeping track of these issues for queuing up and future discussion outside of the measure evaluation call. So, we will certainly keep that on the list.

So, in terms of reliability, I just want to kind of bring us back a bit. It sounds like the concerns about reliability have been resolved.

I also just wanted to point out in terms of a previous evaluation, none of the same issues were raised that were identified in the prior evaluation from last cycle so I just want to kind of put it out there again to see if anyone has any other concerns, results identified in the last cycle assuming that those concerns were not raised again in this cycle that the subgroup is okay with the

developer having addressed those concerns and then if all minds are settled we can go to the reliability vote. So, I just want to kind of put it out there again and see if there is any lingering concern about reliability before we call vote?

David Cella: None here, Dave C.

Sam Simon: Yes. I'm good too. This is Sam Simon.

Ashlie Wilbon: Okay. We only heard from a couple of folks but I am assuming since we didn't hear any other major concerns we will go ahead and direct you guys to click your vote for reliability for this measure to the SurveyMonkey and we will move on to the discussion for validity.

So, again with the list of validity, high level summaries, there were low high votes, 3 moderate votes, 1 low, and 1 insufficient. This again put us in the consensus not reached zone, if you will.

Again, as a reminder, as an intermittent base measure both data element and score level testing are required which the developer submitted. They performed several types of data element validity testing as well as measure score validity testing that are listed here, structural validity, contract validity, and so forth.

In terms of the concerns that were identified more specifically from the prior cycle, again the subgroup members were seeking clarity on whether testing was done at the score level and wanting some clarity around this submission and how that was illustrated or demonstrated.

The consideration for implications for discerning meaningful difference at the clinician level, if the measure was used in accountability purposes assuming

there was some concern that the result seemed to indicate that there was only significant differences in performance at upper and lower end of the distribution.

Again, concerns with the application of risk adjustment and introduction of possible structural bias with the testing of the risk adjustment change for the lack of differentiation at the clinic versus clinician level with probability and again concerns with the ability of the measure to account for how the change scores accomplished based that some providers could achieve change scores based on different types of performance.

So, I guess let me pause there and see if there were any similar concerns that were raised or considered for this review cycle as a point to see whether the developer adequately address these concerns based on a previous prior review. Again, none of these seem to be identified in this cycle so I just want to raise it in case folks had some questions or concerns to reiterate.

Okay, hearing none, some of the concerns that were submitted for the current review cycle again were concerns around the use of the change score versus some external criteria to perform score validity testing. Again, the demonstration of the contract validity was considered to be weak and that there was a compelling evidence of an association between the score and another independent measure or concept of quality care.

There is a note here that NQF is not directive on what type of measures are appropriate for use in score level validity analysis, so there is not a lot of direction there so we will leave this to the developer to determine the approach that they would like to take to adequately demonstrate validity.

There were some conflicting perspectives from some of the methods panel members based on the PAs on the acceptability of the amount of variations in the performance that was described. Some felt that the variation was adequate; others believe it was minimal given the large testing sample and the final concern was around the lack of clarity of which testing results demonstrate clinician and score level validity.

So, I'll pause there and see if any of those concerns were once that you raise, if you'd like to share your thoughts with the group.

David Cella: This is Dave C. again. I was trying to wait for somebody else to chime in but I guess it's me again. I would be the insufficient vote and partly that's because of a notion that, you know, if you have insufficient reliability, concerns about sufficient reliability they are going to carry forward to concerns about validity. So, that was coloring my review. The score, the data element level of validity I thought was mixed, to be honest.

Bottom line, I actually thought it was okay. It seems like this is measuring more or less but it is intended to be measured if we use that definition of validity, but you know it wasn't really a unidimensional model story where these items closer together sufficiently in a way where you can be certain that you know what you're measuring.

At the end of the day though it does appear to perform okay so I would have -- that wasn't the only concern I had. The other concern was that you mentioned to me at least clinician level was not clear to me that the measure carried forward to the clinician. So, I just had these concerns so I scored it insufficient. I suppose I could have scored it as low, but I said insufficient just to carry forward my previous reliability vote. Anyway, that's where I am with it. Maybe, others can help me with it.

(David Behrens): I can jump in here. Actually, I think I'm one of those who gave it moderate, but I'll just walk through a little bit of the thinking on validity. I did have concerns about a number of things. There's a lot of information that was presented here but, you know, a lot of the things under the group validity basically just said the scores differ when you go clinician by clinician. And you can group into three, you can group in the ten but you're basically just showing that, you know, that there's variation. So I didn't see something about like truly independent construct validity.

You know, there is a section called non-group solidity here. And I thought that was going to be it that basically what I see under that I would call risk adjustment. It basically says people who enter with certain characteristics like being younger or male have higher or lower scores. Now I guess you can work backwards and say, you know, those are known groups meaning there's some independent knowledge that suggests that these groups should do better and in fact they do so there we are. But it wasn't known groups in the sense of an independent outcome like return to work or not or something like that.

The reason I think I ended up monitoring it was that -- and I'm happy to be corrected on the rule here -- it seems like in a number of measures (Kevin) established face validity is sufficient to at least get you into moderate. And I did see the work here that was done with panel of physical therapists. I'm not sure the term face validity was used but that's what I saw. So I'm happy to be corrected here on the issue of whether a formal run through of the face validity process is enough to at least get you to moderate.

David Cella: All right, thank you. I think you've articulated all the things that were on my mind and that were causing me concern. And, you know, I - this is (Dave) again, David Cella. I put those very concerns into the category of insufficient

because I didn't see a clear known groups. But I too am willing to accept face validity as one of these you know it when you see it this does seem reasonable sufficient criteria. So that I think is also another general question that would be helpful to hear from (Karen) or someone in terms where NQF is with this?

(Karen): So this is (Karen) from NQF. Face validity is allowed for new measures at the score level. So basically the idea would be asking folks whether or not they believe that the results would be able to differentiate good from poor quality. So that's the kind of thing we're looking for. I am not as familiar with this submission but I believe the face validity that they described was more along the lines of the validity of the items which we would consider more a data element thing.

So if I am correct on that if that's what they're putting forward that does not conform to our requirements first of all four face validity. And then secondly because this is instrument based we actually do require empirical testing of the measure score. So in other words some kind of construct validation that you would accept is what we're looking for.

Ashlie Wilbon: And (Karen) I did just want to clarify your question you had was on Page 32 the face validity kind of demonstration they describe in 2B1I is around the data element.

David Cella: Yes.

(Karen): Yes.

David Cella: Yes, so, you know, and I guess we should see if there are others, you know, that would like to chime in here. This is David Cella again.

(Ditu): So this is (Ditu). I rate it as moderate but I do have concern with performance score validity. There is two way to correlate with the score. One is based on performance level the three level and then the performance decile right? But they are all based on the same data I think if I'm not mistaken. So it's sort of, it's not external criteria. Ideally you want something like (unintelligible) and a non-group criteria.

Ashlie Wilbon: Are there any others that would like to - who voted either I think (Dave) mentioned he voted insufficient but I'm not sure if anyone has spoken already was the person who voted low or someone else who voted moderate that want to share their thoughts about that?

(Dave Nance): (Dave Nance) here. I'm wondering if we could just ask the measure steward a question. There's a fair amount of material here coming from a body of work labeled Functional Staging Change. I didn't see a lot in here that I thought represented some independent validity confirmation but I'm happy to be instructed if I am missing the point there. Is the functional staging of anything other than a classification of the measure score itself?

Deanna Hayes: This is Deanna. The functional staging is a clinical interpretation parameter to interpret the scores. The external marker if you will we did have an independent panel of physical therapists not involved who assisted in validating and the functional staging model. The other to go backwards a bit when you were asking for external markers would it help if I pointed out that - and I'm losing the table now -- there was a table about the MCII the Minimal Clinically Important Improvement. The MCII is based on the global rating of change.

David Cella: That's the clinician's global rating or a patient's global rating?

Deanna Hayes: The patient.

David Cella: So the patients asked are you better, worse or the same and then this…

Deanna Hayes: Yes, yes.

David Cella: ...and then there's further reference to that rating?

Deanna Hayes: It's - yes it's the 15 point local rating.

David Cella: For the, yes, yes, yes.

Deanna Hayes: Does that help at all with the external issue?

(Dave Nance): Yes, okay so this may help. So if this is pointing us to the results table on the bottom of Page 37 I guess if what's embedded in here I just didn't see the labeling quite clearly. If there's some sort of external like global assessment of change and what we see in this table is that there's a strong relationship between the average improvement of the photo score and from separate global assessment of change. I think I might have except that. I just somehow didn't pick that up as being an independent anchor for validity testing at the score level.

Deanna Hayes: My apologies. I probably confused but the MCII with the global rating of change refers to table - a different table not the functional staging.

(Dave Nance): Oh I'm sorry, okay, okay.

Deanna Hayes: I did two things at once trying to let you all talk. The functional staging so I don't - I'll tell you what it is and you can decide if you feel that it's

independent enough. But the stages one through five if you look near the bottom of Page 36 with the operational definitions were conceived by physical therapists with experience treating patients with neck problems. So you may or may not…

David Cella: But those ratings are made on the same patients but without knowledge of the patient's responses to the questions.

Deanna Hayes: The knowledge of the patient's responses…

David Cella: So the functional staging that the physical therapist does are they using the patient responses to make those - that rating or are they giving a rating that's based upon their examination of the patient independent of what the…

Deanna Hayes: I think the answer is both. May I backup and tell you how these were developed a bit more?

David Cella: Yes.

Deanna Hayes: The - so if you look at table - Figure 2B1II the title is at the bottom of Page 35 and the actual table starts at…

David Cella: Right.

Deanna Hayes: …the top of 36. So this was the results that I'm sure David Cella you're familiar with.

David Cella: Yes, yes. Yes, no that's a pretty figure yes.

Deanna Hayes: Isn't it pretty?

David Cella: Yes.

Deanna Hayes: So there were - the point was to clinically classify and characterize. So there were clinical descriptions by physical therapists. So for example I'm a physical therapist. If I look at the far right side of the column I see hey they can do all this high level stuff. And you know in physical therapy terms if you go back to the operational definitions Stage V as a physical therapist or other healthcare provider who works a lot with patients with neck problems that's a big - that's clinically meaningful that I can say this patient can perform vigorous work, sports and so on.

So there are clinically meaningful categories that were developed in an attempt to generally describe the model. The purpose was clinical interpretation. And then the validation involved a blinded - a panel of external physical therapists who were blinded to the actual score that the patient got that they were given.

David Cella: That's - yes that's what would be - yes so you used the ordering of these functional capabilities and data that's available from the ratings of patients to create these five stages. And so, so far it's not really external it's still kind a circular but then if you take - if you then take these five stages - and I'm just trying to kind of go with you here and get this out of the circularity you see what I mean? I mean there's…

Deanna Hayes: I do.

David Cella: ...it's a nice ordering of functions. And that ordering of functions is used to create the five stages. But if that - if you could then show that - and this is where the global rating comes in. I mean that's actually perhaps more

independent because if patients agree that they're better and they've moved from one stage to another then that's a confirmation of the moving from one stage to another.

Deanna Hayes: You know, as you all have pointed out it's a busy document. And based on the feedback we received last fall which we were very grateful for. And (Daniel) worked very hard to address every point so I'm pleased to hear that some of the same comments did not remain. However one of the things that he did was removed some of the detail. And one of the things he removed was a Sankey diagram and a little write up that illustrated empirically how patients - how many patients move from one stage to the other. So we removed that because we thought it was causing problems and it was too much information. But I have it at my fingertips and we'd be happy to put it back in. It's actually a really interesting diagram.

David Cella: Did those patients that moved up to a higher stage say they were better and did the patient that moved down say they were worse?

Deanna Hayes: So I don't think we did that.

David Cella: Because that - if you do that then we would have it. Then I think you would have it if you could show that. Do you see what I mean? So you take people that move - that get better on the staging and then if they independently say they're better and compare - and people that get worse in the staging independently say they're worse than you'd have - that would be strong validity evidence.

(Dave Nance): And send the amendments to that. You'd be doing that at the clinic or group level right not at the individual patient level?

Deanna Hayes: We tended to do these things at all three levels. It sounds like what you're asking for…

David Cella: Yes.

Deanna Hayes: ...is the provider level.

(Dave Nance): Provider level right and just clarify…

David Cella: Yes…

(Dave Nance): ...what it says in here.

David Cella: ...I would do all of them too. Yes, I think that - so how - so I don't know if others want to comment on this but so whether I - so I would be the insufficient person. And if these data were available and submitted I, you know, and they were positive then I would move to moderate or high. If they weren't available or didn't support the story I might be low. So I don't know how we - how I would proceed in terms of my vote here.

Ashlie Wilbon: This is Ashlie Wilbon. Let me just again ask Deanna is what Dave is describing is that something that you have and based on the analysis that you guys have done is that some - is he describing something that maybe you guys had done but it's just not in the submission?

Deanna Hayes: I don't think we've done those analyses. I don't think it would take long to do because we already have all of that.

Ashlie Wilbon: Okay.

Deanna Hayes: So we're happy to do it just - but I hate to delay if this happens to be the only thing and it might cause a delay in endorsement. I also do want to call your attention back to the separate validity testing that involved the Minimal Clinically Important Improvement which had that external anchor of the global rating of change. And I just want to make sure that in a pinch you wouldn't consider that…

David Cella: Yes.

Deanna Hayes: …sufficient.

David Cella: It's like we're in the market.

(Karen): Deanna, this is (Karen) from NQF. Can you tell me again just so I know for my notes what you're referring to? I've lost it.

Deanna Hayes: It's…

(Karen): Do you know what page?

Deanna Hayes: Yes, and I thought there was a table and I can't find the table. But if you go to bottom of Page 31 it should be one clinically important improvement that describes that. So the point is that MCII takes into account the patient saying that they were at least somewhat better. And then we take those values and we're able to show how many of the - how much of the scores aggregated, exceeded the MCII.

(Karen): That looks to me…

Deanna Hayes: That's one way of showing…

(Karen): …at - is that at the patient level or is that rolled up at the clinic and clinician level?

Deanna Hayes: If you go down to middle of Page 32 clinician and clinic performance score level it describes using the MCII and then goes into additional demonstration guide it is showing.

David Cella: I'm having a hard time lining up my page numbers with yours. You're on the testing form from January 6th?

(Karen): Yes, mine says January 3. That's probably when we submitted it.

David Cella: It could just be the way mine…

(Karen): So it might be best if I had the chance…

David Cella: What's the heading number? Is it 2B1.3 or 2B1…

Deanna Hayes: It starts 2B1III.

David Cella: Oh okay, so on mine for some reason that's back down on Page 36. Okay

Deanna Hayes: Let me get to where you are then.

David Cella: Well I'm at 2B1III yes.

Deanna Hayes: Okay.

(Dave Nance): I don't want to, (Dave Nance) here. I don't want to belabor this but I think it's actually pretty important when we come down to the finish line here. In the figure for me it's the bottom of Page 42 this is 2B1VV. I wonder if we could be very clear if I look at the bar on the right for example I'm curious what exactly is the y-axis here? Does this mean that 75.5% of those patients have independently said that they got better or is that 75% achieved a numeric improvement on the photo scale that it - that corresponds to the MCII like ten points or five points something like that. To me that difference is crucial.

Deanna Hayes: Yes, let me take a run at that. So the - if you - that's a great table but if you go back up to Page 41 there's a little table.

David Cella: Can you go by the letters because I think they are different page numbers.

Deanna Hayes: Yes I'm sorry, Table 2B1V-A.

David Cella: Okay.

Deanna Hayes: So you can see that the highest performers their patients exceeded the MCII that 73% whereas the lower performance only about 42% exceeded MCII. Keep in mind that MCII is kind of is by definition a minimal level of improvement. So, you know, patients might or might not achieve, you know, by the time they complete their episode of care if we're using the PROM a patient may have a better score.

So the MCII might be a score improvement of say eight points but that was their minimal their perception of minimally important improvement yet they went on to make more improvement than that. So MCII is minimal. So the low performers only 42% met or exceeded the MCII versus average versus high.

(Dave Nance): All right, but just to clarify were talking about enabling improvement on the photo scale right not on some other thing?

Deanna Hayes: An eight point improvement on the photo neck measure scale based on the global rating of change where the patient - so the global rating of change would be we take the threshold of the global rating of change score of plus three or better. And you start by finding that on the neck measure. A global rating of change for this patient let's say they have at least a three or better corresponds to at least an eight point improvement on the photo scale and I'm making up eight but.

(Karen): So (Dave Nance) going back to your question do you feel like you have an answer to the question that you just asked?

(Dave Nance): Well I guess I do. And I guess I align myself with David Cella. I'd really like to see at the clinic or provider level, you know, pretty simple correlation or, you know, some kind of comparative analysis between two things the average photo score improvement which has been used here to identify deciles, alone, high performers everything else you've got that and then separately the percent of patients who report a global change improvement greater than three. So and I presume this can be done pretty quickly but that is not I think what I'm seeing on Page 42 for example.

David Cella: Is that Page 39 though sorry 2B1IV- V? We're looking at decile of I presume again sort of if these axes were labeled better it would make it clearer but it looks like that decile of photo score. And then what you're plotting on the y-axis is the percentage of patients who met the MCII?

(Dave Nance): Yes.

Deanna Hayes: Yes.

(Dave Nance): Except the MCII is just defined by a unit of change in the photo score. So it's photo score on the X and photo score on the Y.

Sam Stolpe: Right, right

((Crosstalk))

Sam Stolpe: Yes, that's the circular nature right that's what we're trying to get away from.

(Dave Nance): Yes.

Woman: Yes.

(Dave Nance): Yes, I want to see something else on the line.

Sam Stolpe: Right, right yes. And...

(Dave Nance): Yes.

Sam Stolpe: Yes, and that - okay that is helpful. I wanted to make sure I didn't misinterpret. I mean that - I was sort of this is Sam. I did rate it is low. And that was some of my concern is I did find some of this circular as well and so share the concern that (Dave) pretty clearly described, both (Dave)'s.

(Karen): Well we could easily do what you're asking for. Our approach MCII takes into account the global rating of change in a more maybe more complicated way.

David Cella: So this is (Dave) - yes this is David Cella. I have the impression - so we're obviously a little bit - some of us are a little bit uncomfortable. By the same token I have to say I have the impression that you probably have what we need it's just not quite put together that way. I mean it be nice to see the global rating of change, you know, really parsed out from the score, you know, the MCII definition, maybe on the y-axis as an example.

But it seems to me like we're in this -- and (Karen) you can help again on this maybe - we're in this sort of period of review. It's relatively early in that period because this is our first call if we're going to have another one and there is opportunity I think for you to provide more clarification on, you know, getting - helping us get out of the circularity concern. And that would enable this to, you know, to avoid another round of revision prior to going to the full review committee.

Deanna Hayes: You're right (Dave). It's easily done.

(Karen): But unfortunately we can't do it that way. We can't allow kind of additional stuff at the middle even though I know we still have another call. For one thing, you know, we're not allowing it and haven't allowed it for other developers.

David Cella: Oh okay sorry I didn't - I should have known better. I didn't mean…

(Karen): That's okay I mean maybe you will…

David Cella: Yes

(Karen): …figure out eventually how to be able to do this because it is often iterative but we won't be able to do that this time.

David Cella: Well then, you know, there is information in front of us. I suppose we can vote with what we have understanding that, you know, sometimes our votes are, you know, they have whatever degree of leap of faith one might put into it…

(Karen): Right.

David Cella: …or not.

(Karen): And I think I would take, you know, see what's there. See if you feel like that it is doing the kind of contra validation that you would expect at the clinician and clinic level. If not then I would say vote insufficient. And I think Deanna, you know, has heard what you're saying. And I think the only thing that I want to make sure that I'm clear on just so everybody is clear the suggestion is to potentially depending on what you rate things as the suggestion is to compare the performance rate at the clinic and clinician level to this global rate of change value. But I was a little unclear as to whether the global rate of change is actually external of the photo or is it just using the same photo data. So my understanding and correct me if I'm wrong is that you're looking for something external that you can compare not just something derived from the photo measure itself.

David Cella: Right. But my - just add to that my - I understand that the global rating of change is separate from the patient's, you know, functional rating on the data element.

(Karen): Okay. As long as it is okay - I got okay I got a little lost in the conversation.

David Cella: Yes.

(Karen): So okay, all right.

David Cella: Yes, they're separate and that's where the value would be. And so I think our task is to decide if we think that, you know, the probability that that's what we would see, you know, had we the opportunity to see it. And apparently it was in part in the submission. I don't know if we can refer back to the last submission.

Sam Stolpe: Deanna, this is Sam Stolpe. Did you include a correlation or some analysis between that global rate of change and the performance instrument at the clinic level and at the incident level instrument level?

Deanna Hayes: Probably not the way you're asking. We, you know, the global rating of change is inherent in the MCII so I think that's why we weren't looking at it that way. We assumed that that's presumed. And I can see that we're, you know, next time we would call it out more specifically. But the MCII is based on the global rating of change. We can't get there without that separate global rating of change marker.

Ashlie Wilbon: Hi. This is Ashlie Wilbon. I just wonder in the interest of time if we could call a vote based on again reiterating what (Karen) said having you guys vote based on what's in front of your now. And, you know, Deanna hearing kind of specifically what the panel would like to see potential in the future to have you guys go ahead and vote and submit. And then we'll see what comes out and determine next steps from there.

Deanna Hayes: Could I say one more thing?

Ashlie Wilbon: Sure.

Deanna Hayes: If we were to do something like that which as I said it could easily be done. If we look at - you were looking at the table Figure 1 sorry Figure 2B1IV-A validity of performance at the clinician level for example. If we took the data that we had and did the simpler table that I think you're talking about, you know, the line would look like what you're seeing there. It would be an upward diagonal line from left to right which is kind of demonstrated by those three columns there.

(Dave Nance): (Dave Nance) here. If I can just take another swing at it and I realize we do need to kind of close out here, you know, the reason I said what I said about you got, you know, photo on the X and photo on the Y is that, you know, the MCII basically is a way running in the background it's for patients to say, you know, I've experienced a meaningful improvement I need to improve eight points or better on the photo score. Okay, that's behind the scenes. Now what we do is you say okay I'm going to divide my units to ten based on something like average improvement on the photo score.

And then on the y-axis I'm going to do what percent of patients got an eight point improvement or better on the photo score. So it's kind of two mathematical spins at the same concept it's not really anything independent. Now I'm very sympathetic to the idea that buried underneath it is this idea of the global rating of change.

And I think I'm trying to do the sort of the spin the math inside out and say well could you possibly get, you know, a percent over eight without somehow automatically implying a higher percentage of people who would say, you know, I did more than three points better than the global rating. I guess it

would be far cleaner but I understand from (Karen) the process is trouble now is just to take the thing independently. And so again make that the y-axis and then it would be - wouldn't have this circularity problem. But I don't know did that help any?

Deanna Hayes: Yes, and we can as I said I regret that we didn't do that. This approach is what has worked in the past with our NQF submissions so we, you know, took that forward and that seemed like it would work based on the feedback from last fall as well otherwise we would have been happy to do it. And what you're asking could easily be done and it would end up looking very similar to what you're seeing with the general trend.

(Dave Nance): Yes, I think so. And I will say I'm very sympathetic to the idea that you seem - you ask for one thing and then you seem to be asked for another thing and another cycle. And as somebody who's been on the other side of this a couple of times in another context I appreciate the difficulty and I'm trying to be sympathetic to that.

Deanna Hayes: Well and for the sake of, you know, we partner with NQF in ever evolving the science and we learn from every interaction. And yes sometimes it's just a matter of a different scientific method panel wants something different from the last and we're okay with that too. But, you know, for the sake of evolving I do appreciate what I've learned from you today and we would definitely intend to roll that in to the ongoing evolution of the science.

Ashlie Wilbon: Okay, that's the - does the Method Panel feel comfortable voting at this point on validity and moving on to the next discussion?

David Cella: Yes.

Ashlie Wilbon: Okay. So please go ahead and select your validity vote. And you should be able to submit for that measure. And then I am going to hand it over to Sam and (Karen) to lead a discussion for the (paps) measure.

Sam Stolpe: Very good, thank you Ashlie Wilbon. This is Sam Stolpe. We're going to be discussing CAHPS Measure 0005. But and this isn't to single them out but it's because we're going to use this as an illustration for a broader discussion that has implications for each of the CAHPS measures. So I won't belabor the point too much with the description of the measure I think everybody on the call has a fairly intimate familiarity with the CAHPS measures but Measure 0005 is the CAHPS Clinician and Group Survey.

And as you know it's a standardized survey instrument that asks patients to report their experiences with their care providers over a period from the preceding six months. So the survey itself has both an adult and a child form each of those has five major domains. And one thing I wanted to remind the Scientific methods Panel just as a general approach to this measure and these types of measures broadly is that we would consider each of those five domains or sections if you will as an independent measure that you may have a different rating on. So you may rate for example Item 5 which is overall rating of provider which consists of a single item you may rate that one as insufficient and the other four would move on from the adult CG CAHPS Survey 3.0?. But that one would need to be revisited and additional testing added as just an example.

If one composite scores low you can elect to say well we thought that one did not meet certain threshold so we were going to say that, that one's low but the other ones move on. So just another way of saying it this isn't an all or none you have to pass all aspects of the survey instrument in order to say that it's reliable and valid you can take it piecemeal. Okay, so that is actually going to

be pertinent to our discussion but I'll continue just describing what we have here. So type of measure of course is an outcome PRO-PM. And the level analysis is that the clinician and group practice. The measure is risk adjusted. It's an optional case mix risk adjustment based on four factors general health status, mental health status, age and education.

Now let's go ahead and move to directly to the reliability discussion. With this measure passed with moderate reliability but was pulled with - by staff specifically to have a discussion around the broader implications associated with the types of testing that were presented by the measure developers for each of these CAHPS surveys. The measure developer calculated chrome block alpha on each composite of the survey. Now while the Methods Panel Members noted an adequate performance generally the Care Coordination Composite for both the adult and child alpha scores was noted to perform at a .55 and .39 respectively.

And the other thing to note is that because the single items elements of this, you know, have nothing to compare them with specifically within that item we didn't have data element testing presented for global items across all of the CAHPS surveys that were submitted.

So what we'd like to do is have a discussion around why we would possibly consider that as a concern or rate this as insufficient for that particular item if in fact NQF has this requirement that per all PRO-PMs we need to have both data element, and score level reliability and validity testing presented. Well what does that actually mean? So we had one other item to note, and that was related to risk of death. So, but I think what I'd like to do is just highlight that very briefly and say that we'll - we'll cover that in some more detail and just go directly into a discussion related to the reliability of the instrument itself.

So the first general concern was related to the low performance of the care coordination portion of that of the survey. So this was pointed out by several committee members but was not necessarily reflected in the vote. So we wanted to make sure that we had an opportunity for discussion with the added reminder that we can tease out individual performance measures within the set of measures that are presented as a cap survey pro PM.

Okay. But that's a local - hadn't opened up the discussion around the care coordination score that was noted by several committee members to be low. Go ahead. Go ahead.

(Ron Hayes): Sorry. This is (Ron Hayes). Are you opening it up to the caps team or the NQF people?

(Dave): Sorry. For the (unintelligible) panel to discuss.

(Ron Hayes): Okay. All right.

(Dave): So if it's helpful, I can reiterate the question that we're discussing.

Man: I can jump in a little bit. You know, since in the end we're answering overall assessment of the liability, I'd have to say I wasn't bothered by either the two Cronbach's alpha levels that were low. I noticed them, I got them yellow highlighted it here. But when you jump down and you look at the providers for a level of the liability, even for these items, that tends to be high.

And the reason why I think that can work is that, you know, you can take two items and call them care coordination and you have to label them something if you want to put them together. Now, if they don't correlate really highly, is reflected in the Cronbach's alpha. So you might say, well that's, you know, it's

weak evidence that there's an underlying factor or that, the thing's reliable, but then - it seems to me anyway, that if you could show it the score level, that they'd behave in a reliable fashion I was willing to then to say, overall across the two levels, looked okay. So at least that was, that was my line of thinking there

Joe Kunisch: And this is Joe, I think I brought this up as a concern, but you know in the explanation and the 2A2.4, you know, I was, I was okay with it because of the composite and the global ratings had the site level reliabilities. And that's kind of how I looked at all of these. You know, what is it actually measuring, a site versus, you know, provider to provider.

(Ditu): This is (Ditu), I have a question about how chrome book alpha was calculated. If I understand, this scale, so you can calculate Krumbach alpha at patient level based on, you know, (unintelligible) respond category. Oh, you can base on - applied a top box procedure so it be (unintelligible). Or you can do at the provider level. For this survey, my impression is you are not using scale score. It is you are using - for each item we did not have scale, you have tried to (unintelligible) and then you identify for each particular item, proportion of patience response, top response.

So really you can positive score it at attribution level. So it's not typical. It's not like you can positive score at patient level and roll that. You're actually roll up to the facilities level, then roll up to composite. So I noticed across multiple test survey, like at hospital, they calculate Krumbach alpha at the right level and, you know, to become a component at the facility level item score. But here I, my impression is it's at a patient level. Maybe a developer can clarify.

(Dave): That's a question for the major developer, we invite your comments at this time.

(Ron Hayes): I'll let Westat be clear on that. But I think what was said before that last comment is really the key. Often we do estimate Cronbach's Alpha at the patient level and we don't consider that to be the important reliability. We do it because everybody wants to see it and it gives you a little information. Really the key is like the previous person said, the site level is what we're interested in because we're not interested in the reliability at the patient level and the site level reliabilities are provided in table, you know, different tables I'm sure you know, around page 10.

This is one example, but so for care coordination and for - and that's all we can do for the global rating items is to do site level of reliability because we only have a single item. That's what our focus is and that's the basis of all the recommendations for using CAPS.

So I don't know if there's - one thing we should say about care coordination is it is very difficult concept to measure and to get items that are internally consistent. And you know, that's something we struggled with for a long time as far as internal consistency reliability. But again our main focus is to make these things reliable at the unit that it's being compared at or reported at.

Man: (Ditu)'s question also included some elements related to top box scoring, which actually we can have the measure developer weigh in on both that issue as well as (Ditu)'s question related to this topbox scoring. (Ditu), am I missing anything else that was inside of your question?

(Ditu): I just, I mean I cannot require you to provide (unintelligible) from measure need to require those data, and score level reliability. And I also want to get

other people's input on the panel because (unintelligible) there are multiple CPAS measures. So different developers seem to be doing that differently. There are some, I mean I work with a lot of hospital cap measures want to delay - I think everything they did is measure at specify level, it's not - so it is one thing like how you specify measure, how you (unintelligible) doing at a different level.

Man: Okay. So we can maybe put a peg in that one and just go to the measure developer to ask them the question that you had before, which I don't know if it needs to be restated or if the measure developer feels comfortable answering the question.

Naomi: This is Naomi Watt from Westat. If that question is for the Cronbach's alpha, was it the - based on the top box score versus the full scale score? It was the full scale score at the patient level. At the respondent level.

(Ditu): Right. So here we are using top box score, right? (Unintelligible) at the facility level, each item you get a top box score, and then you wrote a, nice (unintelligible). For every facility you calculate a top box score for that item for that facility, and then you have five scores and then you composite that five facility-level items specific into one composite score. Am I understanding that right?

(Ron Hayes): Just one thing to note about that is top box is sometimes reported but not always. So we use more of the distribution as well. So sometimes you will see top box presented here for different purposes. But one of the reasons we have reliability know that's not top box is because we, it's off - the scores are also used in a more continuous way in CAPS.

(Ditu): So the question, how do we (unintelligible) the evidence with the measure as specified. I understand some you can use as a (unintelligible) score or you can use top box score. So what each specified purpose you need to have, you know, relevant evidence, all right?

(Karen): This is (Karen) from AQS. Let me give a shot here. First of all, it sounds like you're saying that when they did the Cronbach's alpha calculations, they were not using top box and they weren't, they were just using the whole scale at the patient level. That's typically what we've seen in the past and that's what we, we kind of expect to see. And then when they do their score-level testing since that is specified for top box, they would, you know, use the top box.

And I think the calculation and, developers can correct me if I'm wrong. I think the calculation is as you described it (Ditu), top box items and then averaging over, across how many items in the domain. So that - it sounds like that's what they've done. That sounds like what we have expected and has accepted in the past.

I guess the question is, is there any reason to think that we should be expecting that patient-level data, at top box versus not top box? And I think that's more of a question for you guys. Again, we've, we are - the way they, it sounds like they've done it is what we've accepted in the past. So the question for you might be have we wrongly accepted that in the past?

(Ditu): I think the question - actually I probably didn't make it clear. So that this is as important, at the patient level you can use survey format to calculate Cronbach alpha or you can use top box format to calculate. They probably will be similar. My concern is more because we are not derive a composite score at patient level, then we actually get a score at item level.

So when you (unintelligible) you are not using composite score at patient level. So you can calculate Cronbach alpha at the facility level instead of at patient level. I think this is - I'm just curious, what other panel members think on this issue.

Joe Kunisch: This is Joe, I guess my question would be what would you expect to see in that, in that difference between the two? I mean, are you expecting significantly different results than there - than the method that they chose?

(Ditu): I don't know, so here, right (unintelligible) do it that way and then some measure doing it this way, there are other measures that do it this way. So I'm just saying that, do we assess it either way, or do you think that's a reasonable thing to do? It is relevant when you talk about composite, how you composite them later, right?

Joe Kunisch: Yes. I agree. But I don't think it should be restricted to one way or the other.

Woman: So (Ditu) it sounds like that is another one of these things that we need to put on our list for methods panel discussion later on. Unless (unintelligible) or (Sam) or (Mary Beth) or others, you know, want to ….

Man: I think we've had an ample discussion of reliability on this particular submission. My suggestion's really more structural to NQF and that is, you know, related to the idea that we can, we can in the case for something submitted with five or six performance measures within it, that we can pull one out one way or the other and have different ratings. Let the format be structured so that that's possible to do.

I mean I was among the people, I rated moderate because there were really no concerns about most of them in a minor to moderate concern about the care

coordination. But that's been discussed. So just if we had some way to actually, the form, I'm not advocating five or six separate forms, but some way ono the form to be able to pull one out explicitly would be good.

Woman: Okay. We might be able to handle it. We'll think about how to do that, but that is actually one reason that we wanted to call this to your attention, just to make sure that everybody knows, even though it's not apparent on the form, that you can always call out these individual performance measures.

Man: I think what we heard is that we did, I mean, several people did call out one.

Woman: Yes. It sounds like even though you called it out, you've pointed out concerns, you're still okay with the moderate ratings for that particular one. I'm not hearing them.

Man: I am. I think we can all, I think there’s probably a readiness to vote, unless anyone wants to say more.

Woman: We actually may not need to vote. This is really just a discussion.

Man: Oh, I see. Okay. All right. Right. Yes, yes.

Man 2: Unless the tide has a very significant shift, which it doesn't sound like there is, then we probably don't need to revote. Unless someone’s calling for it, which I don't get the impression that they are. So if that's the case, then we can move to the other item for discussion.

Woman: Right, the global item.

Man: Correct. So let's just go ahead and revisit any residual things that need to be discussed outside of this, outside of this conversation, but for the global items scores. So just to remind you what we're talking about here, there's several items in the CAPS surveys which have a single item. Now, NQS criteria is that you need to show data element testing and obviously for Cronbach's alpha, you know, there is no internal consistency to be discussed for single items.

But you could make the argument that if it is true that demonstrating internal consistency is sufficient for data element reliability, then you, all single items are sufficient by their nature. And it appears that that's what the committee did by passing these measures. But we just want to call out that we have noted that data element testing was not performed for the single items and allow the committee to discuss whether or not that is appropriate or how we could potentially do it differently if in fact we should. So we'll, we'll turn it over to you to discuss

(Dave): This is (Dave), (Dave Narends) is leading a reliability white paper, which is going to pick up these issues and a lot of that is probably more appropriate for that general discussion than on a particular, I don't know if we're still talking about triple oh five. But it, you know, we've talked about and apparently have not acted on it here, but we've talked about stability, you know, test/retest reliability being actually probably more important than internal consistency, but less convenient or less available.

So I guess, you know, that's what I would say is that ideally we would see, you know that these are stable measures then because that's when we really think about reliability and the kind of reliability that you need is an essential for validity downstream. It's the test/retest reliability.

Man 3: Thanks (Dave), I was thinking along the same lines. I'm not sure I'm willing just as a matter of concept or principle to give advance on reliability to single items, although I understand you can say they're internally consistent because it's just one. But I don't think that does the job.

But I'm curious for (Ronnie), if (Ronnie)'s on the call, there's a huge amount of development work in CAPS and way back in the development was there some test reliability done on some of these single items? Get patient survey and then wait for the memory to fade. Then go back three, four days later, a week later. Ask again. Does that exist? I think it probably does or it might.

(Ronnie): Yes. I mean there's a little bit of that. The problem we have is that this is supposed to be update. So we can't give you the whole history of everything that's been done. I also want to point out that we do have the main reliability here and the site level, the lab for the single item. Now you could have three tests as well.

In fact probably one thing, I mean unfortunately I don't know exactly what data set we would use, but you could look at reliability on a single item for the units, you know, over like a year time period, not a test/retest time period. Even that would probably be pretty stable in a lot of cases and I know in Medicare CAPS there would be an opportunity there, but that's not the data we presented.

So historically, yes, there's probably a little bit, I'm also not a huge fan of test/retest reliability except for a single item when you're using it for individual purposes. But there is a little data, but again, it's historical.

Man 3: Yes, that's okay. And again as I said back in the beginning, in this mix of things I tend to be more swayed by the site level reliability statistics. But since

it's essentially a chat box item, the way you have to structure it to review it, just was curious if it was out there already.

(Mark Elliot): This is (Mark Elliot) from Rand. We published a paper about H CAPS a couple of years ago that shows hospital-level correlations over time. And those are pretty high. So we could send that citation if that would be of general interest.

(Karen): This is (Karen) from AQF. I think you've already have in your submission, sufficient, what was the cost score level reliability. So that's not in question. Really the question is, we didn't see what we would call data elements reliability testing for the single item measures. And typically when we see a test/retest, possibly there are other methods for that.

We're just stating A, that we didn't see anything for that at the patient level for the two single-item measures in this particular one, but it's true across the board for all the CAPS measures and not just calling out CAPS measures. It's, it's pretty, it's not unusual for other instrument-based measures not to give us that sort of stuff for data element reliability for single item.

So the difficulty is that our criteria state - that we need to have it. So our question for our methods panel is, are we asking too much for single items or are we asking the right thing and we just are not seeing it. And of course the implications of that is if we should be asking it and we're not seeing it then for the global items, the rating sheet had been insufficient. That's kind of the implications here.

Man: Would any of our panelists like to opine on that? We'd welcome your opinion.

(Ditu): I have this question. If some developers submit measures based on one item only, how do you evaluate that?

(Karen): That's a really great question. We've kind of, (Peter) in the past, I think in some cases has gone down because we said, hey, we want to see some kind of test/retest or something. But it can be tricky because one might could argue that that one item, only makes sense at the score level. So I think we've, it's been kind of a mixed case in the past. So we've kind of done both, I think.

(Ditu): And related question, it's not part of it (unintelligible), in some others I have seen, right? In some other I have seen, right, they have global rating item and then obviously they assume the validity of the global item. You can say, yes or no, it depends. But then they go on to use global rated item as criteria to evaluate other scale score. So first, may I assume the validity? I'm not talking in particularly to this measure, I have seen others. So how do we handle that?

(Karen): Yes, generally we have allowed that in the past because we, you know, it seemed that they are individual performance measures. So in the past we have accepted that.

Joe Kunisch: So, you know, this is Joe, are you then saying this would be a new standard, if we change what's accepted?

Man: Not necessarily. We have in fact had measures go down because they didn't present data element reliability for a single item, then this would just be carrying forward something that we've done in the past. But I can't say it explicitly. But that's the case (Karen), you just mentioned that it was.

(Karen): Yes, I mean again, it's kind of gone both ways in the past. You know, I think what, you know, our dilemma is, you know, I think in the past sometimes they

have gone down because we've pointed out, hey, we want to see this, you know, maybe test/retest or something along those lines, something - and we don't see it and therefore we're sending it back, right, and not passing it through.

But for example the CAPS measures, these don't have it. But those, right now with the votes that we have right now, would go through. So we're being inconsistent and I think I'm just kind of back to my question. You know, should we be expecting to see some sort of reliability testing at the patient level, for the single items? If so, then we need to think about revoting or maybe just making an overall determination for those if that wasn't there, but if that's asking too much, then we need, you know, say that's too much.

(Dave): This is (Dave). I don't think it's asking too much, but as Ron pointed out, given this is a submission, it's not a new submission. You know there wasn't a request for it so they have it and it's available apparently in a paper that can be sent. And so I don't, you know, I'm trying to separate the discussion of triple oh five from the general question and to the general question that I don't think we should relax that criterion, but I also don't think it should be applied to a vote here because this is not a new submission and we know the data exists.

(Paul): This is (Paul), can I make a comment?

(Dave): Sure.

(Paul): Is it possible that some of the items that, quote unquote went down or were not voted favorably on, were designed to assess differences at the individual level? I think we all agree that CAPS is a unit level. That's why we've placed less emphasis on that. If we were posing it to differentiate individual level experiences, I think that might be more applicable.

(Mark): Related to (Paul)'s comment, just to tack onto something, this is (Mark), if you think about the correlation over time and a single global item measure, say at the hospital level, the sources of variance are both true changes in the hospital's performance and the lack of test/retest reliability at the patient level.

So in some sense, establishing high stability at the facility level is a stricter criteria that encompasses the person-level criterion, because its correlations fall due to due to change - due to either a lack of person-level test/retest reliability or true facility-level change. So in that sense, I would think that good stability at the facility level would actually meaningfully inform the question of whether there's adequate test/retest reliability.

(Dave): You're suggesting that a facility can't have good reliability if the question itself is unreliable and you know, that it's a limiting and - you know, I guess that's something we should take up when we talk about this more generally, but again, are we talking about oh five or are we talking about the general issue? We're talking about oh five right?

Woman: We're kind of talking about…

(Ditu): It's a general issue.

Woman: Yes, it's a general issue, but it could affect oh five and every other CAPS measure in front of you. That's kind of where are, right?

(Dave): Right.

Woman: So if yes.

(Dave): Well, I'm just realizing we have about seven minutes and if we're finished with oh five, then let's decide that and then we might have more time to talk about the general issue with the benefit of the experts from Rand on the call. Are we finished with triple oh five?

Woman: I think so. I mean we were using triple oh five as this kind of an exemplar to discuss three items. Number one, the first item was - can make sure that you guys know that you could, you know, pull individual performance measures, you know, even if it's one NQF number. And that was the care coordination discussion.

The second item was this idea of single item, performance measures and not having any kind of reliability testing presented for the - which is against the (unintelligible) requirements. The third item that (Sam) had alluded to was just the risk adjustment for this measure and the other CAPS measures is kind of the typical way that the CAPS measures operate. So they, and I'll paraphrase, the idea is they offered risk adjustment, and note that some people may use it and other people may choose not to use it.

We were going to make the statement that what we understand to be putting, being put forward for endorsement is the risk-adjusted version. And, the question for the methods panel really was along the lines of, we assume that as you were going through, that you were, you had no concerns with the risk adjustment approach and that you realize that that's what's being put forward for endorsement.

(Dave): So we handled the first, and I think we've handled the second, like, at least my suggestion since I guess we're not voting, my suggestion is that if/when this goes forward to the parent committee, if you will, for review, that they'd be notified that, you know, if we didn't, we didn't evaluate the element reliability

and think it should be reviewed in light of, you know, as an exception perhaps based upon your criteria. And then maybe we have the last five minutes for the third question. Is that fair?

(Karen): Sounds good.

(Dave): Okay. All right.

(Alan): Excuse me, this is (Alan) (unintelligible) at Harvard and I'm going to (unintelligible). I just want to point out that, you know, we're talking about the lack of multiple items to do Cronbach alpha for the (unintelligible). We do different kinds of assessments across multiple items, which is when we left the criterion, the rating items, which are general assessments on the specific report items.

And we do look at the relationships between items but it's not symmetrical, the criterion item of general ratings are looked at it in a different way than the others. And so the regression approach rather than the correlation approach is appropriate there. But I think it gives you a lot of the same kind of information about finding the existence of relationships among items that are (unintelligible).

(Dave): I bet (Dave Nance) is taking taken notes and we'll be taking this up in that reliability paper, huh (Dave)?

(Dave Nance): Was that (Alan) on there just now?

(Dave): Yes.

(Alan): Hi (Dave).

(Dave Nance): Hi (Alan). Yes, I'm happy to take notes, instructive every time I talk to you.

(Dave): Do we have time to handle the risk adjustment question?

Man 2: Yes. Why don't we go ahead and try to tackle that. So, once again, just our interpretation is that the risk adjusted measures, what we're endorsing, we assumed that with your - from your moderate scores that you gave to that, that you're happy with the risk-adjusted method, and just want to confirm that that's true. Because this was listed as optional, so that, that's what prompted the concern.

The assumption is that if it is impactful, that the - if the models are actually showing that the risk adjustment makes a difference, then what we've done in the past is say that we're - we would like to see risk adjustment and that the onus is on the measurement steward to provide the modeling.

(Karen): So maybe in the interest of time, did anybody review this measure and kind of disregard the risk adjustments, or did anybody feel like that the risk adjustment was inappropriate and therefore, may need a different rating on this or any of the other CAPS measures? Not just singling out triple five. We just want to verify.

(Dave): Well, this is (Dave). I certainly didn't disregard it. I did wonder about the optional thing. But I figured that since this is not a new measure, this had gone through this way in the past, I think - in fact I think I even read that. So I thought, I didn't see any reason to change course. So I went along with it. But I didn’t disregard it. I don't really understand why they're not - why it's done, the model that was presented as a great model, but it's not required. I didn't, I didn't really understand that.

(Ron Hayes): So (Dave) do you mean why the CAPS team doesn't require it or why NQF…

(Dave): (unintelligible)

(Ron Hayes): We can't, we here in (unintelligible) in general, but we can't enforce that like on NCQA. So we can say this is a good approach, but that's all we can do.

(Dave): So in other words, CMS and/or NQF or NCQ, they could require the risk adjustment. You're saying?

(Ron Hayes): Well, NCQ may not require it. That would be the other way of looking at it. And CMS uses case metrics routinely.

(Karen): So it is a little confusing. I think it's been confusing over the years. So I, you know, NQF doesn't endorse unadjusted version and adjusted version. So what we would be endorsing and what we've understood that we've endorsed all these years is the risk-adjusted version. Knowing that out in the world, people can choose to not use that. We can't be, you know, NQF has no…

(Dave): Okay, well, that's very helpful. I mean, I was, you know, I was working off of a submission that made it seem like the NQF endorsed measures. Maybe I misread it, but I did get the impression that the NQF-endorsed measure would say, well, you can use it or not, it's up to you and that didn't make sense to me. Now it makes sense.

(Ron Hayes): Right, because we're, our requirement is that you test the specifications. So we need to say what it is specifically that we're talking about. So we just want to make sure that we're talking about the risk-adjusted version and that the measure testing falls in line with that. If that was your assumption.

(Dave): Well, if there was an assumption it was that, meaning I wasn't, I wasn't clear. Maybe I should have been, but I get it now and I thought the risk adjustment was quite good and thorough and I'm glad that it's part of the required measure if it goes forward.

Okay. What do others think?

(Sam): Yes, (Dave), this is (Sam). I had a similar read. I mean, this to me almost seems like an implementation issue. I mean, we're looking at, we're evaluating the risk-adjusted measures so that's kind of a big assumption I proceeded under.

(Dave): Well, I think we're good then. I keep hearing no other comments. We're hitting the hour and some …

(Sam): That's implied consent, we can go with that.

Man 3; That's good. That's why I - I got a really hard stop top of the hour here. I got to run. But I'm okay where we are.

(Dave): All right, we'll cut loose.

(Karen): All right. So I think we're good. Thank you all, and we'll be in touch.

(Dave): Thanks.

(Karen): Thank you all.

(Dave): Bye.

END